RESEARCH ARTICLE

Asian Research Association

Indian Journal of
LANGUAGE AND LINGUISTICS

DOI: 10.54392/ijll2633

Artificial Intelligence in Language Education in the Generative-AI Era: A Systematic Integrative Review of Pedagogical Evidence, Methodological Quality, and Responsible Adoption

Md Riaz Hasan [a, *], Fariha Sultana [b]

[a] School of Computer Science & School of Software, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China

[*] Corresponding author Email: riazhasan.se@gmail.com

DOI: https://doi.org/10.54392/ijll2633

Received: 26-04-2026; Revised: 30-07-2026; Accepted: 05-08-2026; Published: 12-08-2026

Abstract: Artificial intelligence has grown quickly in language teaching, particularly since the introduction of generative AI and large language models. Even so, the existing evidence is dispersed across various technologies, teaching applications, learner groups, and research approaches. This systematic integrative review examined 40 primary empirical or technical studies. It also used 15 systematic reviews, meta-analyses, and bibliometric syntheses to provide contextual evidence. The selected publications were published between 2017 and July 2026, while the majority of the primary evidence appeared after 2022. The review employed descriptive mapping, thematic synthesis, design-sensitive methodological assessment, and framework development. The primary papers were evaluated based on AI technology, educational application, language skill, learner group, research design, outcome type, multilingual representation, methodological quality, and ethical or governance concerns. Writing assistance, automated feedback, speaking practice, learner perceptions, teacher support, and AI-assisted pedagogy were the most extensively researched topics. In contrast, reading, listening, younger learners, low-resource languages, delayed learning outcomes, and validation across contexts received less attention. The evidence points to potential short-term improvements for writing revision, speaking practice, motivation, engagement, and teacher support. However, confidence in these findings is limited because many studies used small sample sizes, brief interventions, single-institution settings, self-reported measures, incomplete model and prompt reporting, and minimal external validation. Generative systems increased interactivity, adaptability, and task coverage but also introduced output variability, hallucination, authorship ambiguity, privacy risk, and reproducibility problems. Based on the integrated findings, the review proposes a responsible-adoption framework connecting AI-system characteristics, pedagogical design, human–AI interaction, contextual conditions, educational outcomes, and institutional governance. The findings indicate that educational value depends less on technological novelty than on task alignment, teacher mediation, learner agency, methodological rigour, multilingual inclusion, and accountable human oversight.

Keywords: Artificial Intelligence, Language Education, Generative AI, Large Language Models, Conversational AI, Human-AI Interaction, Academic Integrity, AI Governance

1. Introduction

Artificial intelligence has quickly evolved from a specialized area of study to a useful part of language instruction. Currently, learning activities can be personalized, written and spoken performance can be assessed, instructional materials can be produced, translation can be supported, lesson planning can be supported, and learner progress can be tracked via AI-supported systems. The variety of technology mediated activities that are accessible to students, educators, organizations, and assessment bodies has expanded with the increasing presence of these systems. The language-intensive aspect of teaching and learning contributes to AI's educational relevance. Repeated practice, prompt feedback, interaction, assessment, and adaptation to individual proficiency are all necessary components of language education. In low resource settings or large classes, it is challenging to consistently meet

these needs. By offering scalable support outside of regular classroom hours, AI technologies can help to partially overcome these constraints. While teachers can utilize AI to create materials, identify learner needs, and manage routine feedback tasks, students can get explanations, practice speaking, edit writing, or complete adaptive activities on their own. AI assisted language instruction cannot be classified into a single pedagogical or technological category. The phrase refers to traditional machine learning models, automated writing assessment systems, deep neural networks, transformer-based natural language processing, task-oriented chatbots, generative AI, and large language models. These instruments differ in goal, adaptability, transparency, dependability, and educational risk. Their impacts should be evaluated in terms of the specific task, learner group, instructional design, and amount of human supervision (Rahman *et al.,* 2024; Liang *et al.,* 2024; Huang *et al.,* 2023; Sharadgah and Sa'di, 2022).

The use of AI in language training has progressed through multiple overlapping technological periods. Earlier applications relied primarily on classical machine learning and rule-based language processing. These systems were utilized for classification, learner prediction, error detection, automated scoring, and early forms of personalization. Their teaching functions were limited, but their results were frequently consistent within well defined tasks. During this time, automated writing assessment emerged as one of the most widely used applications. These systems used grammatical, lexical, syntactic, and discourse-level elements to provide scores or corrective feedback. They increased the speed and availability of writing assistance, but had lower proficiency to interpret context, rhetorical aim, and learner intention (Rahman *et al.,* 2024; Liang *et al.,* 2024; Huang *et al.,* 2023; Sharadgah and Sa'di, 2022).

Deep learning methods later strengthened the ability of AI systems to process complex spoken and written input. Speech recognition, machine translation, semantic classification, and language production were all enhanced by neural architectures. By allowing systems to represent relationships across longer sequences, transformer-based NLP significantly strengthened contextual language modelling. While generative transformer architectures allowed for more flexible text creation, BERT-type models supported classification, semantic analysis, and scoring. Conversational AI made interaction a key component of education. Vocabulary practice, grammatical exercises, speaking rehearsal, dynamic assessment, and guided feedback were all supplemented by task-oriented, retrieval-based, and scripted chatbots. Though their discourse flexibility remained limited, these systems provided opportunities for repeated practice in low-pressure environments. A more significant shift was brought about by the development of generative AI and LLMs. Within a single interface, systems like ChatGPT and similar models can produce explanations, comments, dialogue, translations, questions, summaries, lesson plans, and longer written content. Their accessibility and broad functional range have accelerated adoption in language instruction. However, their probabilistic behaviour, limited transparency, and ability to produce convincing but inaccurate content have also created new educational and governance concerns.

AI offers several opportunities for personalised learning. Adaptive systems can identify learner challenges, adjust task difficulty, organize activities, and provide explanations that are tailored to individual needs. Generative systems can also generate examples for various competence levels, respond to learner inquiries, and alter feedback based on the prompt. Another significant benefit is immediate feedback. AI systems can help with repeated drafting, suggest revisions, and identify grammatical errors in written work. Speech-enabled devices and conversational agents can give repeated opportunities for communication practice. This rapid response may stimulate engagement and shorten the time between student performance and correction. AI also enables interaction at scale. In large classes, students may have limited opportunity for individual speaking practice and detailed feedback. Conversational agents and tutoring systems enable students to practice at their own pace and continue learning beyond the classroom. Automated assessment may also help teachers reduce their workload by enabling formative evaluation, diagnosis, and scoring. Multilingual support is another useful feature. LLMs, multilingual chatbots, and translation systems can offer bilingual scaffolding, cross-linguistic explanations, and wider access to learning resources. However, their performance varies between languages, dialects, and accents, especially in low-resource contexts. In this situation, teacher assistance is just as essential. Artificial intelligence can help with lesson planning, question preparation, content adaptation, feedback drafting, and professional growth. However, these advantages rely on institutional support, thorough evaluation, and proper teacher preparation. AI may eliminate some repetitive tasks, but it also introduces new responsibilities in academic integrity, privacy, output verification, and assessment redesign (Huang *et al.,* 2023; Sharadgah and Sa'di, 2022; Rahman *et al.,* 2024; Liang *et al.,* 2024).

The literature is still dispersed despite significant advancements in research. A single technology category, such as chatbots, intelligent teaching systems, automated writing evaluation, or LLMs, is the subject of numerous

studies. This limited arrangement makes comparing different generations of AI more difficult. Another constraint arises from the recent heavy emphasis on ChatGPT. Reviews of ChatGPT are valuable for understanding recent adaptation, but they may neglect previous work on machine learning, natural language processing, speech technologies, and conversational systems. These early breakthroughs laid much of the scientific and educational groundwork for modern applications. Without a cross-generational comparison, it is difficult to discern if newer systems represent genuine educational progress or simply broader technical capabilities. Existing reviews also frequently highlight affordances. Benefits like personalization, efficiency, interactivity, and scalability are frequently highlighted, even when the data is primarily based on impressions, demonstrations, or technical performance. Without increasing learner competency, a system might produce fluent feedback or obtain high scoring agreement. Therefore, it is crucial to distinguish between technical performance and instructional efficacy.

Additionally, multilingual equity is not thoroughly investigated. English and other high-resource languages are the subject of a large portion of the evidence. There is little focus on low-resource languages, regional variations, multilingual classrooms, and accent variety. Because of this, assertions of multilingual competence might not accurately reflect educational quality in different linguistic environments. Another area with poor integration is governance. Rather than being circumstances that influence educational validity, hallucinations, bias, privacy, academic integrity, authorship, fairness, and institutional accountability are frequently considered as distinct issues. Reviews seldom make a direct connection between these hazards and institutional policy, learner agency, teacher supervision, or pedagogical design (Rahman *et al.,* 2024; Liang *et al.,* 2024; Huang *et al.,* 2023; Sharadgah and Sa'di, 2022).

Previous reviews have provided valuable accounts of chatbots, automated writing evaluation, intelligent tutoring systems, generative AI, and large language models in language education. These reviews have clarified major applications, adoption patterns, and emerging concerns. However, the evidence is still divided by research goal and technology category. A single family of systems is typically the focus of technology-centered evaluations. A single language competence, learner group, or educational environment is frequently the focus of pedagogical reviews. The integration of system features, pedagogical design, human-AI interaction, learning evidence, multilingual representation, methodological quality, and institutional governance remains insufficiently integrated. AI in language instruction has been examined from a number of angles in recent publications. Chatbots, learning a second language, automated evaluation, generative AI, multilingual apps, and technology adoption are some of these. However, there are significant differences in their analytical coverage. Certain reviews focus primarily on technology applications or publication patterns. Others look at ethical issues, intervention results, or learner views. Six analytical domains i.e., technological features, pedagogical integration, result validity, methodological quality, multilingual representation, and governance, are combined into a single evidence structure in this review.

The tendency to view technological capability as proof of educational value is another issue. Without enhancing long-term language proficiency, a system may offer seemingly trustworthy scores, provide fluent comments, or maintain lengthy interactions. Different types of proof include technical accuracy, user satisfaction, perceived utility, engagement, and learning achievement. It is incorrect to consider them to be comparable. In the age of generative AI, this distinction is particularly important. Although LLM-based systems can accomplish multiple instructional tasks through a single interface, their results may differ depending on the prompts, sessions, and model versions. Reproducing such outputs can likewise be challenging.

Additionally, languages, learner groups, and educational situations receive uneven attention in the literature. A large portion of the present evidence is focused on English language learning, university students, academic writing, speaking practice, and learner perceptions. Much less focus is placed on reading, listening, younger students, low-resource languages, regional linguistic variations, and long-term educational effects. Ethical issues such as hallucinations, bias, academic integrity, authorship, privacy, fairness, uneven access, and institutional responsibility. However, these issues are not always directly connected to methodological quality or instructional validity.

Therefore, this review focuses on AI-supported language instruction in the context of generative AI. Conversational systems, deep learning, automated evaluation, and machine learning were previously employed as conceptual and historical comparators. The review does not claim that the evidence is evenly distributed across technological generations. Instead, it examines how various educational roles and system features influence how present findings are interpreted. The following research questions are addressed in the review:

RQ1. How is the development of AI-supported language education represented in the selected literature, and how has the research emphasis changed in the generative AI era?

RQ2. Which AI technologies, system characteristics, language skills, and pedagogical applications have been investigated?

RQ3. How have AI systems been integrated into language teaching, learning, feedback, assessment, and teacher support?

RQ4. What linguistic, educational, affective, behavioural, teacher-related, and assessment outcomes have been reported?

RQ5. How do conventional task-specific systems, conversational systems, and LLM-based applications differ in flexibility, pedagogical role, evidential support, reproducibility, and risk?

RQ6. How do the applications and reported outcomes vary across languages, educational levels, learner groups, geographic contexts, and language-resource conditions?

RQ7. What research designs, evaluation methods, methodological limitations, and reporting practices characterise the evidence base?

RQ8. What concerns related to ethics, academic integrity, privacy, fairness, authorship, access, and institutional governance have emerged?

RQ9. Which research and implementation gaps require priority attention?

The review makes four contributions. First, it makes a distinction between technological performance and objectively measured educational outcomes as well as learner and teacher perceptions. Second, it looks at technical and primary empirical research independently of secondary evidence syntheses. This division maintains clarity in the interpretation of the evidence and lowers the possibility of double counting. Third, it evaluates the methodological limitations of educational interventions and technical AI studies through design-sensitive criteria. Fourth, it develops an integrative framework that connects AI-system characteristics, pedagogical design, human–AI interaction, contextual moderators, educational outcomes, multilingual equity, and institutional governance. The framework is presented as a synthesis-derived model for responsible adoption rather than as a validated causal model.

2. Conceptual and Theoretical Foundations

The term "artificial intelligence" in language education refers to a broad category of technologies that vary greatly in terms of computational architecture, task flexibility, pedagogical function, and level of human supervision. Therefore, before comparing their educational applications and results, it is vital to establish clear conceptual limits. The technological hierarchy that underpins the review is established in this section, which also identifies the main conversational and generative systems, places AI within well-established traditions of technology enhanced language learning, and lists the theoretical stances that are employed to interpret human-AI interaction (Yang and Kyun, 2022; Jeon *et al.,* 2023; Huang *et al.,* 2023).

2.1 Artificial Intelligence, Machine Learning, and Deep Learning

Artificial intelligence, the broadest term, describes computing systems created to carry out tasks typically associated with human intelligence, such as perception, prediction, reasoning, language processing, decision support, and adaptive interaction. AI can be used in language education to analyze student performance, offer feedback, create educational materials, facilitate communication, automate evaluation, or customize learning tasks. Machine learning, a branch of artificial intelligence, uses data to find patterns instead of depending solely on manually defined rules. While unsupervised approaches find latent structures or clusters, supervised machine learning employs labeled examples to categorize, predict, or score new inputs. For automated essay scoring, error categorization, learner modeling, performance prediction, and adaptive sequencing, earlier language-education applications mostly relied on machine-learning techniques (Huang *et al.,* 2023; Yang and Kyun, 2022). Deep learning is a branch of machine learning that uses multilayer neural networks. Large amounts of text, speech, image, or multimodal data can be used to enable systems to learn complex representations. Speech recognition, machine translation, semantic categorization, automated feedback, and language production were all improved by deep-learning systems.

Therefore, deep learning is a specialized area of machine learning, and artificial intelligence encompasses machine learning as part of the technical hierarchy. Deep learning and transformer architectures are the foundation of both modern large language models and generative AI (Yang and Kyun, 2022; Huang *et al.,* 2023). These categories should not be treated as educational classifications. When consistency and task control are needed, an older machine-learning system might be more suitable than a generative model, and two systems might have similar architectures but serve different instructional functions.

2.2 Natural Language Processing and Transformer Models

Computational techniques for analyzing, interpreting, creating, and transforming human language are referred to as natural language processing. Automated writing evaluation, grammatical error detection, essay grading, translation, speech transcription, question formulation, sentiment analysis, semantic comparison, and conversational engagement are examples of NLP applications in language education (Wang *et al.,* 2024; Jeon *et al.,* 2023). In the past, rule-based processing, statistical language models, and manually created linguistic features were frequently used in NLP systems. Deep learning shifted this approach towards distributed representations called embeddings. Words, sentences, or documents are encoded as numerical vectors using embeddings, which use spatial proximity to convey linguistic similarity. By enabling a word's representation to adapt to its surrounding language, contextual embeddings enhanced this process (Wang *et al.*, 2024; Jeon *et al.,* 2023). An attention mechanism used in transformer models allows a system to recognize relationships between words in long sequences. Transformers enable large-scale pretraining and process contextual interactions more effectively than previous recurrent structures. BERT-type models are encoder-based transformers that are primarily used for language-understanding tasks, including text analysis, error detection, classification, and semantic similarity. They can account for both prior and following context because of their bidirectional processing. Language sequences can be predicted and generated by generative transformer systems. They serve as the foundation for many modern LLMs and GPT-family models. These systems can be modified by prompting, fine-tuning, retrieval augmentation, or instruction alignment after being pretrained on large textual corpora. Their ability to generate explanations, feedback, dialogue, examples, translations, questions, and extended content makes them useful for teaching. However, these systems produce probabilistic outputs. Even fluent responses may therefore be inaccurate, inconsistent, or unsuitable for a particular learner or learning context (Wang *et al.,* 2024; Jeon *et al.,* 2023).

2.3 Conversational AI and large language models

Conversational AI encompasses a variety of technologically diverse systems. Scripted chatbots adhere to predetermined dialogue paths or rules based on keywords. Their responses are predictable and can be aligned with specific learning objectives. But when learner input exceeds their designed scope, they are unable to respond effectively (Du and Daniel, 2024; Jeon *et al.,* 2023; Li *et al.,* 2024; Huang *et al.*, 2022). Task-oriented systems are designed for specific educational purposes. Grammar correction, dynamic evaluation, pronunciation assistance, vocabulary practice, and speaking rehearsal are examples. These systems might make use of speech recognition, rule-based techniques, machine learning, or retrieval mechanisms. Nevertheless, their interaction is still restricted to the activity or educational objective for which they were intended (Du and Daniel, 2024; Jeon *et al.,* 2023; Li *et al.,* 2024; Huang *et al.,* 2022). Retrieval-based systems do not create new language; instead, they choose answers from an existing database. Although the available response set limits their conversational flexibility, they provide greater control over responses and lower the likelihood of hallucinations. New responses are dynamically generated by generative conversational systems. They may maintain more extensive communication, modify explanations, simulate roles, and respond to unexpected queries. Although their adaptability makes them appealing for language instruction, their results need to be verified because they could contain pedagogical, linguistic, or factual errors (Du and Daniel, 2024; Jeon *et al.,* 2023; Li *et al.,* 2024; Huang *et al.,* 2022). Generative neural systems trained on large corpora and parameterized at a scale that allows for broad linguistic capability across tasks are known as large language models. Not every LLM application is conversational, and not every conversational system is an LLM. Without serving as a conversational partner, an LLM can serve as a writer, scorer, translator, lesson creator, or analytical tool. This distinction is important because educational risk depends on the role assigned to the model, not only on its technological category.

2.4 Technology-Enhanced Language Learning

AI-supported language instruction extends the established practices of computer-assisted and mobile-assisted language learning. CALL includes the use of computers for communication, practice, assessment, feedback, and language instruction. MALL allows for situational, flexible, and self-paced learning by extending these features to mobile devices (Huang *et al.,* 2022; Sharadgah and Sa'di, 2022; Yang and Kyun, 2022). By modeling learner performance and modifying instruction accordingly, intelligent tutoring systems offered adaptive guidance. Automated feedback systems provided prompt feedback or assessment for writing, grammar, and pronunciation. Learner data were used by adaptive-learning platforms to change the order, level of difficulty, or content of tasks. By facilitating open-ended engagement, improving responsiveness, and broadening the spectrum of analysable language, AI amplifies these functions (Huang *et al.,* 2022; Sharadgah and Sa'di, 2022; Yang and Kyun, 2022). However, technological improvement does not necessarily translate into educational improvement. Curriculum, task design, teacher mediation, assessment, and learner agency are all part of larger instructional systems that still incorporate AI applications. Whether or not they foster meaningful language use, reflection, communication, and independent competence determines their worth.

2.5 Relevant Learning Theories

Several learning theories provide valuable interpretive foundations. According to sociocultural theory, language acquisition occurs through involvement, scaffolding, and mediated interaction. AI has the potential to serve as a mediational tool, but its usefulness will depend on how it supports interaction rather than separates students from human contact. (Li *et al.,* 2024; Guo *et al.,* 2024a; Jeon, 2023; Yang and Kyun, 2022). According to constructivism, students actively construct their knowledge through experience, interpretation, and reflection. According to this viewpoint, AI is beneficial for education when it encourages investigation, editing, and problem-solving as opposed to providing definitive solutions for passive acceptance (Jeon, 2023; Yang and Kyun, 2022; Guo *et al.,* 2024a; Wei, 2023). Activity theory places technology in the context of a larger system that includes students, instructors, resources, regulations, goals, and institutional settings. This viewpoint is especially pertinent to research looking at how AI affects writing, feedback, classroom instruction, and teacher accountability. Self-regulated learning describes how students organize, track, assess, and modify their learning. AI may strengthen self-regulation through personalised guidance and immediate feedback. However, it may weaken this capacity when learners transfer essential cognitive tasks to the system (Guo *et al.,* 2024a; Jeon, 2023; Wei, 2023; Li *et al.,* 2024). Self-determination theory emphasises relatedness, competence, and autonomy. By enabling students to practice at their own pace, AI may promote competence and autonomy. However, meaningful choice and social connection may be weakened by over-automation or decreased human interaction. When evaluating whether AI-generated explanations and feedback decrease unnecessary cognitive effort or cause overload through long, complicated, or poorly calibrated information, cognitive load theory is helpful. The Community of Inquiry framework aids in the explanation of teaching presence, social presence, and cognitive presence in AI-mediated learning environments. Perceived utility and usability are addressed by the Technology Acceptance Model. TPACK and AI-TPACK, on the other hand, concentrate on how educators incorporate content, pedagogical, and technological knowledge in practice (Li *et al*., 2024; Guo *et al.,* 2024a; Jeon, 2023; Yang and Kyun, 2022).

2.6 Human-Ai Roles in Language Education

The reviewed studies assign several distinct roles to AI. It may act as an instructor by providing explanations, practice opportunities, and flexible support. It may also function as an assistant that helps learners or teachers complete specific tasks while human users retain control. It scores, diagnoses, or offers comments in its capacity as an assessor. It participates in problem-solving, planning, and iterative writing as a collaborator. It supports communication and speaking practice as a conversational partner. It creates lesson plans, examples, questions, and draft text as a content generator.

As a tool for decision-making, it offers data that educators or organizations can analyze before taking action (Ji *et al*., 2023; Li *et al*., 2024; Ji *et al*., 2024; Al-khresheh, 2024). The educational risks associated with these roles vary. Strong pedagogical alignment is typically necessary for tutor and conversational-partner roles, whereas high dependability, transparency, and human review are necessary for evaluator and decision-support roles. The operational boundaries between AI, machine learning, deep learning, conversational AI, generative AI, and LLMs are

defined in Table 1, which also connects these categories to their primary roles in language education (Ji *et al.*, 2023; Li *et al.,* 2024; Ji *et al.,* 2024; Al-khresheh, 2024).

Table 1. Definitions and operational boundaries of AI-related technologies in language education

Concept	Operational definition	Typical language-education functions	Boundary for this review
Artificial intelligence (AI)	Umbrella term for computational systems performing prediction, classification, language processing, generation, adaptation, or decision support.	Tutoring, feedback, scoring, content generation, learner modelling, translation, speech processing and teacher support.	Includes rule-based, machine-learning, deep-learning, conversational and generative systems with direct language-education relevance.
Machine learning (ML)	Subset of AI that learns patterns from data to classify, predict, rank or score.	Automated scoring, error classification, performance prediction and early personalisation.	Excludes manually programmed software without adaptive, predictive or inferential functions.
Deep learning (DL)	Subset of ML based on multilayer neural networks that learn complex representations from large datasets.	Speech recognition, neural translation, semantic analysis, automated feedback and language generation.	Treated as a computational architecture rather than a pedagogical category.
Natural language processing (NLP)	Computational processing, interpretation and generation of human language.	Text analysis, scoring, feedback, translation, question generation and dialogue.	Includes statistical, neural and transformer-based methods when linked to education.
Conversational AI	Systems that support dialogue through scripted, retrieval-based, task-oriented or generative mechanisms.	Speaking practice, tutoring, dynamic assessment, question answering and classroom interaction.	Not all conversational AI uses LLMs; rule-based and retrieval systems are retained separately.
Generative AI	Models that create new text, speech, images or multimodal outputs in response to user input.	Feedback, explanations, drafting, question generation, lesson planning and simulation.	Includes generative systems beyond LLMs; output is probabilistic and requires verification.
Large language models (LLMs)	Large transformer-based generative models pretrained on extensive corpora and adapted through prompting, alignment or fine-tuning.	Open-ended tutoring, writing support, scoring, translation, dialogue and teacher assistance.	Not all LLM uses are conversational; analysis depends on assigned educational role and risk.

3. Review Methodology

3.1 Review Design

This study adopted a systematic integrative review design. The design was selected because the evidence included educational interventions, qualitative studies, surveys, mixed-method investigations, technical validation studies, systematic reviews, meta-analyses, and bibliometric syntheses. The study topics, units of analysis, outcome measurements, and methodological quality criteria of these types of evidence varied significantly. For the entire body of evidence, a statistical meta-analysis was therefore inappropriate. Rather than a comprehensive bibliometric census of all works on AI in language teaching, the review sought to give an organized and critical synthesis. Empirical educational studies and technical studies directly pertaining to language instruction, learning, feedback, assessment, or teacher support comprised the main analytical corpus. Bibliometric studies, meta analyses, and systematic reviews

were kept as secondary evidence. These resources aided in the review's positioning, comparison of current classifications, identification of recognized patterns, and contextualization of the primary study findings.

Primary and secondary evidence were kept separate when reporting participant characteristics, intervention outcomes, study design frequencies, and patterns of methodological quality. This separation reduced the possibility of double counting because some primary studies may also have appeared in the included reviews. The review consequently interprets earlier machine-learning, automated-assessment, deep-learning, and chatbot technologies as historical and conceptual comparators. It does not treat the final corpus as an evenly distributed empirical sample across technological generations (Huang *et al.,* 2023; Wu, 2024; Liang *et al.,* 2024; Rahman *et al.*, 2024). The synthesis addressed six connected domains: AI-system characteristics, pedagogical applications, educational and technical outcomes, methodological quality, multilingual and contextual representation, and ethical or institutional governance. The resulting framework was developed from patterns identified across these domains.

3.2 Protocol and Reporting Standards

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 framework guided the review procedure. The reporting of information sources, study identification, eligibility evaluation, inclusion decisions, and presentation of the final study-selection procedure were all guided by PRISMA 2020. PRISMA-S principles also guided the reporting of database platforms, search concepts, field limitations, supplemental searching, and record-management practices (Rethlefsen *et al.*, 2021; Page *et al.,* 2021). A PRISMA-style flow diagram that illustrates the identification, screening, eligibility evaluation, and ultimate inclusion of studies illustrates in figure 1 the study-selection process. Prior to the final analytical step, the review protocol specified the research objectives, publication period, search concepts, eligibility criteria, evidence categories, data-extraction fields, quality-appraisal criteria, and synthesis techniques. In order to minimize arbitrary inclusion decisions and preserve uniformity across technological, linguistic, educational, and methodological categories, this protocol-based approach was employed.

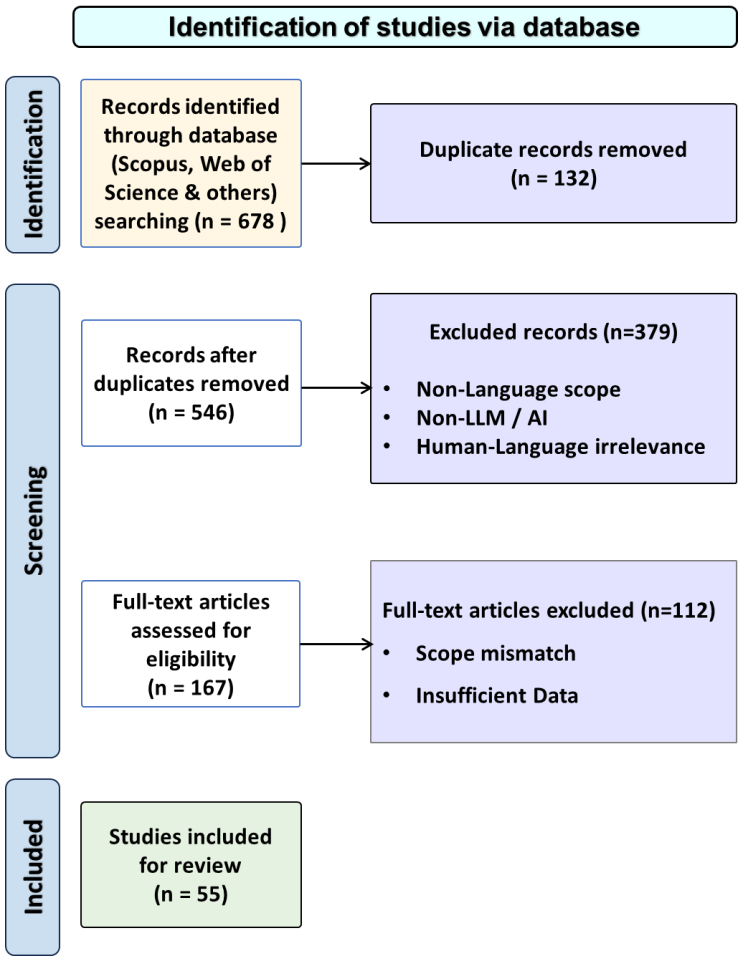

Figure 1. PRISMA 2020 Flow diagram for study identification and selection.

3.3 Information Sources

The literature search encompassed databases for transdisciplinary, educational, linguistic, psychological, and technological research. Scopus, Web of Science Core Collection, ERIC, PsycINFO, Linguistics and Language Behaviour Abstracts, IEEE Xplore, and the ACM Digital Library were the main sources of information. These databases were chosen because applied linguistics, computer-assisted language learning, educational technology, language assessment, psychology, human-computer interaction, and computer science all conduct research on AI in language education (Albedah, 2025; Huang *et al.,* 2023; Rahman *et al.*, 2024). In order to increase coverage and find pertinent research that might not have been found by database searching alone, additional retrieval techniques were employed. Backward reference searching, forward citation tracking, reviewing reference lists in pertinent systematic reviews and meta-analyses, searching publisher platforms, DOI-based verification, and focused searches for research on multilingual education, low-resource languages, teacher perspectives, academic integrity, and emerging LLM applications were some of these methods (Albedah, 2025; Huang *et al.*, 2023; Rahman *et al.*, 2024). The publication period covered 2015 to 2026. From traditional machine learning, automated writing evaluation, speech recognition, and intelligent tutoring systems to transformer-based NLP, conversational AI, generative AI, and LLM-supported teaching and assessment, this time frame was chosen to document the evolution of AI-supported language education.

3.4 Search Strategy

Four conceptual categories representing AI technology, linguistic domains, educational contexts, and pedagogical applications were integrated into the search strategy. The basic structure was maintained while search phrases were modified to fit the syntax and field structures of specific databases. Artificial intelligence, machine learning, deep learning, natural language processing, automated language processing, transformer models, conversational AI, chatbots, intelligent tutoring systems, generative AI, large language models, ChatGPT, GPT-based systems, BERT-based models, speech recognition, and machine translation were all covered in the AI-technology block. The language-domain block included terms such as language, linguistics, multilingualism, bilingualism, second-language learning, foreign-language learning, academic writing, vocabulary, grammar, speaking, pronunciation, listening, reading, translation, literacy, and language assessment.

Education, teaching, learning, learners, students, teachers, classrooms, pedagogy, curriculum, instruction, feedback, assessment, tutoring, professional development, teacher education, and technology-enhanced learning were all included in the educational-context block. Automated writing evaluation, automated essay scoring, automated feedback, personalized learning, adaptive learning, language tutoring, conversational practice, pronunciation assessment, vocabulary learning, academic writing support, lesson creation, teacher support, learning analytics, AI literacy, and academic integrity were all included in the application block. The central search structure integrated terms related to education and the language domain with phrases related to AI technology. Application-specific searches were also conducted to identify studies on particular language skills and specialised educational uses. In order to limit the retrieval of purely technical NLP studies without a direct educational or human-language dimension, search strings were adjusted to strike a compromise between sensitivity and specificity.

3.5 Eligibility Criteria

Research that was written in English, published between January 2015 and July 2026, and presented as peer-reviewed journal articles, complete conference papers, systematic reviews, meta-analyses, or bibliometric reviews qualified. An identified AI, machine learning, deep learning, NLP, chatbot, conversational AI, generative AI, or LLM-based technology had to be investigated or critically analyzed in eligible studies. It had to be directly related to language instruction. Language teaching, learning, automated feedback, language evaluation, academic writing, speaking, pronunciation, reading, listening, vocabulary, grammar, translation, multilingual education, teacher support, or any other obviously educational language-related application had to be covered in the study. The research objective, technological intervention, educational or linguistic setting, study methodology, and main outcomes had to be sufficiently reported in empirical studies. Experimental, quasi-experimental, survey-based, qualitative, mixed-method, case-study, design-based, classroom, technical-validation, and learning-analytics investigations were among the primary empirical studies. Technical studies that assessed models or systems directly related to language instruction were included. Bibliometric studies, meta-analyses, and systematic reviews were kept as secondary

evidence. The review was contextualised, existing taxonomies were compared, established research trends were identified, evidence gaps were validated, and reference tracing was supported. Only when conceptual publications directly influenced the review's theoretical or instructional framework were they utilized. Publications were excluded when they were editorials, commentaries, book reviews, news items, short abstracts, duplicate reports, non-peer-reviewed documents, or purely technical NLP studies without a direct educational, learner, teacher, assessment, or human-language-learning application. Studies were also excluded when their methodological or outcome information was insufficient for reliable classification after all reasonably accessible full-text sources had been checked. Perceived novelty was not used as an independent eligibility criterion. Eligible studies were retained even when they reported null, negative, repetitive, or implementation-related findings. The inclusion and exclusion criteria applied in this review are summarized in Table 2.

Table 2. Inclusion and exclusion criteria

Domain	Inclusion criteria	Exclusion criteria
Publication characteristics	Peer-reviewed journal articles, full conference papers, systematic reviews, meta-analyses and bibliometric reviews published in English between January 2015 and July 2026.	Editorials, commentaries, book reviews, news items, short abstracts, duplicate reports and non-peer-reviewed documents.
Technology	Identifiable AI, ML, DL, NLP, speech, chatbot, conversational-AI, generative-AI or LLM system.	Technology use without an AI component or superficial mention of AI.
Educational relevance	Direct connection to language teaching, learning, feedback, assessment, multilingual education, teacher support or academic integrity.	General educational administration, admissions, dropout prediction or unrelated learning analytics.
Language relevance	Writing, speaking, pronunciation, listening, reading, vocabulary, grammar, translation, literacy or broader language-learning activity.	Pure NLP benchmarking without pedagogical, learner, teacher, assessment or human-language relevance.
Evidence type	Empirical, technical validation, qualitative, mixed-method, review or meta-analytic evidence relevant to the research questions.	Insufficiently described, peripheral or redundant records that did not support the analytical framework.

3.6 Study-Selection Process

All identified records underwent successive screening stages. First, titles, author names, publication years, and DOI data were used to check bibliographic entries for duplication. The predetermined scope and eligibility criteria were then applied to titles and abstracts. In order to verify their technological focus, language-education environment, type of evidence, and contribution to the research questions, potentially relevant studies were evaluated in further depth. There were 55 studies in the final evidence base. Of these, 15 were categorized as systematic reviews, meta-analyses, or bibliometric syntheses, and 40 as primary empirical or technical research. The analysis of technology, applications, study designs, participant profiles, pedagogical integration, educational outcomes, methodological quality, and ethical concerns was based on the primary studies. In order to construct the existing review landscape, compare previous classifications, and contextualize the conclusions of the primary evidence, the secondary studies were analyzed independently. At the eligibility stage, choices were made based on contribution, methodological clarity, adequacy of presented evidence, and direct connection to the review topics. Maintaining analytical breadth in writing, speaking, listening, reading, assessment, vocabulary, teacher assistance, multilingual education, and academic-integrity research was another goal of the selection process.

3.7 Data Extraction and Coding

A systematic literature matrix was created to ensure consistent extraction and comparison. Author, year, title, journal or conference, publication type, DOI, and source details were among the bibliographic variables. AI category, technical generation, tool or model, modality, adaptive or generative capabilities, and the role assigned to the AI system were among the technological variables (Huang *et al.,* 2023; Yang and Kyun, 2022; Liang *et al.*, 2024). Language proficiency, educational application, instructional setting, teacher engagement, student role, feedback arrangement, and the type of human–AI interaction were among the pedagogical variables. Country or region, educational attainment, participant group, target language, multilingual environment, and language-resource status were examples of contextual factors (Huang *et al.,* 2023; Yang and Kyun, 2022; Liang *et al.,* 2024). Research design, sample or corpus, length of intervention, comparative condition, theoretical framework, data-collection tools, analytical techniques, and type of outcome evidence were examples of methodological variables. Linguistic and educational outcomes, affective outcomes, behavioral and learning-process outcomes, teacher-related outcomes, and technical or assessment results were the categories into which outcome variables were divided. Hallucinations, dependability, bias, fairness, privacy, academic integrity, authorship, transparency, explainability, student data protection, human oversight, digital inequality, accessibility, and institutional policy were all covered by ethical and governance coding. To guarantee that emerging themes could be consistently compared across technology generations, coding categories were iteratively improved (Huang *et al.,* 2023; Yang and Kyun, 2022; Liang *et al.,* 2024).

3.8 Quality Assessment

The methodological quality of technical AI research and empirical educational research was evaluated independently. The Mixed Methods Appraisal Tool informed the design-sensitive criteria used to appraise educational studies. The evaluation took into account the research question's clarity, design appropriateness, sampling adequacy, measurement validity, outcome data completeness, analytical procedures' suitability, coherence between the evidence and conclusions, and handling of potential bias or confounding (Hong *et al.,* 2018). Task formulation, dataset provenance, dataset size, model identification, training and evaluation methods, baseline comparison, human evaluation, external validation, error analysis, reproducibility, and ethical reporting were considered when evaluating technical AI studies. The model version, prompt design, system settings, repeated-run stability, comparison with human or non-LLM baselines, and assessment of hallucination or inconsistency were also evaluated in studies utilizing LLMs. Quality evaluation was performed to identify recurrent methodological limitations throughout the body of evidence and to gauge the degree of confidence in individual findings. A single quality problem was not, by itself, a reason for excluding studies. Rather, limitations were taken into account during the synthesis process, especially when interpreting assertions about causality, generalizability, dependability, and educational efficacy.

3.9 Data Synthesis

The synthesis integrated comparative, historical, thematic, and descriptive analyses. Publication years, evidence types, technology stages, study designs, educational levels, language proficiency, pedagogical applications, and outcome categories were all summarized using descriptive frequencies. These variables were treated as characteristics of the included evidence base rather than as measures of global research productivity. Descriptive, thematic, comparative, and framework-building techniques were employed in the synthesis. Publication year, evidence type, study design, educational level, language proficiency, target language, technology category, pedagogical application, and outcome type were all included in the descriptive analysis. Therefore, the stated frequencies should not be interpreted as a census of research activity worldwide, as they only reflect the chosen studies. Five variables were used by the thematic synthesis to categorize technologies: computational architecture, degree of generativity, style of interaction, educational role, and chronological period. Conceptual overlap across categories like deep learning, conversational AI, multimodal systems, and LLMs was lessened because of this structure. Applications were coded as tutoring and personalisation, writing support, automated feedback, assessment, speaking and pronunciation, reading and listening, vocabulary and grammar, translation, teacher support, and academic-integrity management. A single study could contribute to more than one application or outcome category. Tables and figures therefore indicate whether categories are mutually exclusive and identify the relevant denominator (Rahman *et al.,* 2024; Liang *et al.,* 2024; Wu, 2024).

Reported outcomes were separated into technical performance, objectively measured learning outcomes, learner perceptions, affective outcomes, behavioural outcomes, teacher outcomes, and assessment outcomes. Greater interpretive weight was assigned to studies with appropriate comparators, validated measures, transparent reporting, and direct educational outcomes. Technical accuracy, human–model agreement, satisfaction, or intention to use were not treated as direct evidence of learning improvement. Primary and secondary evidence were synthesised separately. The 40 primary studies supplied the main evidence for educational outcomes, technical performance, methodological quality, multilingual representation, and governance concerns. The 15 secondary papers were used to clarify the contribution of the current research, locate previous classifications, and compare published findings. The framework was created using a hybrid deductive-inductive methodology. The data-extraction framework and the study questions provided the original categories. Inductive analysis was then used to identify recurring findings on learner agency, teacher mediation, contextual fit, reliability, verification, equity, and governance. Higher-order dimensions were created by combining related codes. During interpretation, conflicting and unfavourable results were also kept. Instead of being presented as established causal pathways, the associations that emerged are given as propositions derived from the synthesis.

4. Descriptive Profile of the Evidence Base

4.1 Publication Distribution over Time

Forty primary empirical or technical studies and fifteen secondary evidence syntheses published between 2017 and July 2026 made up the final corpus. They were distributed unevenly over time. Three secondary reviews were published in 2022, one included study was published in 2017, and 51 papers were published from 2023 onward. Between 2018 and 2021, no primary study was published in the final corpus. This pattern indicates that the review represents more strongly the generative-AI period than the previous phases of AI-assisted language instruction. As a result, an equally weighted empirical comparison between consecutive technology generations is not supported by the corpus. Earlier automated writing systems, machine-learning applications, deep-learning methods, and task-oriented chatbots are used mainly to establish conceptual and historical continuity. Claims about sequential development are accordingly restricted to changes in research emphasis within the selected evidence base. The post-2022 literature focused primarily on ChatGPT and other LLM-based applications for writing support, feedback, speaking practice, teacher assistance, assessment, academic integrity, and learner perceptions. Figure 2 presents a conceptual technological timeline

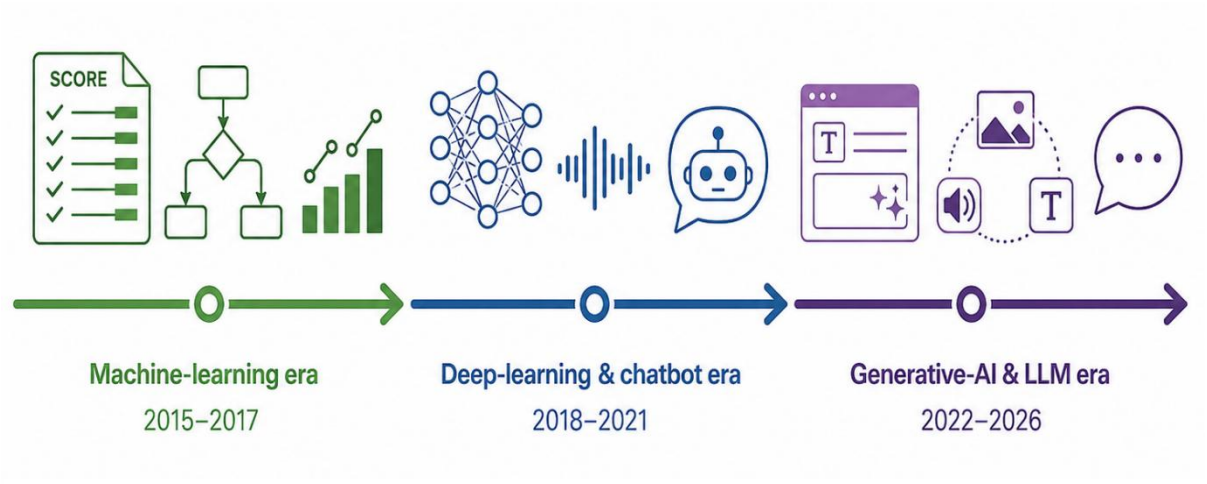

Figure 2. Timeline for Principal Technological generation of AI.

4.2 Evidence-Type Distribution

The 55 papers included 15 secondary studies and 40 primary empirical or technical investigations. Experimental and quasi-experimental studies, mixed-method research, qualitative and perception-based studies, classroom interventions, technology-acceptance studies, system-development evaluations, and technical validation studies were included among the main sources of evidence. Systematic reviews, meta-analyses, bibliometric reviews, and integrated review studies made up the secondary evidence (Albedah, 2025; Wu, 2024; Rahman *et al.,* 2024).

The empirical data varied in terms of methodology. Language achievement, speaking performance, writing quality, listening comprehension, learner motivation, anxiety, and engagement were mostly assessed using controlled and quasi-experimental approaches. Learner and instructor experiences with ChatGPT, chatbots, and AI-supported feedback were often investigated through qualitative research. Mixed-method research integrated learner perceptions, surveys, interviews, and performance measurements. Technical studies evaluated system dependability, human-AI agreement, personalized question generation, automated essay assessment, and feedback provision. The review was contextualised, prior classifications were compared, and recurrent methodological and thematic gaps were found by analyzing the 15 secondary research independently. When analyzing participant outcomes or intervention effects, they were not integrated with the primary research (Albedah, 2025; Wu, 2024; Rahman *et al.,* 2024; Uygun and Demir, 2026).

4.3 Journal and Disciplinary Distribution

AI-supported language education is placed at the nexus of educational technology, applied linguistics, computer-assisted language learning, language evaluation, psychology, and general education research, as evidenced by the distribution of the chosen publications among 30 journals. Computers and Education: Artificial Intelligence was the most frequently represented source, with seven publications. Five studies were contributed by Education and Information Technologies. Computer Assisted Language Learning and Language Learning & Technology each supplied three papers, whereas Interactive Learning Environments, System, and Humanities and Social Sciences Communications each contributed four (Huang *et al.,* 2023; Liang *et al.,* 2024; Rahman *et al.,* 2024). A significant portion of the evidence was published in educational technology journal, especially in research on chatbots, personalized learning, generative AI, and teacher adoption. Second-language learning, writing, speaking, and pedagogical integration were the primary topics of journals in applied linguistics and language education. CALL-focused publications addressed learner experience, chatbot design, technology-mediated engagement, and classroom application. Evidence on automated scoring, assessment anxiety, reliability, and feedback validity was provided by language-assessment journals. Motivation, self-regulated learning, anxiety, engagement, social presence, and acceptance of technology were all studied in psychology publications. In addition to reflecting the field's dual technical and educational nature, this multidisciplinary distribution adds to the diversity of study strategies, terminology, and outcome measures (Huang *et al.,* 2023; Liang *et al.,* 2024; Rahman *et al.,* 2024).

4.4 Educational Settings and Learner Groups

The included studies covered primary, secondary, tertiary, teacher education, and professional learning contexts, with tertiary education being the most commonly represented setting. Research on academic writing, automatic feedback, ChatGPT acceptance, language evaluation, self-regulated learning, and learner perceptions frequently featured university students. Studies including elementary EFL learners, sixth-grade pupils, young learners, and lower-secondary speaking practice were included in the less frequent primary and lower-secondary contexts (Albedah, 2025; Sharadgah and Sa'di, 2022; Lo *et al.*, 2024). Teacher-focused research looked at faculty attitudes, teacher readiness, teacher-AI collaboration, content development, feedback techniques, innovation, and professional development. Research on teacher education focused on how teacher educators and beginning language instructors responded to generative AI. Conventional classrooms, fully online settings, blended learning, mobile-assisted learning, platform-based study, after-class assignments, and AI-supported individual practice were among the learning environments. The variety of settings suggests that AI is being used both as an informal learning tool and as a formal teaching component. However, judgments about sustained use throughout school systems and younger learner populations are limited by the preponderance of short-term university based studies (Albedah, 2025; Sharadgah and Sa'di, 2022; Lo *et al.,* 2024).

4.5 Language-Skill Distribution

The distribution of the evidence base across language skills was uneven. Thirteen publications covered a wide range of language-learning applications or multiple skills. Automated writing evaluation, argumentative writing, peer feedback, academic writing, essay scoring, teacher feedback, and ChatGPT-supported revision comprised the largest single application cluster. The next important aspect was speaking, which included speaking anxiety, pronunciation-related interaction, willingness to communicate, conversational practice, and fluency (Albedah, 2025;

Jeon *et al.,* 2023; Huang *et al.,* 2023; Du and Daniel, 2024). Only two of the chosen studies addressed listening, and one study specifically addressed reading comprehension. Grammar was frequently included as part of larger language-learning outcomes rather than as a separate research focus, whereas vocabulary emerged primarily through dynamic assessment and individualized practice. Automated scoring, written evaluation, feedback validity, dependability, and learner reactions to AI-assisted assessment were all studied. This distribution demonstrates that speaking and writing have received significantly more attention than listening and reading (Albedah, 2025; Jeon *et al.,* 2023; Huang *et al.,* 2023; Du and Daniel, 2024).

4.6 Geographic Distribution

The evidence was concentrated in Asian EFL and English-learning environments when study settings could be clearly identified. Studies involving students and educators from Saudi Arabia, China, Vietnam, Iran, and other Asian educational settings were included in the collection. Many primary studies were carried out in a particular country or institutional setting. Other publications presented evidence from global or multinational reviews. In research on academic writing, EFL, ESL, and English-medium education, English was the most frequently used target language. The results cannot easily be applied to low-resource languages, multilingual school systems, or areas like Africa, Latin America, and portions of South Asia because the research revealed little geographic and linguistic variety. Therefore, rather than being a measure of institutional production, the regional distribution should be seen as a concentration of research activity (Albedah, 2025; Sharadgah and Sa'di, 2022; Wu, 2024).

4.7 Interpretation of Evidence Strength

A straightforward sequence in which one generation of AI replaces another is not supported by the evidence currently available. Deep learning architectures, task-specific machine learning systems, automated assessment tools, conversational agents, speech-enabled apps, generative models, and educational platforms continue to coexist. Additionally, there is conceptual overlap between these groups. Deep learning mainly describes a computational architecture. Generative AI describes a kind of output capability, whereas conversational AI refers to a mode of interaction. An LLM can function as a content generator, writing assistant, tutor, assessor, or conversation partner. Thus, computational architecture, generativity, interaction mode, instructional role, and chronological evolution are all separated by the taxonomy. These technologies are not viewed as generations that are mutually exclusive.

5. Taxonomy of AI Technologies

The selected evidence shows that, as opposed to a straightforward replacement sequence, AI in language instruction has evolved through multiple partly overlapping technological phases. Because they have distinct pedagogical goals, conversational agents, generative models, deep learning architectures, traditional machine learning systems, and AI-supported educational platforms continue to coexist. More interaction and content-generation possibilities are offered by more modern LLM-based systems. They also bring up increased output variability, decreased transparency, and more complicated governance issues. In contrast, earlier systems often performed narrowly defined tasks with more consistent and reliable results. Thus, the taxonomy described here categorizes technologies based on their operational implications, instructional roles, and computational features (Huang *et al.,* 2023; Yang and Kyun, 2022; Rahman *et al.,* 2024; Liang *et al.,* 2024).

5.1 Conventional Machine Learning

The early analytical and evaluation layer of AI-supported language instruction was made up of conventional machine-learning programs. In order to forecast learner performance, classify texts, score written responses, detect errors, or customize learning sequences, these systems usually depended on supervised classification, regression, decision rules, or statistical feature extraction. Task specificity was their main advantage. Such systems could generate comparatively consistent outputs within that area after being trained and validated for a specific purpose (Alam *et al.,* 2023; Bai and Hu, 2017). One of the most well-known uses was automated writing evaluation. Previous methods produced scores or corrective feedback by analyzing lexical, grammatical, syntactic, and discourse features. Research on automated writing evaluation revealed that students could utilize machine-generated feedback to edit their writing, but the input's value varied according to its precision, clarity, and compatibility with learning goals. Feedback that is flawed or too general may cause students to disregard, misunderstand, or blindly accept system

recommendations. Additionally, learner classification, vocabulary diagnosis, language performance prediction, and early forms of adaptive learning were all done using conventional systems (Alam *et al.,* 2023; Bai and Hu, 2017). Although these systems were limited by predetermined features, fixed task boundaries, and limited responsiveness to unexpected learner input, they generally provided higher procedural stability than open-ended generative models. As a result, their instructional impact was greatest when the model was combined with instructor assistance and the goal construct was precisely defined.

5.2 Deep-Learning and Transformer Systems

Deep-learning systems have increased the ability of language technologies to process complex linguistic patterns. More advanced text categorization, speech recognition, machine translation, semantic analysis, and language production were made possible by neural networks, recurrent neural networks, long short-term memory models, and later transformer architectures. Neural translation, automated pronunciation analysis, speech transcription, context-sensitive error detection, and adaptive feedback were all assisted by these techniques in language training (Xiao, 2025; Wang *et al.,* 2024; Jeon *et al.,* 2023). By modeling relationships across longer sequences, transformer-based systems such as BERT-type architectures, improved contextual language representation. They were therefore appropriate for educational NLP tasks like reading-comprehension analysis, automated scoring, semantic similarity, and feedback generation. Transformer models could learn language representations directly from large corpora and lessen reliance on manually created linguistic features, in contrast to conventional feature-based systems (Xiao, 2025; Wang *et al.,* 2024; Jeon *et al.,* 2023). Another significant use of deep learning is speech recognition. It provided automated pronunciation feedback, conversational practice, speech-based assessment, and listening support. Additionally, neural machine translation has become more significant in multilingual classrooms and scholarly writing. However, the quality of training data, pronunciation variety, accent identification, and the degree of language coverage continued to determine how effective it was as a teaching tool. High-resource languages and standardized variants often performed better than low-resource languages, regional accents, or code-switched speech.

5.3 Conversational AI and Task-Oriented Chatbots

A more interactive approach to language learning was developed by conversational AI. Early educational chatbots were task-oriented, rule-based, scripted, or retrieval-based. Their dialogue formats were frequently created with particular educational objectives in mind, such as dynamic evaluation, speaking rehearsal, vocabulary development, or grammatical exercises. For students who struggle with speech anxiety, task-oriented systems could offer repeated opportunities for interaction without the social pressure that comes with human communication (Du and Daniel, 2024; Jeon *et al.,* 2023; Fathi *et al.,* 2024; Huang *et al.,* 2022; Annamalai *et al.,* 2023). Chatbots for speaking practice, vocabulary diagnosis, writing support, and classroom engagement were among the reviewed studies. These features were expanded by speech-enabled chatbots, which let students practice spoken conversation and receive immediate responses. In certain instances, the instructor was still in charge of task design, monitoring, and interpretation when conversational agents were integrated into teacher-student-AI interaction frameworks (Ericsson and Johansson, 2023; Hsu *et al.,* 2023; Jeon *et al.,* 2023; Fathi *et al.,* 2024). Controlled interaction was the main benefit of task-oriented chatbots. Their results matched particular learning objectives and were reasonably predictable. Restricted discourse flexibility, repeated interaction patterns, little sensitivity to learner intent, and trouble sustaining authentic dialogue were some of their drawbacks. These limitations distinguish task-oriented chatbots from generative conversational systems, which offer greater linguistic flexibility but less predictable outputs.

5.4 Generative AI and Large Language Models

Generative AI and LLMs marked a major shift from narrowly defined educational automation to open-ended language generation. ChatGPT and related GPT-family systems emerged as the most prevalent technology in the post-2022 evidence base. The examined research also included other commercial models, such as Claude and PaLM. According to Albedah (2025), Guo and Wang (2024), Escalante *et al.* (2023), Lo *et al.* (2024), and others, these systems provided writing support, feedback generation, essay scoring, lesson planning, conversational tutoring, translation, paraphrasing, question generation, reading-comprehension support, and teacher professional development. In contrast to previous systems, LLMs are capable of simulating conversation, responding to a variety

of prompts, offering comprehensive context-sensitive explanations, and carrying out many linguistic activities via a single interface. This adaptability allows for more extensive educational usage. Without switching between specialized tools, learners can ask for clarification, examples, revisions, role-play, or customized feedback (Pack *et al.,* 2024; Li *et al.,* 2024; Guo and Wang, 2024; Escalante *et al.,* 2023). But the same adaptability also leads to a great deal of uncertainty. Prompts, model versions, sessions, and decoding parameters can all affect LLM outcomes. These algorithms may generate examples that are culturally inappropriate, inconsistent scores, content that is convincing but false, or feedback that is too advanced for the learner. In a number of instances, technical research on LLM based essay evaluation showed encouraging concordance with human raters. Nevertheless, dependability varied between models and repeated attempts at scoring. Therefore, rather than being viewed as reliable expert systems, LLMs should be viewed as probabilistic educational tools.

5.5 Multimodal and Embodied AI

Language acquisition is extended beyond text-based interaction using multimodal and embodied technologies. Applications for speech recognition integrate automated transcription, pronunciation analysis, and feedback with spoken input. Younger learners may find social robots' physically embodied conversational interfaces especially interesting. Multimodal systems can promote vocabulary development, spoken communication, reading comprehension, and contextualized training by integrating text, speech, visuals, and audio (Xiao, 2025; Lin *et al.,* 2024). These systems have educational potential because of their ability to imitate authentic communicative circumstances while also providing greater sensory connection. However, hardware access, classroom infrastructure, cost, technical upkeep, and teacher, student, and parent approval all affect their use. Novelty effects should not be confused with long-term educational efficacy, even though embodied systems may boost engagement (Xiao, 2025; Lin *et al.,* 2024).

5.6 AI-Supported Educational Infrastructure

AI in language teaching serves as infrastructure rather than a visible tutor or chatbot. The organization and monitoring of instruction are shaped by adaptive platforms, automated evaluation systems, learning analytics, personalized recommendation engines, and teacher-support technologies. These systems can create practice materials, recognize mistakes, predict learner needs, sequence activities, and give teachers dashboards (Guo *et al.,* 2024b; Cao and Phongsatha, 2025; Ma and Chen, 2024; Wang *et al.,* 2024). The evidence base contained examples of personalized reading comprehension systems, AI-supported peer feedback, automated essay grading, adaptive English-learning platforms, and systems that coupled learner modeling with generated questions. Lesson planning, feedback creation, content creation, and professional development were examples of teacher-support software. These systems can minimize regular effort, but their effectiveness is dependent on integration with curriculum, assessment standards, and teacher judgment (Ma and Chen, 2024; Wang *et al.,* 2024; Wei, 2023; Guo *et al.,* 2024b).

5.7 Cross-Cutting Technological Properties

The technologies varied in a number of cross-cutting aspects. While LLMs offered better adaptability and task breadth, conventional systems often gave greater stability, transparency, and repeatability within specific tasks. Commercial systems frequently offered greater overall performance and simpler access, but they were less transparent. More room for inspection, localization, and institutional control was made possible by open-source alternatives (Pack *et al.,* 2024; Wu, 2024; Albedah, 2025; Bao *et al.,* 2025). Language coverage was also significantly increased by Transformer and LLM systems. However, equivalent performance across languages was not ensured by support for several languages. Stronger support was nevertheless given to high-resource languages. Because model versions and system behavior could change over time, reproducibility was also more challenging for proprietary LLMs (Pack *et al.,* 2024; Wu, 2024; Albedah, 2025; Bao *et al.,* 2025). The cost ranged from low cost text based tools to speech, multimodal, and embodied systems that required a lot of resources. Human monitoring was still crucial in every category. Teachers have to preserve learner data, prevent inappropriate dependence, validate outputs, analyze assessment results, and integrate technology with learning objectives.

The technical taxonomy should not assume that newer systems automatically replace older ones. The balance of capability, control, dependability, cost, and pedagogical risk varies with each generation. The primary

technology categories and their representative applications are compiled in Table 3. Their evolution over several, somewhat overlapping AI generations is depicted in Figure 3.

Table 3. AI technology taxonomy and representative language-education applications

Technology category	Defining characteristics	Representative applications	Reference	Principal strength	Principal limitation
Conventional ML and automated systems	Narrow, task-trained classification, prediction or scoring with structured outputs.	Automated writing evaluation, scoring, feedback uptake and diagnostic assessment.	Bai and Hu (2017); Alam *et al.* (2023)	Stability and task control.	Limited contextual flexibility and open-ended explanation.
Conversational AI and task-oriented chatbots	Scripted, retrieval-based or bounded dialogue systems; some speech-enabled.	Speaking practice, dynamic assessment, vocabulary support, classroom interaction and motivation.	Fathi *et al.* (2024); Ericsson and Johansson (2023); Ji *et al.* (2024); Kwon *et al.* (2023); Annamalai *et al.* (2023)	Repeated low-pressure interaction.	Restricted dialogue range and pragmatic sensitivity.
Generative AI and LLMs	Probabilistic, prompt-driven generation across multiple language tasks.	Writing feedback, tutoring, scoring, academic writing, lesson support and teacher development.	Escalante *et al.* (2023); Guo and Wang (2024); Guo, Pan, *et al.* (2024); Karataş *et al.* (2024); Cong-Lem *et al.* (2024); Al-khresheh (2024); Li *et al.* (2024a); Mohamed (2024); Teng (2024); Wang *et al.* (2024)	Broad task coverage and contextual response.	Hallucination, variability, opacity and authorship risk.
Multimodal, speech and embodied AI	Integration of speech, audio, text or physical embodiment.	Listening, pronunciation, oral practice and social-robot-supported learning.	Lin *et al.* (2024); Xiao (2025)	Richer communicative interaction.	Infrastructure cost, recognition bias and limited validation.
AI-supported platforms and applications	Adaptive platforms, mobile applications and system-level learning support.	Personalisation, blended learning, engagement, reading support and self-regulation.	Cao and Phongsatha (2025); Ma and Chen (2024); Mingyan *et al.* (2025); Wei (2023)	Integration with ongoing learning activity.	Platform dependence and uncertain transfer.

6. Application and Pedagogical Taxonomy

The reviewed research suggests that has been evaluated suggests that pedagogical roles, rather than just technology labels, are the best way to understand AI applications in language instruction. Depending on the instructional design, the same technology system can function as a tutor, feedback provider, assessor, conversational partner, content creator, or decision-support tool. On the other hand, large language models, task-oriented chatbots, adaptive platforms, and traditional machine-learning algorithms can all provide comparable educational functions.

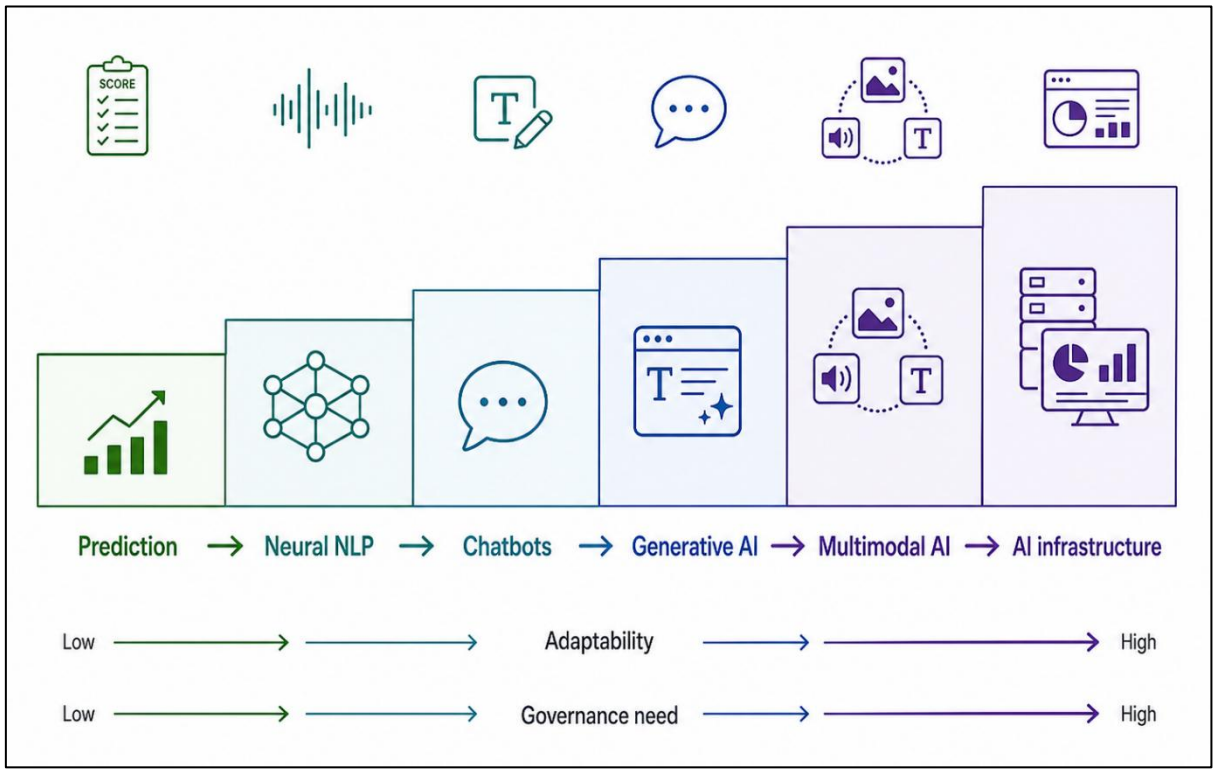

Figure 3. Technology taxonomy across AI generations.

Thus, the taxonomy created in this section connects technology capabilities to measured outcomes, learner agency, pedagogical function, teacher involvement, and implementation constraints.

6.1 Intelligent Tutoring and Personalised Learning

AI-powered tutoring and personalization systems were created to adjust task difficulty, pace, feedback, and instructional content to the needs of learners. While more recent systems use conversational AI and generative models to deliver context-sensitive explanations and practice exercises, earlier systems depended on learner modeling, rule-based sequencing, or performance prediction. This area includes AI-mediated language applications, vocabulary-diagnosis tools, adaptive English-learning platforms, and personalized reading systems (Jeon, 2023; Wang *et al.,* 2024; Ma and Chen, 2024; Cao and Phongsatha, 2025). The primary instructional function of these systems was to give scaffolded practice and instant support. By choosing assignments, repeating exercises, asking for clarifications, or moving at their own pace, students demonstrated moderate to high agency. In classroom-based implementations, teacher engagement ranged from active orchestration to minimal supervision in self-directed systems. Language proficiency, learner engagement, self-regulated learning, vocabulary growth, and perceived personalization were among the reported results. The main drawback was that algorithmic adaptation frequently relied more on short-term success metrics than on a thorough comprehension of learner needs. Therefore, if the system changed difficulty without taking into consideration motivation, culture, past knowledge, or instructional aims, personalization might become shallow (Ma and Chen, 2024; Wei, 2023; Jeon, 2023; Wang *et al.,* 2024).

6.2 Writing Assistance and Automated Feedback

The application area with the highest representation was writing assistance. Grammar mistakes were found, corrections were suggested, feedback was generated, argumentative writing was supported, and academic

production was aided by commercial writing tools, chatbots, LLMs, and conventional automated writing-evaluation systems. The technologies included open-ended generative feedback via ChatGPT and associated systems, as well as feature-based automated evaluation (Guo and Wang, 2024; Guo, Li, *et al.,* 2024a; Guo, Pan, *et al.,* 2024b; Escalante *et al.,* 2023). AI has played a variety of pedagogical roles, from writing collaborator to error detector. The system offered learners the option to accept, reject, or edit the corrective feedback in some studies. In others, AI facilitated iterative drafting, teacher feedback, or peer feedback. Because instructional value depended on whether students critically assessed recommendations rather than just adopting them, learner agency was crucial. In order to clarify inaccurate recommendations, explain feedback, and link it with evaluation standards, teacher engagement remained crucial (Kwon *et al.,* 2023; Bai and Hu, 2017; Guo, Li, *et al.,* 2024a; Guo, Pan, *et al.,* 2024b). Measured outcomes included writing quality, revision behaviour, feedback uptake, argumentative-writing performance, learner preference, and perceived usefulness. Positive effects on writing and feedback quality were documented in a number of trials, albeit these benefits varied. Automated feedback could be inaccurate, generic, overly corrective, or insufficiently sensitive to rhetorical purpose. Concerns like authorship, reliance, and the line separating acceptable help from replacement of student work were also raised by generative systems.

6.3 Automated Assessment and Essay Scoring

AI-supported assessment covered automated essay scoring, dynamic vocabulary assessment, writing evaluation, and comparisons between AI-generated scores and human judgements. Traditional scoring systems often used learned prediction models and preset linguistic features. In contrast, LLM-based systems used prompting to provide overall scores or ratings based on rubrics (Jeon, 2023; Pack *et al.,* 2024; Biju *et al.*, 2024; Li *et al.,* 2024). These systems were used for both assessment and diagnosis. They could highlight learning challenges, facilitate formative assessment, and offer quick scoring. Because automated judgments needed to be interpreted and validated, particularly in high-stakes situations, teacher involvement remained crucial. Compared to summative assessment, learner agency was typically higher in formative use, where students may utilize the results to revise their work (Jeon, 2023; Pack *et al.,* 2024; Biju *et al.,* 2024; Li *et al.,* 2024). Scoring accuracy, agreement with human raters, intrarater and interrater reliability, anxiety, motivation, writing performance, and acceptance of AI-supported assessment were among the results that were presented. Promising outcomes for LLM based scoring were reported in a number of studies, especially with more sophisticated models. Reliability, however, differed between scoring attempts and systems. This volatility is still a significant drawback. When a system generates varying scores between sessions or model versions, it cannot be regarded as being on par with a verified human assessment procedure. Therefore, transparent rubrics, frequent validation, human moderation, and explicit limitations on high-stakes use are necessary for automated assessment.

6.4 Speaking, Pronunciation, And Conversational Practice

The second main application cluster was speaking and conversational practice. Task-oriented chatbots, voice-enabled systems, mobile applications, and generative-AI chatbots were employed to mimic conversation, provide opportunities for repeated speaking practice, and lessen the social strain that comes with in-person interactions. While some systems allowed for open-ended dialogue, others offered structured speech activities (Ericsson and Johansson, 2023; Hsu *et al.,* 2023; Huang *et al.,* 2024; Fathi *et al.,* 2024). AI was mostly used in education as an interaction facilitator, practice tutor, or conversational partner. Because students were in charge of the timing, repetition, and content of practice, learner agency was typically high. From creating assignments and keeping an eye on interactions to incorporating chatbot activities into group or classroom-based learning, teachers were involved in a variety of ways (Hsu *et al.,* 2023; Huang *et al.,* 2024; Ji *et al.,* 2024; Jeon, 2024). Speaking performance, fluency, readiness to communicate, enjoyment, anxiety, confidence, and learner experience were among the outcomes assessed. Speaking skills improved and anxiety decreased in a number of studies, particularly when AI engagement was paired with structured educational exercises. Open ended interaction did not, however, consistently enhance authentic communication, pragmatic competence, or pronunciation accuracy. Major problems continued to include poor conversational coherence, limited sensitivity to accents, speech recognition failures, and inadequate corrective feedback.

6.5 Reading and Listening Support

Writing and speaking were significantly more common than reading and listening. Personalized reading comprehension systems, generated comprehension questions, adaptive difficulty selection, chatbot-assisted listening practice, and speech recognition-based listening treatments were among the AI uses in these domains (Xiao, 2025; Huang *et al.,* 2024; Wang *et al.,* 2024). AI served as an adaptive assistance system, comprehensive tutor, or question generator in reading apps. While teachers or experts continued to play a role in validating generated content, learners were given texts and questions that matched their projected proficiency. Learner engagement, personalization, reading comprehension performance, and question quality were among the reported results. One major drawback was that improved reading achievement was not always demonstrated by technical validation of created questions (Xiao, 2025; Huang *et al.,* 2024; Wang *et al.,* 2024). In order to improve decoding and comprehension, listening apps used chatbots, speech systems, or AI-powered tasks. While teacher involvement was frequently restricted to task design and supervision, learners exerted agency through repeated listening and interaction. Listening comprehension, flow, anxiousness, and involvement were among the reported results. However, the limited number of studies restricts generalizations about efficacy across listening genres and ability levels.

6.6 Vocabulary and Grammar Learning

Applications for grammar and vocabulary were frequently integrated into writing assistance, chatbot, and tutoring programs rather than being considered separate technologies. In response to learner responses, dynamic assessment chatbots diagnosed vocabulary knowledge and adjusted prompts or support. Grammar corrections, examples, explanations, and help with paraphrase were offered via writing tools and generative systems (Alam *et al.,* 2023; Bai and Hu, 2017; Jeon, 2023; Kwon *et al.,* 2023). AI had a diagnostic, remedial, and explanatory function in education. The degree to which students actively understood explanations or merely accepted corrections differed depending on their level of learner agency. Teachers were still required to differentiate between appropriate language use and superficial grammatical perfection. Reported outcomes included vocabulary growth, improved grammatical accuracy, diagnostic precision, and greater learner confidence. However, these systems often placed more emphasis on sentence-level correctness than on contextual, pragmatic, or discourse-level appropriateness (Jeon, 2023; Kwon *et al.,* 2023; Bai and Hu, 2017; Alam *et al.,* 2023).

6.7 Translation and Multilingual Support

Machine translation, multilingual chatbots, and LLMs were employed for translation, paraphrase, cross-linguistic explanation, and educational resource access. These tools could support multilingual learners by delivering real-time translations, short explanations, and cross-language comparisons. Their pedagogical role ranged from translation help to more general language scaffolding (Albedah, 2025; Li *et al.,* 2023). Learner agency was generally high because students could request alternatives, compare different outcomes, and employ translation to aid comprehension. Teacher engagement was still required to avoid overdependence and ensure that translation aided learning rather than replaced meaningful language use. The primary reported outcomes were better access, comprehension, writing support, and learner perceptions. However, performance was uneven between languages. Culturally distinct meanings, dialects, and low-resource languages received less consistent support, but multilingual ability was often stronger for high-resource languages (Albedah, 2025; Li *et al.,* 2023).

6.8 Teacher Planning and Content Generation

Generative AI enhanced teacher-facing applications by assisting with lesson preparation, question generation, worksheet building, examples, assessment items, feedback templates, and differentiated resources. Instead of acting as a direct tutor for students, AI served as a content creator or planning assistant in these applications (Al-khresheh, 2024; Moorhouse, 2024; Moorhouse and Kohnke, 2024; Mohamed, 2024). Because generated materials needed to be chosen, modified, and validated, teacher agency was crucial. Although the technology could expand the variety of resources available and shorten preparation times, its usefulness relied on pedagogical knowledge. Perceived efficiency, inventiveness, creativity, and a lighter burden were among the reported results. The main drawback was the potential for generated content to be erroneous, inadequately curriculum-aligned, culturally incorrect, or targeted at an inappropriate proficiency level (Moorhouse, 2024; Moorhouse and Kohnke, 2024; Nugroho *et al.,* 2024; Al-khresheh, 2024).

6.9 Teacher Professional Development and Readiness

Additionally, teacher professional development both examined and used AI. Research looked at teacher preparedness, attitudes, adoption, cooperation with AI, and the incorporation of generative systems into language-teacher training. AI-supported teaching platforms, ChatGPT, and generative-AI tools were among the technologies (Al-khresheh, 2024; Mohamed, 2024; Moorhouse, 2024; Moorhouse and Kohnke, 2024). The pedagogical role of AI was to facilitate reflection, resource building, experimentation, and professional learning. Teachers were greatly assisted by teacher educators in developing responsible-use procedures, redesigning assessments, and evaluating outputs. Readiness, confidence, inventiveness, perceived utility, and risk awareness were among the reported results. The main obstacle was the disparity in AI literacy. Although they were often aware of the potential advantages, teachers were not systematically prepared in terms of prompt design, verification, privacy, assessment integrity, and governance (Moorhouse, 2024; Moorhouse and Kohnke, 2024; Nugroho et al., 2024; Al-khresheh, 2024).

6.10 Academic-Integrity and Governance-Related Applications

Detection, disclosure guidelines, assessment redesign, authorship clarification, and institutional policy building were among the AI-related academic integrity applications. According to Cong-Lem *et al.* (2024) and Shen and Chen (2025), control of learner and instructor use was often the pertinent application, rather than the technology itself. The role of education was both preventive and regulatory. Redesigning assessments, demanding transparency, identifying appropriate aid, and upholding human monitoring fell under the purview of educators and institutions. Instead of allowing unlimited access, learner agency was articulated through appropriate use.

Academic dishonesty awareness, instructor reactions, student behaviors, and opinions of legitimate versus unacceptable AI assistance were among the results (Cong-Lem *et al.,* 2024; Shen and Chen, 2025). The primary restriction was a lack of consistent institutional standards. Policies frequently evolved more slowly than classroom procedures, leaving educators to decide for themselves what constitutes appropriate use. The application taxonomy by technology, pedagogical function, stakeholder, language proficiency, and assessed outcome is summarized in Table 4. The connections between AI capabilities, educational applications, educators, students, and institutional players are shown in Figure 4.

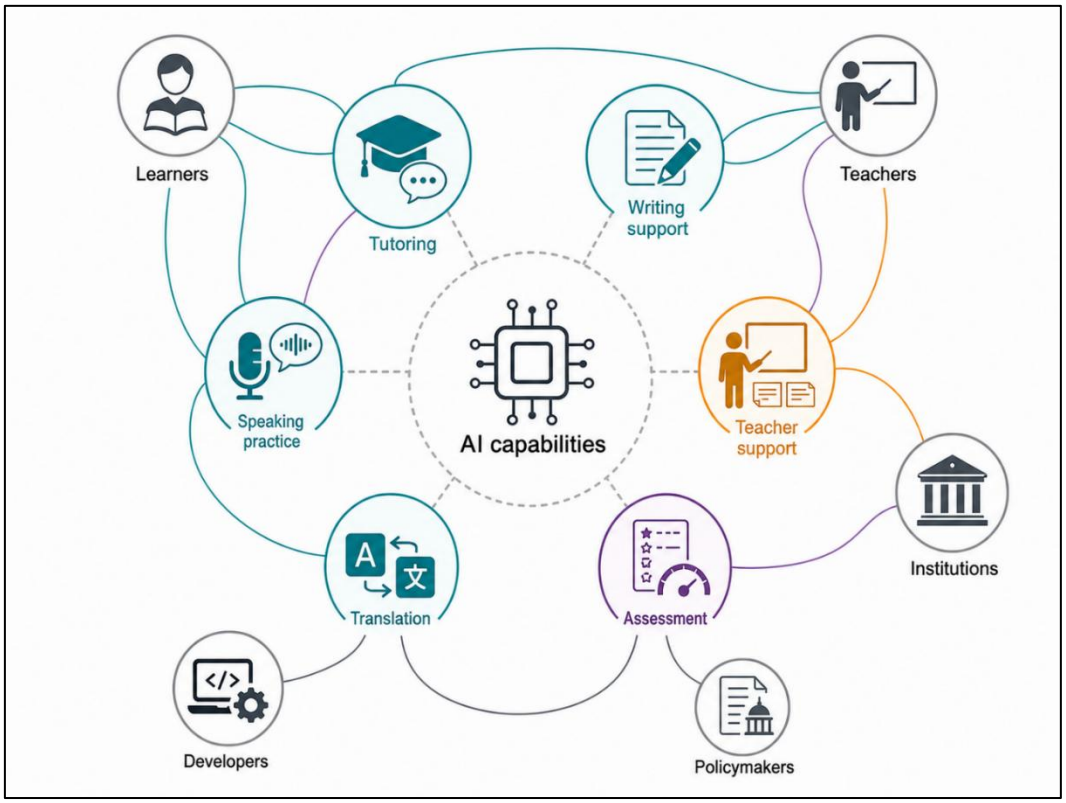

Figure 4. Stakeholder ecosystem with AI capabilities.

Table 4. Application taxonomy by pedagogical role, stakeholder and language skill

Application	Technology	Pedagogical role	Teacher involvement	Learner agency	Typical outcomes	Principal limitation
Intelligent tutoring/ personalisation	Adaptive platforms, chatbots and LLM systems.	Tutor, scaffold and recommender.	Task design, monitoring and interpretation.	Moderate to high.	Achievement, self-regulation and perceived personalisation.	Adaptation may rely on shallow performance signals.
Writing assistance/ feedback	Automated writing evaluation, Grammarly and LLMs.	Feedback provider, editor or collaborato.	Rubric alignment and verification.	High when suggestions are critically evaluated.	Writing quality, revision and feedback uptake.	Inaccuracy, overcorrection and authorship ambiguity.
Automated assessment	Scoring models and LLM evaluators.	Evaluator and diagnostic support.	Validation, moderation and appeal.	Low in summative use; higher in formative use.	Agreement, reliability, anxiety and performance.	Instability, construct misalignment and fairness risk.
Speaking/ pronunciation	Chatbots, speech AI and generative dialogue.	Conversational partner and practice tutor.	Task orchestration and corrective support.	High.	Fluency, willingness to communicate and anxiety.	Accent sensitivity and limited pragmatic feedback.
Reading/ listening	Adaptive systems, generated questions and speech tools.	Comprehension scaffold.	Material validation and task sequencing.	Moderate.	Comprehension, engagement and flow.	Sparse evidence and weak longitudinal validation.
Vocabulary/ grammar	Dynamic-assessment chatbots and writing tools.	Diagnostic and corrective support.	Explanation and contextualisation.	Moderate.	Vocabulary growth and grammatical accuracy.	Surface correctness may displace contextual learning.
Translation/ multilingual support	Neural translation and multilingual LLMs.	Cross-linguistic scaffold.	Control of dependence and cultural fit.	High.	Access, comprehension and multilingual support.	Uneven low-resource-language quality.
Teacher planning/ content generation	Generative AI and teacher-support platforms.	Assistant and content generator.	Teacher retains full authorship and validation.	Teacher-led.	Efficiency, creativity and resource production.	Curriculum misalignment and fabricated content.
Professional development/ readiness	Generative AI and AI-integrated teacher education.	Reflective assistant and professional-learning tool.	Teacher educators structure critical use.	Teacher-led.	Readiness, confidence and innovation.	Uneven AI literacy and institutional guidance.

Integrity/ governance	Disclosure systems, assessment redesign and policy mechanisms.	Decision-support and regulatory function.	Central institutional and teacher responsibility.	Responsible, disclosed use.	Integrity awareness and policy clarity.	Inconsistent standards and unreliable detection.

7. Educational, Linguistic and Stakeholder Outcomes

A wide range of linguistic, educational, affective, behavioural, teacher-related, and evaluation results were reported in the 40 primary investigations. However, depending on the study design, length of the intervention, outcome measure, and the difference between directly measured performance and participant perception, the quality of the evidence varied significantly. Technically sound systems did not always result in superior educational outcomes, and positive perceptions toward AI did not always translate into observable learning improvements. Thus, the synthesis makes a distinction between technical validation, self-reported experience, observable behaviour, and objective performance (Wu, 2024).

7.1 Writing Outcomes

The language skill that was studied the most was writing. Automated feedback, academic writing, argumentative writing, peer review, essay scoring, and learner reactions to AI-generated recommendations were all included in the analyzed studies. Grammatical accuracy, lexical choice, organization, idea generation, revision efficiency, and instant feedback were the most commonly mentioned advantages across various applications (Escalante *et al.,* 2023; Guo and Wang, 2024; Guo, Li, *et al.,* 2024a; Guo, Pan, *et al.,* 2024b). Studies on AI-supported peer and instructor feedback revealed that technology could increase feedback timeliness and availability while also facilitating iterative revisions. Students could discuss ideas, revise drafts, and seek support outside regular classroom hours. When AI was integrated into structured pedagogical activities rather than being used as an uncontrolled writing alternative, some research found improvements in writing quality and feedback quality (Guo, Li, *et al.,* 2024a; Guo, Pan, *et al.,* 2024b; Kwon *et al.,* 2023; Bai and Hu, 2017).

It became clear that revision behaviour was a crucial mediating mechanism. Students' reactions to automated feedback were not consistent. While some people automatically accepted edits or disregarded criticism they felt was incorrect, others critically assessed ideas and only included pertinent modifications. This suggests that learner proficiency, feedback literacy, task design, instructor mediation, and system quality all had an impact on feedback uptake. Learners had generally good evaluations, particularly of convenience, immediacy, and reduced anxiety during drafting. However, increased writing proficiency was not correlated with perceived utility. Many studies did not show transfer to autonomous writing and instead mostly depended on self-reported satisfaction or preference. Because it was sometimes difficult to distinguish between feedback, rewriting, and content creation, generative systems also raised authorship issues. The most reasonable conclusion is that AI can aid in the development of writing when it is utilized for formative feedback, guided revision, and explanation; however, educational benefit is less certain when the system takes the place of the learner's own planning and composition (Alam *et al.,* 2023; Steiss *et al.,* 2024; Teng, 2024; Escalante *et al.,* 2023).

7.2 Speaking and Pronunciation Outcomes

Fluency, conversational interaction, readiness to communicate, enjoyment, anxiety, and confidence were all studied in speaking research. Without the social pressure that comes with interacting with people, task-oriented chatbots, speech-enabled apps, mobile tools, and generative conversational agents offered frequent opportunities for oral practice (Fathi *et al.,* 2024; Ericsson and Johansson, 2023; Hsu *et al.,* 2023; Jeon, 2024). Speaking performance was found to improve in a number of intervention studies, especially when AI exercises were combined with collaborative learning, teacher assistance, or structured assignments. The ability to practice repeatedly, regulate the speed of engagement, and participate without worrying about being judged by their peers right away was valued by the learners. These conditions were linked to a greater readiness to communicate and reduced speaking anxiety

(Hsu *et al.,* 2023; Jeon, 2024; Ji *et al.,* 2024; Mingyan *et al.,* 2025). However, interaction quality differed between systems. Scripted chatbots provided predictable practice, but limited conversational freedom. Generative systems promoted greater discourse, but they occasionally provided linguistically incorrect responses, insufficient remedial feedback, or responses that did not match learner intent. Speech-recognition errors also affected the accuracy of pronunciation feedback, particularly for students with non-standard accents or lower proficiency. Speaking performance was assessed more consistently than pronunciation. Some studies used oral-task ratings that did not distinguish between pronunciation, intelligibility, grammar, and interactional competence. The research thus supports AI as a valuable practice partner, although it does not demonstrate equivalent performance across all dimensions of spoken proficiency. Human feedback remained critical for pragmatic appropriateness, fine-grained pronunciation, and genuine interaction (Tai and Chen, 2024; Fathi *et al.,* 2024; Ericsson and Johansson, 2023; Hsu *et al.,* 2023).

7.3 Reading and Listening Outcomes

Reading and listening were significantly underrepresented in the evidence base. The limited reading evidence focused on personalized question development, adjustable task difficulty, and comprehension support. AI systems could generate reading questions, evaluate student performance, and adjust task difficulty based on expected abilities. According to technical research, generated questions can be as challenging and high-quality as expert-developed materials (Huang *et al.,* 2024; Wang *et al.,* 2024; Xiao, 2025). Nevertheless, enhanced reading comprehension was not a direct result of system performance. In certain research, student achievement was not as important as the quality of generated content. This distinction is crucial since even a technically sound question-generation system might not increase comprehension strategies or be poorly integrated into instruction (Huang *et al.,* 2024; Wang *et al.,* 2024; Xiao, 2025). Chatbots, voice recognition software, and AI-assisted tasks were employed in listening experiments to enhance decoding, understanding, flow, and anxiety. Positive benefits on listening performance and engagement were reported by some research, especially when students could repeat assignments and get help right away. However, significant generalization is hindered by the small number of studies, diversity in task design, and inadequate longitudinal evidence. In contrast to writing and speaking, reading and listening thus constitute obvious study gaps.

7.4 Vocabulary and Grammar Outcomes

Grammar and vocabulary goals were often integrated into broader language-learning programs. Adaptive algorithms and dynamic-assessment chatbots were employed to diagnose vocabulary understanding and offer progressive assistance. Grammar explanations, corrections, examples, and help with paraphrasing were provided using generative AI and writing tools (Jeon, 2023; Kwon *et al.,* 2023; Bai and Hu, 2017; Alam *et al.,* 2023). Numerous studies have shown improvements in vocabulary and grammar proficiency, especially when students were given frequent practice and prompt feedback. Identifying knowledge gaps and adjusting subsequent prompts were made easier with the use of vocabulary tools. The quickness and accessibility of grammar support were highly appreciated, particularly while writing (Jeon, 2023; Kwon *et al.,* 2023; Bai and Hu, 2017; Alam *et al.,* 2023). The primary constraint was the inclination of AI systems to prioritize sentence-level accuracy. Grammatically correct output was not always natural in context, rhetorically effective, or pragmatically appropriate. Learners could also become dependent on automated correction without developing a sound understanding of grammar. Better outcomes were more likely when feedback required self-correction, reflection, or explanation rather than passive acceptance.

7.5 Affective and Behavioural Outcomes

Affective outcomes appeared frequently throughout the evidence base. AI-supported education was investigated in terms of motivation, engagement, enjoyment, anxiety, self-efficacy, autonomy, and self-regulated learning (Ebadi and Amini, 2024; Alotaibi *et al.,* 2025; Cao and Phongsatha, 2025; Ghafouri, 2024). Motivation and engagement were frequently increased when AI systems gave immediate responses, interactive tasks, personalized support, or possibilities for self-paced practice. Learners frequently described these tools as accessible, convenient, and nonjudgmental. Conversational systems were found to be particularly associated with increased satisfaction and decreased anxiety during speaking practice (Cao and Phongsatha, 2025; Ghafouri, 2024; Ma and Chen, 2024; Wei, 2023). Self-efficacy increased in certain experiments, particularly when students believed that AI assisted them in

practicing independently or completing challenging tasks. However, these benefits were not uniform among all participants. Learners with lower proficiency may overestimate the correctness of the generated information. Confidence gained through continuous AI support does not always translate into autonomous performance. Positive impacts on autonomy and self-regulation were more noticeable when learners chose their own tasks, tracked their progress, and strategically used feedback. In a controlled study, students using an AI-supported platform outperformed those receiving traditional education in terms of achievement, motivation, and self-regulated learning. However, when students depend on AI to generate ideas, make corrections, or finish tasks, their autonomy may also be compromised. According to the data, AI only promotes autonomy when students are still in charge of making decisions and evaluating themselves (Xiao, 2025; Ebadi and Amini, 2024; Alotaibi *et al.,* 2025; Cao and Phongsatha, 2025). Similar conditions applied to anxiety outcomes. Some learners found that AI interaction lessened their social and speaking anxiety, while concerns about incorrect feedback, evaluation surveillance, or unfamiliar technology could lead to new forms of anxiety. Task design, trust, digital competency, and the perceived repercussions of error all influenced affective benefit.

7.6 Teacher Outcomes

Teacher-related outcomes included workload, readiness, planning, trust, innovation, accountability, and professional learning. For lesson planning, resource production, feedback drafting, question creation, and differentiated content, teachers often saw generative AI as beneficial (Al-khresheh, 2024; Mohamed, 2024; Moorhouse, 2024; Moorhouse and Kohnke, 2024). Regular planning and feedback activities showed the greatest potential for workload reduction. However, the necessity to confirm accuracy, modify materials, keep an eye on student usage, and restructure exams somewhat outweighed the time saved on preparation. Thus, instead of eliminating teacher workload, AI reallocated it (Moorhouse, 2024; Moorhouse and Kohnke, 2024; Nugroho *et al.,* 2024; Al-khresheh, 2024). There were significant differences in readiness. Stronger AI literacy among educators was associated with a higher likelihood of experimenting with and pedagogically integrating new tools. Others voiced doubts regarding academic integrity, prompt design, privacy, and institutional expectations. As a result, professional development was essential for pedagogical decision-making, evaluation, governance, and technological operation. A key prerequisite was found to be trust. When AI produced results that were clear, precise, and simple to change, teachers were willing to use it. When systems delivered inconsistent assessment results, produced culturally inappropriate content, or had hallucinations, trust declined. When educators viewed AI as a collaborator rather than an independent substitute, innovation was at its peak. Nonetheless, the instructor continued to bear responsibility, especially when it came to assessment, feedback and consequential decisions (Mohamed, 2024; Moorhouse, 2024; Moorhouse and Kohnke, 2024; Nugroho *et al.,* 2024).

7.7 Assessment Outcomes

Assessment studies examined model variability, scoring consistency, validity, reliability, and agreement with human raters. While LLMs provided more flexibility in understanding open-ended responses, automated scoring methods provided speed and scalability (Biju *et al.*, 2024; Li *et al.,* 2024; Pack *et al.,* 2024). Several sophisticated models demonstrated good intrarater reliability and encouraging agreement with human raters. Performance varied, though, depending on the models, prompts, and repeated scoring instances. When proprietary systems were modified, a model that seemed trustworthy in one session could yield different scores in subsequent sessions (Biju *et al.,* 2024; Li *et al.,* 2024; Pack *et al.,* 2024). Validity also depended on the assessment task. A model could show strong agreement with human scores and still fail to represent the intended construct or treat learner groups fairly. Automated systems may reproduce cultural bias, rely on superficial linguistic features, or apply poorly aligned rubrics. Human moderation therefore remained necessary, especially in placement, certification, and summative assessment.

7.8 Mixed and Null Findings

The evidence was not uniformly positive. Several studies found only partial or skill-specific advantages. AI-assisted learning resulted in more consistent improvements in vocabulary, grammar, and writing than in speaking and listening. Some interventions increased student satisfaction or motivation without yielding comparable improvements in measured performance. Other research showed technical quality, such as accurate question generation or agreement with human scoring, but did not look into whether learning had improved (Escalante *et al.,*

2023; Karataş *et al.,* 2024; Wu, 2024; Li *et al.,* 2024). Perception-based research frequently found enthusiasm, convenience, and a willingness to adopt AI. However, these results remained susceptible to self-report bias and novelty effects. Short intervention periods also limit the ability to draw conclusions about retention and long-term development. In several situations, no discernible influence was observed for certain abilities or learner groups (Wu, 2024; Li *et al.,* 2024; Pack *et al.,* 2024; Escalante *et al.,* 2023). The overall trend indicates conditional effectiveness rather than consistent benefit across all applications and circumstances. When AI was integrated into structured instruction, guided by teachers, and matched with a specified pedagogical goal, it provided better results. When systems were employed without explicit learning objectives, when results were solely assessed by perception, or when technical proficiency was viewed as being on par with academic achievement, the benefits were diminished. The outcome patterns are summarised in Table 5. The relationships between technology, stakeholder groups, and result domains are mapped in Figure 5.

Table 5. Educational, linguistic, affective and stakeholder outcomes

Outcome domain	Reported positive pattern	Important qualification	Ref
Writing	Improved feedback availability, revision support, grammatical accuracy and writing quality in structured uses.	Benefits depended on feedback literacy, teacher mediation and independent learner engagement.	Escalante *et al.* (2023); Guo and Wang (2024); Guo, Pan, *et al.* (2024); Kwon *et al.* (2023); Bai and Hu (2017); Alam *et al.* (2023); Steiss *et al.* (2024); Teng (2024)
Speaking/pronunciation	Greater practice, fluency, willingness to communicate and reduced anxiety in some interventions.	Pronunciation, pragmatics and authentic interaction were less consistently established.	Fathi *et al.* (2024); Ericsson and Johansson (2023); Hsu *et al.* (2023); Jeon (2024); Karataş *et al.* (2024); Mingyan *et al.* (2025); Tai and Chen (2024)
Reading/listening	Evidence of personalised questions, comprehension support, listening gains and improved flow.	Very small evidence base and limited delayed outcomes.	Wang *et al.* (2024); Huang *et al.* (2024); Xiao (2025)
Vocabulary/grammar	Immediate diagnosis, correction and repeated practice supported short-term gains.	Risk of passive acceptance and sentence-level emphasis.	Jeon (2023)
Affective/behavioural	Motivation, engagement, enjoyment, confidence, autonomy and self-regulation often improved.	Perceptions could reflect novelty; autonomy could decline through dependence.	Ebadi and Amini (2024); Alotaibi *et al.* (2025); Cao and Phongsatha (2025); Ma and Chen (2024); Wei (2023); Xiao (2025)
Teacher	Potential gains in planning efficiency, readiness, innovation and professional learning.	Verification and governance redistributed rather than eliminated workload.	Al-khresheh (2024); Mohamed (2024); Moorhouse and Kohnke (2024); Nugroho *et al.* (2024)
Assessment	Promising scoring agreement and rapid feedback.	Validity, repeatability, subgroup fairness and model variability remained concerns.	Biju *et al.* (2024); Li *et al.* (2024a); Pack *et al.* (2024)
Integrity/governance	Greater awareness of acceptable use, dependence and assessment redesign.	Policies and disclosure norms remained inconsistent.	Cong-Lem *et al.* (2024); Shen and Chen (2025)

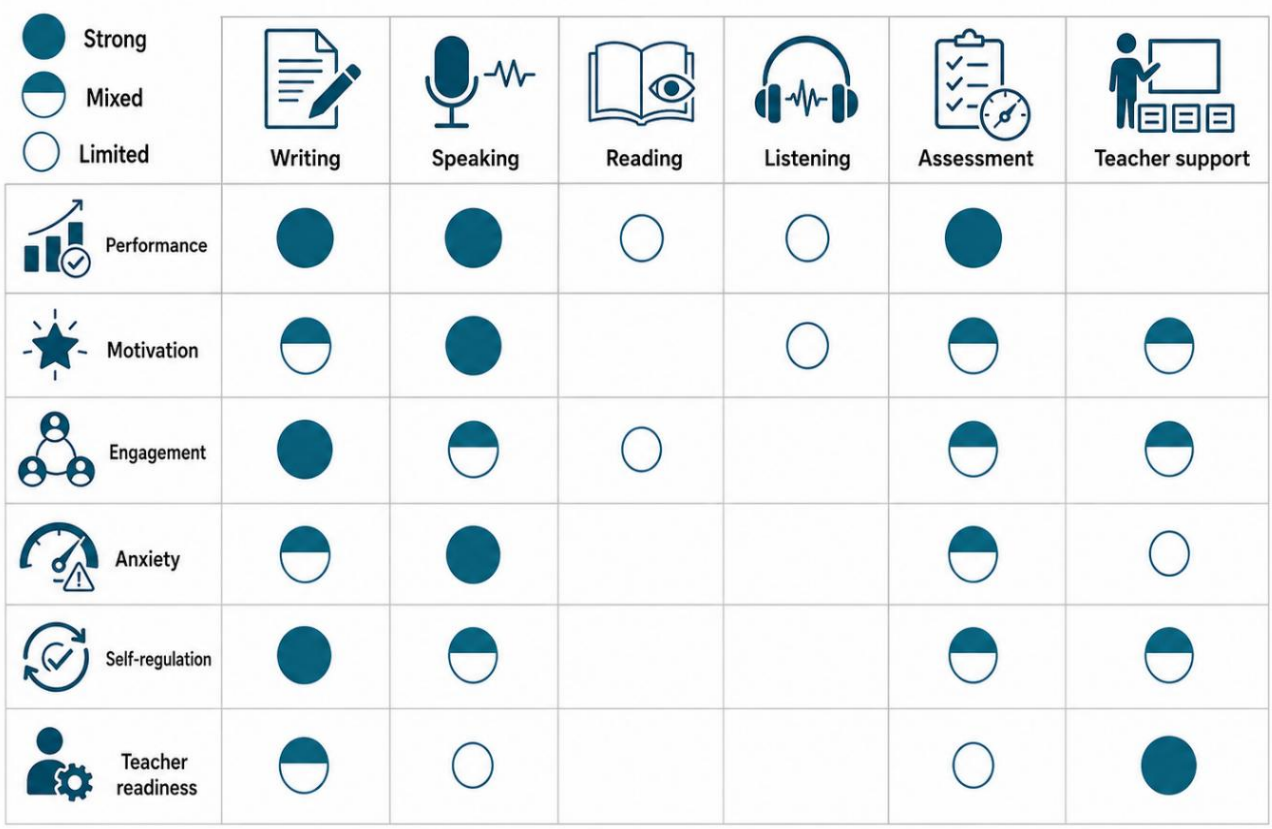

Figure 5. Relationship Map between among technologies, stakeholder groups and outcome domains.

8. Cross-Generational Comparison of AI Systems

The comparison in this section is interpretive rather than a balanced head-to-head empirical analysis. The final corpus contains far more post-2022 studies on LLMs than earlier primary investigations. Differences across technology categories may therefore reflect both actual system characteristics and unequal research coverage. Conventional systems are discussed mainly through historical evidence and task-specific applications. LLM-based systems, in contrast, are represented by a larger and more recent body of educational, perception-based, and technical research.

8.1 Conventional Machine Learning and Automated Systems

Traditional machine-learning and automated systems were typically developed for narrow, well-defined tasks. Their applications included automated writing evaluation, scoring, learner classification, error detection, vocabulary diagnosis, and early forms of personalization. These systems frequently used supervised models trained for a specific educational goal, manually created linguistic features, or structured inputs (Bai and Hu, 2017; Alam *et al.,* 2023). Their primary strength was operational stability. Once a model, rubric, or decision rule was constructed, the system was able to deliver somewhat consistent results within the given task. The limited range of outputs also made it easier to match scores and structured feedback to educational goals. At the functional level, these systems were frequently easier to understand, particularly when specific linguistic indicators or established grading criteria were applied (Bai and Hu, 2017; Alam *et al.,* 2023). Their flexibility, however, was limited. When learner input deviated from the typical task structure, used uncommon language, or required contextual interpretation, performance frequently suffered. These systems could detect surface-level problems, but they were less capable of explaining their reasoning, responding to open-ended queries, or sustaining extended interaction. As a result, their instructional value was greatest in well-defined situations where task control and consistency were more crucial than conversational breadth.

8.2 Deep-Learning and Transformer Systems

Deep learning algorithms increased AI's ability to model complicated linguistic links. Neural architectures aided in text synthesis, machine translation, semantic classification, speech recognition, and automatic feedback. Transformer-based systems improved contextual language representation by learning associations between longer sequences and large corpora (Jeon *et al.,* 2023; Wang *et al.,* 2024; Xiao, 2025). These advancements enhanced specific NLP applications in language education. Transformer models enabled better context-sensitive scoring, error detection, reading comprehension analysis, audio processing, and translation. Compared to classic feature-based systems, they required less manual linguistic engineering and could generalize more effectively across diverse language inputs (Jeon *et al.,* 2023; Wang *et al.,* 2024; Xiao, 2025). Despite this, deep-learning systems relied substantially on task-specific optimization, model design, and training data. Their internal decision-making processes were likewise less transparent than those used in many traditional systems. Performance remained uneven across languages, accents, and learner groups because training datasets frequently favoured high-resource languages and standardized forms. As a result, these systems enhanced technical competence while decreasing interpretability, emphasizing the importance of external evaluation.

8.3 Conversational Ai

Conversational AI has shifted the focus from automated analysis to interaction. Chatbots that were scripted, retrieval-based, task-oriented, and speech-enabled were used to practice vocabulary, speaking, grammar, dynamic assessment, and guided writing. Their key pedagogical contribution was to provide repeated interaction in a low-pressure setting (Huang *et al.,* 2022; Du and Daniel, 2024; Jeon *et al.,* 2023; Ericsson and Johansson, 2023). Because task-oriented chatbots had limited dialogue pathways and response sets, they were typically more predictable than LLM-based systems. They were therefore appropriate for formative exercises and organized practice. Without the social anxiety that comes with interacting with people, learners could practice, adjust interaction speed, and repeat activities. Additionally, educators could create chatbot exercises based on certain learning goals (Jeon *et al.,* 2023; Ericsson and Johansson, 2023; Hsu *et al.,* 2023; Huang *et al.,* 2022). The primary limitation was the narrow range of supported tasks. Earlier chatbots frequently failed when learners moved beyond predefined dialogue patterns. Their responses may become repetitious, shallow, or irrelevant to the learner's intended meaning. Accent sensitivity, limited pragmatic understanding, and recognition errors all had an impact on speech-enabled devices. As a result, conversational AI improved opportunities for interaction but did not reliably replicate the intricacy of real-world human speech.

8.4 Large Language Model Based Systems

A far more comprehensive interaction model was introduced by LLM-based systems. Within a single interface, ChatGPT, GPT-family models, Claude, PaLM, and associated systems might produce explanations, feedback, examples, discussion, translations, summaries, questions, lesson plans, and assessment responses. They were particularly appealing for writing assistance, tutoring, teacher preparation, and conversational practice because of their wide task scope (Albedah, 2025; Li *et al.,* 2024; Escalante *et al.,* 2023; Lo *et al.,* 2024). Flexibility was the main benefit of LLMs. Teachers and students could change prompts, ask for different answers, change the level of difficulty, simulate roles, and mix several activities. Compared to scripted systems, their context sensitivity allowed for richer interaction, and their generative ability allowed for customized replies without the need for a function that was specifically designed for every application (Escalante *et al.,* 2023; Li *et al.,* 2024; Pack *et al.,* 2024; Lo *et al.,* 2024). However, the LLM behaviour was probabilistic rather than fixed. Model version, prompt wording, conversation history, and decoding settings could all influence the output. A system might provide useful feedback in one instance but produce inaccurate or inconsistent advice in another. Proprietary models may also change over time without clear documentation, which reduces reproducibility. LLMs therefore expand educational possibilities, but they offer lower output stability and control.

8.5 Capability Risk Relationship

A comparison of technological generations demonstrates an ongoing trade-off between capability and risk. As systems transitioned from traditional models to LLMs, flexibility expanded faster than reliability. Traditional systems were limited in scope yet often reliable for specialized purposes. LLMs were more adaptive, but they were

more difficult to evaluate consistently over multiple runs and changing tasks (Albedah, 2025; Bao *et al.,* 2025; Pack *et al.,* 2024; Shen and Chen, 2025). Additionally, personalization varied with generation. Previous approaches used student ratings or response patterns to modify predetermined content. Although such personalization frequently relied on conversational signals rather than verified learner models, LLMs could produce more responsive and context-sensitive help. As a result, it could seem personalized without truly reflecting learning needs, misconceptions, or proficiency (Pack *et al.,* 2024; Shen and Chen, 2025; Albedah, 2025; Bao *et al.,* 2025).

In general, transparency decreased as model complexity rose. While transformer and LLM outputs resulted from extremely complicated internal representations, conventional scoring systems could frequently be explained by clear rules or distinguishable traits. This made it challenging to justify the results of a score, correction, or recommendation. Compared to proprietary LLMs, reproducibility was higher in fixed systems. Previous systems were able to undergo repeated testing in stable environments. Providers may alter models without maintaining previous behavior, thus LLM results may differ between sessions. For assessment, where consistency is crucial, this proved especially troublesome (Albedah, 2025; Bao *et al.,* 2025; Pack *et al.,* 2024; Shen and Chen, 2025). Newer systems increased language coverage, but multilingual breadth did not ensure language-to-language performance parity. Low-resource languages, dialects, and code-switched communication continued to be less supported than high-resource languages. As a result, more recent systems preserved underlying data inequities while increasing nominal access.

Human oversight was also necessary for pedagogical value. Because of hallucinations, inconsistencies, and improper production, LLMs required even more rigorous verification than conventional systems, which required teachers to interpret scores and feedback. As technology advanced, teacher supervision became more crucial rather than less (Albedah, 2025; Bao *et al.,* 2025; Pack *et al.,* 2024; Shen and Chen, 2025). Additionally, the dangers related to academic integrity and authorship were more prominent in generative systems. While LLMs might produce significant chunks of the work directly, earlier technologies could only grade or modify student work. This hindered the evaluation of independent learner competency and blurred the line between support and substitution.

8.6 When Newer Systems Do Not Produce Better Outcomes

Stronger educational results did not always follow from greater technological competence. LLMs and sophisticated chatbots were useful for writing assistance, engagement, or conversational practice, according to several studies, however the advantages varied according to the skill. Speaking, listening, and pragmatic competence frequently showed less improvement than writing, vocabulary, and grammar. Positive experiences were sometimes described by students, but there was no indication that their independent performance had improved (Escalante *et al.,* 2023; Wu, 2024; Li *et al.,* 2024; Pack *et al.,* 2024). Furthermore, educational efficacy was not assured by technical supremacy. High-quality questions, good agreement with human raters, or clear explanations may all be produced by a model, but these results did not prove that students retained information or gained a better understanding. In a similar vein, open-ended interaction may boost participation while subjecting students to unreliable or inadequately calibrated feedback (Li *et al.,* 2024; Pack *et al.,* 2024; Steiss *et al.,* 2024; Escalante *et al.,* 2023). Older systems could outperform newer ones when the educational task required consistency, transparency, or strict alignment with a validated rubric. General-purpose LLMs may not be as reliable as automated scoring or diagnostic techniques tailored to a particular architecture. On the other hand, tasks requiring interaction, explanation, content creation, or flexible support were better suited for LLMs. Therefore, a linear hierarchy in which LLMs are a generally superior form of educational AI is not supported by the evidence. The educational task, learner group, risk level, necessary transparency, and amount of instructor supervision all influence the best system (Wu, 2024; Li *et al.,* 2024; Pack *et al.,* 2024; Steiss *et al.,* 2024). While Figure 6 shows the growing relationship between capability, flexibility, uncertainty, and governance requirements across technology generations, Table 6 compares the features of pre-LLM and LLM-based systems.

9. Multilingual Equity and Contextual Variation

Significant disparities in the linguistic and geographical dispersion of AI-supported language instruction were found in the evidence base. The majority of the examined studies were focused on English-language environments, especially English as a foreign or second language, despite the fact that modern AI systems are often touted as multilingual.

Table 6. Comparative analysis of pre-LLM and LLM-based systems

Dimension	Conventional ML/automated systems	Deep-learning/transformer systems	Conversational AI	LLM-based systems
Task scope	Narrow and predefined.	Specialised but context-sensitive.	Bounded dialogue or task flows.	Broad and open-ended.
Output form	Structured scores, labels or corrections.	Predictions, representations and specialised outputs.	Interactive responses within dialogue constraints.	Generated explanations, dialogue, feedback and content.
Flexibility	Low.	Moderate.	Moderate within task boundaries.	High.
Personalisation	Rule- or score-based adaptation.	Data-driven learner or language representation.	Interaction-based but often shallow.	Prompt- and context-sensitive, but not necessarily based on validated learner models.
Reliability/stability	Generally high within validated task.	Moderate to high within fixed model and domain.	Relatively predictable in scripted systems.	Variable across prompts, runs and versions.
Transparency	Functional rules/features may be inspectable.	Lower due to model complexity.	Higher in scripted systems; lower in neural systems.	Low, especially in proprietary models.
Reproducibility	Relatively strong with fixed model/data.	Possible with disclosed architecture and data.	Moderate.	Often weak because of nondeterminism and model updates.
Language coverage	Usually narrow.	Improved but data dependent.	Often limited to programmed languages/tasks.	Broad nominal coverage with unequal quality.
Main educational value	Consistency and controlled diagnosis.	Improved specialised NLP performance.	Repeated low-pressure interaction.	Flexible tutoring, generation and cross-task support.
Primary risks	Construct reduction and limited context.	Opacity and training-data bias.	Restricted authenticity and recognition errors.	Hallucination, authorship ambiguity, privacy and inconsistent assessment.
Human oversight	Interpretation and validation.	Technical and pedagogical validation.	Task design and interaction monitoring.	Continuous verification and governance.

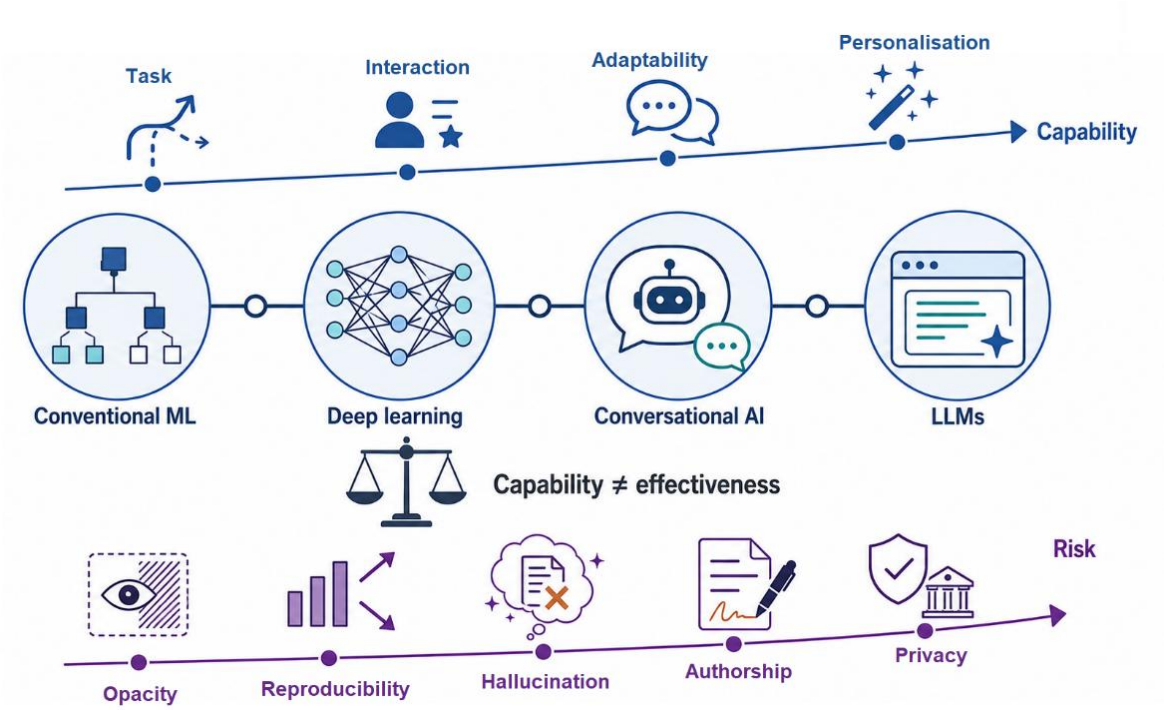

Figure 6. Relationship between technological capability, pedagogical flexibility, output uncertainty, and governance requirements across AI technology generations.

This focus influenced learner groups, cultural contexts, and educational problems in addition to the technologies that were assessed. Thus, multilingual equity became a design issue as well as an empirical gap (Albedah, 2025; Li *et al.,* 2023).

9.1 Dominance of English and High-Resource Languages

English was the principal target language across the 40 primary studies. The evidence concentrated on English as a foreign language, English as a second language, academic writing, English-medium instruction, speaking practice, feedback, assessment, and teacher use. Low-resource languages and endangered languages were largely absent, while studies centred on other high-resource languages remained comparatively limited. This distribution means that the review primarily reflects AI use in English-oriented educational settings rather than multilingual language education in its full diversity (Li *et al.,* 2023; Albedah, 2025). The dominance of English may be associated with the availability of digital corpora, commercial systems, research funding, institutional infrastructure, and established language-learning markets. However, the language being learned, the language used to interact with the AI system, the learners' first languages, and the language of publication should be reported separately. An English-learning study may still involve multilingual learners or bilingual prompting. A supplementary study-level table should therefore document these distinctions wherever the original studies provide sufficient information (Albedah, 2025; Li *et al.,* 2023). This disparity is especially crucial for LLMs. The multilingual output quality of these systems varies, despite their ability to produce text in multiple languages. Fluency in widely represented languages does not guarantee grammatical accuracy, pragmatic appropriateness, cultural sensitivity, or instructional usefulness in lower-resource languages.

9.2 Multilingual and Cross-Linguistic Applications

Only limited number of studies examined translation, multilingual use, cross-linguistic support, or AI-assisted multilingual learning were few in number. This research showed that AI can offer multilingual access to learning resources, contrastive explanations, instant translation, and paraphrase. Additionally, users might request bilingual explanations, compare linguistic forms, and switch between languages using generative systems (Albedah, 2025; Li *et al.,* 2023). By facilitating rapid comprehension and reducing access hurdles, these tools can help multilingual learners. Additionally, they might help educators create differentiated resources for classrooms with a variety of language backgrounds. However, rather than using controlled learning studies, many multilingual applications were

assessed through perceptions, demonstrations, or informal use (Albedah, 2025; Li *et al.,* 2023). Cross-linguistic support was most effective when AI served as a scaffold rather than a replacement for target-language production. Although bilingual explanations and translation could aid students in understanding challenging material, an over-reliance on them ran the danger of decreasing effective language use. Therefore, the pedagogical benefit of multilingual AI depended on how and when students switched between languages, whether or not teachers facilitated cross-linguistic use, and whether or not translation aided comprehension.

9.3 Underrepresentation of Low-Resource Languages

In the final evidence base, low-resource languages were significantly underrepresented. Few studies specifically examined educational applications for languages with inadequate benchmark datasets, limited digital corpora, poor speech recognition support, or minimal commercial investment. The lack of AI-supported resource generation, tutoring, translation, and documentation for these languages is a significant drawback (Albedah, 2025; Li *et al.,* 2023). Additionally, the existing claims about multilingual inclusion are based primarily on nominal system capabilities rather than proven educational efficacy due to the absence of data. A model may produce inconsistent, culturally inappropriate, or subpar output when processing input in a low-resource language. Similarly, non-standard varieties, code-switched communication, and regional accents may cause speech systems to perform poorly (Albedah, 2025; Li *et al.,* 2023). A cumulative disadvantage results from underrepresentation. Languages with fewer digital resources often receive weaker model support, which are therefore used less in teaching and produce fewer datasets for further development. AI may therefore increase rather than lessen language inequality in the absence of intentional intervention.

9.4 Cultural and Linguistic Bias

In both learner-facing and teacher-facing apps, cultural and linguistic bias emerged as a pervasive issue. Training data may contain dominant cultural assumptions that are reflected in AI-generated examples, feedback, lesson material, and assessment responses. This may result in feedback that favors standardized variants over regional or acceptable alternatives, incorrect examples, or limited definitions of acceptable language (Pack *et al.,* 2024; Cong-Lem *et al.,* 2024; Albedah, 2025). In automated evaluation, bias is particularly significant. Even when meaning is obvious, systems that were predominantly trained on dominant language varieties may penalize dialectal, regional, or multilingual characteristics. Similarly, without acknowledging rhetorical variation, writing tools may promote adherence to a limited academic standard (Albedah, 2025; Cong-Lem *et al.,* 2024; Pack *et al.*, 2024). Linguistic bias also affects speech technologies. Recognition errors may be detrimental to learners whose pronunciation deviates from the accent patterns represented in the training data. The technology might misidentify otherwise intelligible speech in some situations. Because learners could receive inaccurate feedback that reflects model constraints rather than actual language issues, this raises concerns about equity and education.

9.5 Geographic Concentration

A large share of the primary research was conducted in Asian EFL situations. Studies commonly involved with instructors and students in Saudi Arabia, China, Vietnam, Iran, and other Asian countries. On the other hand, there was little data from South Asia, Africa, and Latin America. Additionally, many studies were limited to a specific university, institution, or country (Rahman *et al.,* 2024; Albedah, 2025; Sharadgah and Sa'di, 2022). Although this focus demonstrates the widespread use of educational technology in Asian EFL contexts, it also restricts generalizability. Different educational systems have different curricula, assessment cultures, teacher autonomy, infrastructure, language policies, and attitudes toward AI. Results from technologically advanced universities might not be applicable to rural institutions, multilingual public schools, or resource-constrained environments (Rahman *et al.,* 2024; Albedah, 2025; Sharadgah and Sa'di, 2022). A small number of cross-national comparison studies were also included in the evidence base. As a result, it was challenging to ascertain whether variations in results were caused by the technology itself or by institutional, cultural, and educational factors.

9.6 Digital Inequality and Infrastructure

Devices, internet connectivity, platform availability, subscription fees, teacher proficiency, and institutional infrastructure all affected access to AI-supported language instruction. Despite the apparent widespread availability

of commercial LLMs, their efficient usage frequently necessitates digital literacy, appropriate hardware, dependable internet access, and, occasionally, paid features (Albedah, 2025; Al-khresheh, 2024; Mohamed, 2024). Institutions and students were both impacted by digital inequality. Students with wider access to technology could use sophisticated tools more often and on their own. Others relied on shared infrastructure or restricted free versions. Additionally, curriculum integration, privacy compliance, technological support, and professional training presented challenges for educators in settings with limited resources (Albedah, 2025; Al-khresheh, 2024; Mohamed, 2024). Because speech, multimodal, and embodied systems demand more bandwidth, hardware, and maintenance than text-based tools, infrastructure constraints were especially pertinent. Disparities between marginalized educational environments and well-resourced institutions may grow as a result of these circumstances.

9.7 Implications for Inclusive Design

Nominal multilingual capacity alone is insufficient for inclusive AI-supported language instruction. Languages, dialects, accents, skill levels, and cultural contexts should all be taken into consideration when evaluating systems. Instead of depending exclusively on general-purpose models trained on unbalanced corpora, low-resource language development should incorporate local educators, linguists, and communities (Albedah, 2025; Li et al., 2023; Pack *et al.,* 2024). Transparency about language coverage and recognized limitations should also be addressed through inclusive design. Teachers and students need to know exactly where a system works well and where human verification is required. Speech systems should be assessed for accent fairness, and assessment instruments should be verified across language groups prior to high-stakes use (Albedah, 2025; Li *et al.,* 2023; Pack *et al.,* 2024).

Table 7. Multilingual and contextual distribution

Dimension	Observed pattern	Equity implication	Priority response
Target language	English, EFL and ESL contexts dominated.	Findings cannot be generalised to all languages.	Conduct comparative validation across high- and low-resource languages.
Multilingual use	Some translation, bilingual explanation and multilingual content applications.	Nominal multilingual capability may exceed demonstrated educational quality.	Evaluate accuracy, pragmatics and learning outcomes by language.
Low-resource languages	Substantially underrepresented.	Existing data inequalities may be reinforced.	Support community-led datasets, localisation and open resources.
Geography	Strong concentration in Asian tertiary and EFL settings.	Limited transferability to other regions and education systems.	Use multicountry and multicentre designs.
Educational level	Tertiary education predominated; younger learners were uncommon.	Developmental suitability and child safety remain uncertain.	Expand carefully governed primary and secondary studies.
Setting	Classroom, online, blended, mobile and platform-based contexts were represented.	Effects may be confounded by infrastructure and teacher support.	Report implementation conditions and compare settings.
Access/infrastructure	Commercial access, devices, connectivity and teacher competence varied.	AI adoption may widen institutional and learner inequalities.	Provide equitable access, alternatives and institutional support.
Culture/linguistic variety	Limited direct evaluation of dialects, accents and local rhetorical norms.	Models may privilege dominant language varieties.	Validate subgroup performance and involve local experts.

AI should encourage multilingual repertoires rather than enforce monolingual standards from a pedagogical standpoint. When carefully combined, translation, translanguaging, and bilingual scaffolding can be beneficial for education. In order to prevent AI adoption from favouring already privileged students, educational institutions should also address pricing, accessibility, privacy, and technical assistance. The language, geographic, and contextual representation of the included research is summarized in Table 7. Overall, the evidence indicates that AI has considerable potential to support multilingual education. However, the current research remains too concentrated on English and high-resource settings to justify broad claims of equal linguistic parity.

10. Methodological Quality and Reporting Adequacy

The methodological quality of the 40 primary studies varied considerably across educational, technical, qualitative, and mixed-method research. Although methodological development did not always advance at the same rate as the technology, the evidence indicated that interest in AI-assisted language instruction was expanding. Strong causal designs, longitudinal follow-up, repeatable technical reporting, and external validation were only applied in a small number of investigations. Nevertheless, a number of studies offered helpful preliminary data regarding usability, user acceptance, and immediate learning advantages. As a result, the evaluation took into account not only if favourable effects were documented but also whether the results were reliable, repeatable, and transferable (Albedah, 2025; Bao *et al.,* 2025; Wu, 2024).

10.1 Research-Design Distribution

The major evidence consisted of experimental, quasi-experimental, quantitative descriptive, qualitative, mixed-method, technological validation, case-study, and survey-based designs. The majority of research on speaking, writing, listening, vocabulary, motivation, and self-regulated learning was experimental or quasi-experimental. These studies usually contrasted pre-intervention performance, alternative technology, or traditional education with AI-supported training (Albedah, 2025; Wu, 2024). Learner experiences, teacher preparedness, technological acceptance, and classroom integration were often investigated using qualitative and perception-based approaches. To give a more comprehensive picture of the effects of schooling, mixed-method studies integrated achievement scores, questionnaires, interviews, or system-use data. Technical research assessed human-AI agreement, automated scoring, generated feedback, question quality, and model reliability (Albedah, 2025; Wu, 2024). Although this design diversity was suitable for a new interdisciplinary subject, it also made direct comparisons between research more difficult. While technical research focused on accuracy or system performance, educational experiments frequently prioritized learning outcomes. Although they were unable to prove causal effectiveness, qualitative research provided significant contextual understanding. Therefore, multidimensional synthesis rather than a straightforward ranking of technologies was supported by the evidence base.

10.2 Sample Size and Setting

Small classroom cohorts, interview groups, and larger survey and platform-based datasets were among the sample sizes. Numerous studies were carried out in a particular course, institution, or country. Small samples were especially prevalent in teacher-readiness studies, exploratory qualitative research, and classroom interventions (Albedah, 2025; Bao et al., 2025; Wu, 2024). Single-site research decreased external validity but provided in-depth contextual information. Results from a single university, school, or language program may be impacted by local curricula, student proficiency, digital infrastructure, teacher competency, or institutional attitudes towards artificial intelligence. Transferability was not always determined by adequately describing these requirements (Albedah, 2025; Bao *et al.,* 2025; Wu, 2024). Larger text or scoring datasets were occasionally employed in technical research, although corpus size by itself did not ensure instructional usefulness. Even though a model hasn't been tested with real students or classroom assignments, it might do well on a sizable benchmark. Therefore, rather than relying solely on numerical size, sample adequacy has to be evaluated in connection to the research question.

10.3 Intervention Duration

The majority of instructional interventions were brief. Many lasted one training unit, a few sessions, or several weeks. These designs were helpful for assessing short-term success, engagement, or immediate viability, but they were insufficient to establish long-term language development (Albedah, 2025; Wu, 2024). For complex

goals including writing proficiency, speaking fluency, self-regulation, and learner autonomy, short interventions were particularly troublesome. Improvements seen right away following an AI-supported activity could be more indicative of task familiarity, novelty, intense practice, or transient motivation than of long-term competency (Albedah, 2025; Wu, 2024). The extent to which learners maintained their gains following the intervention or applied them to autonomous tasks without AI support was only partially investigated. Therefore, when assessing claims of long-term educational effectiveness, cautions is required due to the predominance of short-term studies.

10.4 Control Groups and Comparison Conditions

There were differences in the quality of comparison conditions. Some studies used wait-list controls, alternate feedback techniques, non-AI technology, or traditional training. Others (Wu, 2024; Steiss *et al.,* 2024; Tai and Chen, 2024; Xiao, 2025) relied on pre-test and post-test comparisons without a distinct control group. It was challenging to separate the impact of AI from more practice, teacher attention, novelty, or task repetition in the absence of suitable comparison conditions. Control groups did not always receive equivalent instructional time or feedback intensity, even when control groups were included (Wu, 2024; Steiss *et al.,* 2024; Tai and Chen, 2024; Xiao, 2025). Direct comparisons between AI supported and human instruction were uncommon. This was an important limitation because improvement from baseline alone does not establish the educational value of AI. Stronger study designs should compare AI-assisted instruction with well-designed human feedback, existing educational technologies, and blended human-AI approaches.

10.5 Objective versus Self-Reported Outcomes

The discrepancy between self-reported impressions and objectively evaluated results was a significant methodological distinction. Writing scores, speaking performance ratings, vocabulary tests, listening comprehension scores, automated scoring agreement, and system accuracy metrics were among the objective criteria. Perceived utility, contentment, motivation, confidence, anxiety, intention to adopt, and teacher preparation were all measured by self-report (Wu, 2024; Alotaibi *et al.,* 2025; Mohamed, 2024; Teng, 2024). Results based on perception were generally more favourable than those based on performance. AI was frequently characterized by educators and students as practical, efficient, interesting, and convenient. These optimistic opinions, however, did not always translate into quantifiable progress (Wu, 2024; Alotaibi *et al.,* 2025; Mohamed, 2024; Teng, 2024). Additionally, self-reported data were susceptible to novelty effects, social desirability, and expectation bias. Because the technology is new or because they think positive attitudes are expected, participants may react favourably. More robust research combined perceptions with performance information, behavioural records, interviews, or long-term data.

10.6 Longitudinal Evidence

Longitudinal evidence was limited. The majority of research examined results at a particular point in time or right after an intervention. Few included follow-up observations, repeated measurements, or delayed post-test (Ericsson and Johansson, 2023; Wu, 2024). Due to the cumulative nature of language learning, this gap was significant. When support is removed, short-term improvements can decrease, and continued use of AI could have consequences that weren't apparent at first. Dependency, diminished autonomous problem solving, shifts in feedback literacy, or changing trust are a few examples (Ericsson and Johansson, 2023; Wu, 2024). To determine whether teachers transition from experimental use to steady pedagogical integration, longitudinal designs are also required. Conclusions regarding retention, transfer, behavioural change, and institutional implementation were constrained by the lack of long-term follow-up.

10.7 Model and Prompt Reporting

AI models and prompts were not consistently reported, especially in experiments in LLM based studies. The model family was identified in several tests, but the precise version, access date, system parameters, and prompt sequence were not disclosed. Despite the fact that model capabilities can vary significantly between versions and subscription levels, others have defined ChatGPT in a generic manner (Lo *et al.,* 2024; Albedah, 2025; Pack *et al.,* 2024). For replication, timely reporting was frequently insufficient. Seldom were complete system prompts, user prompts, follow-up instructions, rubric formulations, temperature settings, or response-selection processes fully documented. This omission was significant in scoring or feedback investigations since small prompt adjustments

could significantly influence results (Lo *et al.,* 2024; Albedah, 2025; Pack *et al.,* 2024). The issue of model upgrades also affected technical research that used proprietary platforms. If the provider changes the system after data collection, the study may no longer be reproducible. Researchers should therefore report the exact model version, access date, prompt design, and generation settings.

10.8 Reproducibility and External Validation

Reproducibility was limited in both technical and educational studies. Reusable datasets, scoring scripts, prompt libraries, and detailed protocols were rarely available. Proprietary tools created an additional barrier to independent verification (Albedah, 2025; Bao *et al.,* 2025; Pack *et al.,* 2024). External validation was also uncommon. Numerous models were assessed within the same dataset, learner group, language variety, or institution. It was challenging to ascertain whether results would generalize without testing in different contexts, populations, or languages (Albedah, 2025; Bao *et al.,* 2025; Pack *et al.*, 2024). There were few replication studies. This could be somewhat explained by the speed at which technology is developing, yet ongoing innovation without replication erodes the body of cumulative evidence. To determine whether these results remain consistent across model versions, institutions, proficiency levels, and cultural contexts, more independent research is needed.

10.9 Common Methodological Weaknesses

Small sample sizes, single-institution designs, brief interventions, a lack of delayed post-tests, reliance on self-report measures, inadequate prompt documentation, ambiguous model versions, restricted replication, and a lack of external validation were the most frequent flaws. Other issues included inconsistent outcome measures, inadequate control of confounding variables, incomplete reporting of comparison conditions, and overinterpretation of technical or perception-based results (Albedah, 2025; Bao *et al.,* 2025; Wu, 2024). These flaws lessen trust in general causal assertions, but they do not invalidate the body of evidence. Clear research objectives, suitable comparison groups, objective and subjective assessments, open technical reporting, and cautious interpretation were all included in the best studies. Multicenter designs, longer interventions, delayed evaluations, preregistered procedures, open materials, human benchmarks, subgroup analysis, and external validation should be prioritized in future research (Albedah, 2025; Bao *et al.,* 2025; Wu, 2024). The methodological characteristics, quality markers, and reporting constraints of the primary studies are summarised in Table 8. A methodological-quality heat map covering the main design domains is shown in Figure 7, which also emphasizes the evidence base's strengths and recurrent flaws.

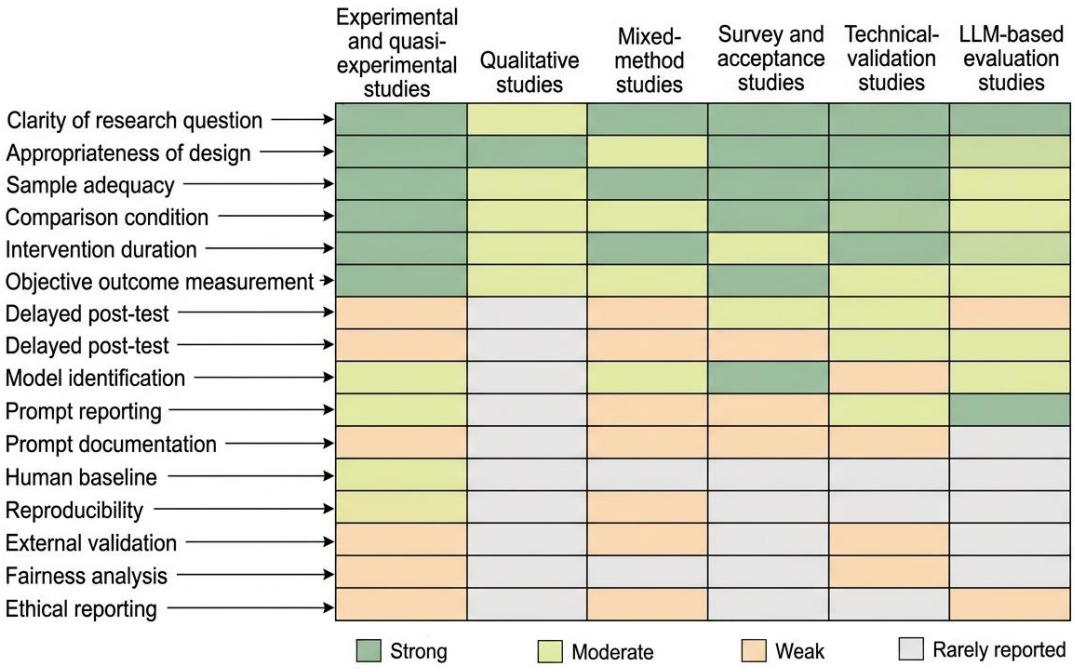

Figure 7. Methodological comparison map.

Table 8. Common reporting deficiencies and corrective requirements

Reporting deficiency	Consequence	Minimum corrective requirement
Unclear model or product version	Results cannot be attributed to a stable system.	Report exact model, version, provider and access date.
Incomplete prompts	Outputs and findings cannot be replicated.	Provide system/user prompts and follow-up sequence.
Missing generation settings	Variability cannot be interpreted.	Report temperature, decoding, number of runs and selection rule.
Insufficient intervention description	Pedagogical mechanism remains unclear.	Describe duration, frequency, task sequence, teacher role and learner instructions.
Weak comparison-condition detail	Observed differences may reflect unequal treatment.	Report instructional time, feedback exposure and baseline equivalence.
Overreliance on self-report	Effectiveness may be overstated.	Add validated performance or behavioural indicators.
No delayed post-test	Retention and transfer cannot be established.	Include delayed assessment after AI support is reduced or removed.
Limited subgroup analysis	Bias and differential performance remain hidden.	Report results by proficiency, language, accent or relevant learner group.
No external validation	Generalisability is uncertain.	Test in an independent institution, dataset, language or cohort.
Unavailable materials/data	Independent verification is constrained.	Share prompts, instruments, code, anonymized data or detailed protocols where feasible.

11. Ethics, Academic Integrity, and Governance

The governance analysis operates at three levels. The first addresses issues including hallucinations, assessment reliability, academic integrity, privacy, learner reliance, and unequal access that were specifically mentioned in the included primary studies. The second identifies patterns that surfaced from cross-study comparisons, namely the connection between output variability and verification burden. The third offers normative suggestions based on the integrated evidence. These consist of institutional data-governance practices, disclosure regulations, appeal processes, and risk-based oversight. These suggestions should not be interpreted as interventions that underwent consistent testing in all of the included research. Beyond technical performance, the increasing use of AI in language instruction has sparked ethical, scholarly, and institutional concerns. AI can help with feedback, evaluation, writing, speaking practice, teacher planning, and personalized learning, according to the analyzed studies. Concerns over hallucinations, prejudice, authorship, privacy, justice, responsibility, and unequal access accompany these advantages, though. Because AI-generated outputs can directly affect learner knowledge, writing production, assessment outcomes, and judgments of linguistic correctness, these dangers are particularly significant in language education. Thus, coordinated effort at the learner, teacher, institutional, and system-design levels is necessary for responsible adoption (Albedah, 2025; Cong-Lem *et al.,* 2024; Shen and Chen, 2025).

11.1 Hallucination and Epistemic Reliability

In generative AI and LLM-based applications, hallucinations have grown to be a significant concern. Inaccurate explanations, made-up references, deceptive examples, or inappropriate corrections in clear and compelling language can all be produced by these algorithms. Because the answers frequently seem confident and

authoritative, learners with low proficiency or subject knowledge may find it challenging to identify such inaccuracies (Lo *et al.,* 2024; Albedah, 2025; Cong-Lem *et al.,* 2024; Shen and Chen, 2025). Therefore, epistemic dependability is both a technological and pedagogical matter. Inaccurate recommendations can create grammatical errors, change meaning, or promote improper rhetorical choices in writing and feedback applications. Unsupported interpretations might have an impact on comments or ratings in language evaluations. If outputs are not independently verified during teacher planning, fabricated content could wind up in the classroom (Lo *et al.,* 2024; Albedah, 2025; Cong-Lem *et al.,* 2024; Shen and Chen, 2025). Systematic verification is the main mitigation. Explanations, examples, references, scores, and feedback produced by AI should be regarded as preliminary results that need to be reviewed by humans. A generative model should never be the only one used to make important judgments. Institutions should also set up verification processes commensurate with the educational risk and mandate the explicit disclosure of system limitations.

11.2 Bias and Cultural Representation

Cultural presumptions, prevailing language norms, unequal representation of linguistic variety, and unbalanced training data can all lead to bias. Dialectal, regional, multilingual, or culturally distinctive forms may be viewed as inadequate by AI systems that were largely trained on high-resource languages and standardized English (Albedah, 2025; Li et al., 2023; Pack *et al.,* 2024). Such bias can influence what is presented as appropriate, natural, formal, or academically acceptable in language instruction. While voice recognition software may incorrectly identify non-standard accents, automated feedback may favour dominant rhetorical conventions. Additionally, generated examples may convey culturally limited information or perpetuate prejudices (Albedah, 2025; Li *et al.,* 2023; Pack *et al.,* 2024). Evaluation across languages, dialects, accents, and cultural contexts is necessary for mitigation. While schools should refrain from viewing model output as an impartial language authority, developers should report subgroup performance. To guarantee that systems accurately represent linguistic diversity, local educators and language experts should be involved in the tool selection, validation, and adaptation processes.

11.3 Academic Integrity and Undisclosed AI Use

The concepts of authorship, autonomous labor, and plagiarism have become more complex due to generative AI. Previous technologies were mostly used to score or correct writing. On the other hand, LLMs are capable of producing concepts, sentences, paragraphs, translations, and complete assignments. According to Cong-Lem *et al.* (2024) and Shen and Chen (2025), this raises questions regarding when AI provides appropriate help and when it takes the place of the learner's own input. Because submitted work may no longer accurately reflect the learner's true competence, undisclosed use can reduce assessment validity. Planning, drafting, rewriting, and reflecting are all part of the learning process in writing-intensive language courses, therefore this issue is especially significant (Cong-Lem *et al.,* 2024; Shen and Chen, 2025). Since detection tools are still ineffective and access is ubiquitous, strict prohibition alone is unlikely to be effective. Defining permissible and prohibited uses, demanding disclosure, creating activities to make the learning process apparent, and incorporating oral defense, version history, reflective commentary, or supervised components are examples of more defensible strategies. Policies pertaining to academic integrity should provide a clear distinction between content replacement, editing, translation, collaboration, and support.

11.4 Authorship and Ownership

Authorship concerns go beyond misconduct. Concerns around creative contribution, ownership of generated content, and accountability for mistakes are also brought up by AI-assisted writing. Although AI systems cannot take on academic responsibility, users may use significant AI-generated content in assignments, instructional materials, or publications (Cong-Lem *et al.,* 2024; Shen and Chen, 2025; Teng, 2024). The human user should continue to bear responsibility. It is the responsibility of both students and teachers to make sure that any work that is turned in or disseminated is correct, unique, and suitable. Institutions should also establish when and how AI help must be acknowledged, as well as the ownership terms associated with commercial platforms. In order to avoid creating copyright uncertainty, care must be taken when created outputs resemble protected property or when prompts contain unpublished student or institutional content (Cong-Lem *et al.,* 2024; Shen and Chen, 2025; Teng, 2024).

11.5 Privacy and Student-Data Protection

Sensitive data is frequently used in AI-assisted language learning. Student writing, speech recordings, assessment results, demographic information, performance metrics, and behavioural traces are examples of this. Users may not fully comprehend how such data is handled, saved, preserved, or utilized when it is entered into commercial platforms (Albedah, 2025; Cong-Lem *et al.,* 2024; Mohamed, 2024). When organizations use external tools without doing a thorough data-governance evaluation, privacy hazards rise. Even if they are unable to give meaningful consent for data processing, students may feel pressured to use these platforms as part of a course (Albedah, 2025; Cong-Lem *et al.,* 2024; Mohamed, 2024). Institutions should prefer systems with clear procedures for data access, deletion, and retention should be preferred by institutions. They should refrain from uploading private student content and collect only the bare minimum of identifiable information. Before incorporating AI into assessment or institutional learning systems, contractual protections, approved-use procedures, and data-protection impact evaluations are necessary.

11.6 Fairness in Automated Assessment

One of the most important issues with automated grading and feedback is fairness. A system may consistently disadvantage specific linguistic groups, competency levels, dialects, or writing styles while producing ratings that appear to be accurate. Since human scoring may sometimes contain bias, agreement with human raters does not prove fairness on its own (Biju *et al.,* 2024; Li *et al.,* 2024; Pack *et al.,* 2024). Therefore, it is important to evaluate assessment methods across relevant learner groups and linguistic variety. Prior to use, institutions should look at construct alignment, error patterns, score consistency, and subgroup performance. In high-stakes situations, automated outputs should be viewed as supplemental evidence rather than the last word. Additionally, learners must have access to review, explanation, and appeal procedures (Biju *et al.,* 2024; Li *et al.,* 2024; Pack *et al.,* 2024).

11.7 Human Oversight

The most consistent governance requirement across applications was human oversight. According to Ji *et al.* (2023), Li *et al.* (2024), Ji *et al*. (2024), Pack *et al.* (2024), and others, teachers are still in charge of ensuring that the use of AI is in line with pedagogical goals, verifying created outputs, interpreting automated scores, and safeguarding learner data. Oversight should be meaningful rather than symbolic. Without adequate technical and pedagogical knowledge, a teacher cannot effectively oversee the use of AI. A clear division of responsibilities is therefore necessary. Humans should retain control over interpretation, adaptation, and decisions with significant consequences. AI may assist by generating, recommending, classifying, or scoring (Ji *et al.,* 2023; Li *et al.,* 2024; Ji *et al.,* 2024; Pack *et al*., 2024).

11.8 Institutional Policy and Governance

Institutional policy development often lagged behind classroom adoption. Teachers were left to decide on permissible use, disclosure requirements, privacy protections, and evaluation procedures on their own in the absence of explicit guidelines. As a result, students experienced uncertainty and inconsistency among courses (Cong-Lem *et al.,* 2024; Mohamed, 2024; Moorhouse, 2024; Shen and Chen, 2025). Approved tools, forbidden data uses, disclosure criteria, evaluation guidelines, supervision duties, and protocols for reporting errors or harm should all be outlined in effective governance. Due to the quick changes in model capabilities and commercial terms, policies should be changed on a regular basis. Academic leaders, instructors, students, IT personnel, legal counsel, librarians, and ethics experts should all be involved in governance (Cong-Lem *et al.,* 2024; Mohamed, 2024; Moorhouse, 2024; Shen and Chen, 2025).

11.9 Teacher and Learner AI Literacy

AI literacy is essential for both risk mitigation and effective education. Prompt design, output verification, bias identification, privacy protection, assessment redesign, and pedagogical integration are all skills that educators must possess. System constraints, authorship obligations, source verification, disclosure requirements, and the dangers of improper dependence should all be understood by learners (Al-khresheh, 2024; Mohamed, 2024; Moorhouse, 2024; Moorhouse and Kohnke, 2024). Technical proficiency is insufficient on its own. Critical thinking,

ethical consciousness, and the capacity to differentiate trustworthy information from merely fluent output are all necessary for responsible AI literacy. Therefore, operational skills should be linked to academic integrity, governance, and pedagogy in professional development (Moorhouse, 2024; Moorhouse and Kohnke, 2024; Nugroho *et al.,* 2024; Al-khresheh, 2024).

11.10 Equity and Access

Subscription fees, internet quality, device availability, language assistance, disability access, and institutional infrastructure all contribute to unequal access to AI. Students who can afford more sophisticated systems might benefit more than others who rely on limited or free versions. There are comparable disparities between institutions with and without adequate resources (Albedah, 2025; Li *et al.,* 2023; Mohamed, 2024). Affordable access, user-friendly interfaces, multilingual support, alternative forms of paticipation, and institutional backing when AI usage is necessary are all necessary for equitable adoption. According to Albedah (2025), Li *et al.* (2023), and Mohamed (2024), educational institutions should refrain from basing instruction or assessment on resources that certain students cannot equally access. In general, ethical and governance issues should not be viewed as external constraints considered only after implementation. These are prerequisites for educational validity. While Figure 8 shows an ethics and governance structure linking system design, pedagogical usage, human oversight, institutional policy, and learner protection, Table 9 summarizes the main risks, repercussions, and mitigation solutions.

Table 9. Ethical risks, consequences and mitigation strategies

Risk	Educational manifestation	Potential consequence	Mitigation and governance response
Hallucination/epistemic unreliability	Fabricated references, inaccurate explanations, unsuitable feedback or scoring rationales.	Mislearning, invalid assessment and loss of trust.	Human verification, source checking, uncertainty disclosure and prohibition of autonomous high-stakes decisions.
Cultural/linguistic bias	Preference for dominant varieties, accents and rhetorical norms.	Unfair correction, marginalisation and culturally narrow instruction.	Subgroup validation, local expert review and plural linguistic norms.
Academic integrity	Undisclosed generation or substitution of learner work.	Invalid assessment of independent competence.	Permitted-use rules, disclosure, process evidence, oral defence and assessment redesign.
Authorship/ownership	Unclear human contribution and use of generated or protected material.	Misattribution, copyright uncertainty and weak accountability.	Human responsibility, acknowledgement rules and institutional guidance.
Privacy/data protection	Student writing, speech or assessment data entered into external platforms.	Unauthorised retention, reuse or disclosure.	Data minimisation, approved tools, contractual safeguards and impact assessment.
Assessment fairness	Unequal scoring across language groups, dialects or proficiency levels.	Discriminatory or invalid decisions.	Construct validation, subgroup testing, human moderation and appeal.
Opacity/reproducibility	Unknown model changes and unexplained outputs.	Inconsistent practice and non-auditable decisions.	Version control, logging, audit trails and transparent reporting.
Overdependence	Learners delegate planning, revision or problem solving.	Reduced autonomy and uncertain independent competence.	AI-free tasks, reflection, staged scaffolding and monitoring.

Unequal access	Differences in subscriptions, devices, connectivity and language support.	Widening educational inequality.	Institutional provision, accessible alternatives and inclusive procurement.
Governance gaps	Inconsistent teacher decisions and unclear accountability.	Policy conflict, unmanaged risk and stakeholder uncertainty.	Risk-tiered policy, defined roles, review cycles and incident procedures.

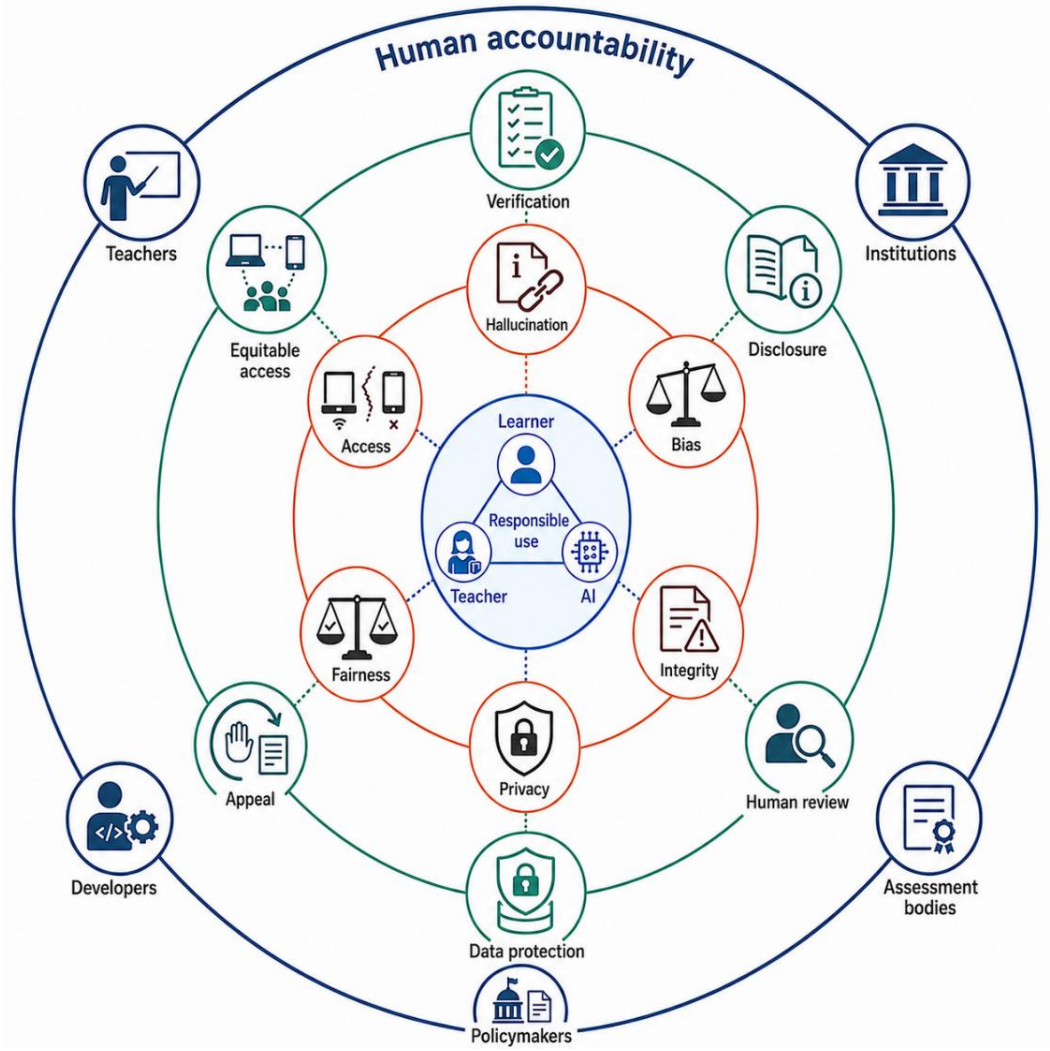

Figure 8. Ethics and governance framework.

12. Integrative Framework for Responsible AI-Supported Language Education

The synthesis indicates that the educational value of AI cannot be explained by technological capability alone. Outcomes were shaped by the interaction among system characteristics, educational task, pedagogical design, learner agency, teacher mediation, contextual conditions, methodological quality, and institutional governance. The framework developed in this review integrates these factors into a provisional model for responsible AI-supported language education. It is intended to support interpretation, research design, and institutional decision-making rather than to represent a statistically validated causal model (Liang *et al.,* 2024; Ji *et al.,* 2023; Li *et al.,* 2024; Bao *et al.,* 2025). The framework was constructed through a hybrid deductive–inductive procedure. The initial deductive categories were derived from RQ2–RQ8 and the predefined extraction domains. During synthesis, recurrent inductive codes were added for learner agency, feedback literacy, teacher mediation, contextual fit, output verification, reproducibility, multilingual inclusion, and institutional accountability. Related codes were consolidated into seven

higher-order dimensions. Negative, contradictory, and methodologically weak findings were retained rather than removed from the framework-building process. The seven dimensions are AI-system characteristics; language and educational task; pedagogical integration; human–AI interaction; educational and technical outcomes; contextual moderators; and ethical and governance conditions. Relationships directly supported by the primary evidence are presented as evidence-derived associations. Feedback loops and governance pathways are presented as conceptual propositions developed from the integrated synthesis. Future research should test these relationships through longitudinal, comparative, and multi-context studies.

12.1 Framework Dimensions

There are seven interconnected dimensions in the framework. Characteristics of AI systems, such as technical generation, model type, modality, adaptability, transparency, reliability, reproducibility, language coverage, and cost, are covered in the first dimension. While generative models offer more flexibility but necessitate more stringent verification and supervision, conventional machine-learning methods typically give higher stability within specific tasks. The language and educational tasks constitute the second dimension. The type of task determines whether AI is appropriate. Different degrees of language sensitivity, engagement, consistency, and risk control are needed for writing, speaking, listening, reading, vocabulary, grammar, translation, evaluation, feedback, tutoring, and teacher assistance. A system that does well in low-stakes writing exercises might not be appropriate for summative evaluation or pronunciation diagnosis.

Pedagogical integration is the subject of the third dimension. The way AI is integrated into the curriculum, instruction, feedback, evaluation, and learner assistance determines the educational value. When teachers give clear instructions, activities are well-planned, scaffolding is suitable, and learning objectives are clear, the AI use is more defensible. Unstructured access could increase convenience but not result in significant learning. The fourth dimension is the interaction between humans and AI. The paradigm acknowledges that AI can serve as a collaborator, tutor, assistant, evaluator, conversational partner, or content creator. Teachers and students must maintain the power to make decisions at the same time. Whether AI improves or detracts from learning depends on learner agency, feedback literacy, critical review, and teacher mediation. Linguistic and educational outcomes are included in the fifth dimension. Language competence, writing quality, speaking fluency, vocabulary and grammatical growth, motivation, engagement, anxiety, self-efficacy, autonomy, self-regulation, teacher preparedness, and assessment quality are some of these. Technical performance, adoption intentions, perceptions, and objectively assessed results are all considered distinct types of evidence in this paradigm.

The sixth dimension consists of contextual moderators. Learner proficiency, educational attainment, language-resource status, infrastructure, culture, teacher preparedness, and institutional capability are some of these. These elements affect social acceptance, linguistic dependability, educational suitability, and system access. Conditions related to ethics and governance are covered in the seventh dimension. Hallucination control, bias mitigation, academic integrity, authorship, privacy, fairness, openness, human oversight, equitable access, and institutional policy are a few of these. These criteria establish the legitimacy, dependability, and accountability of AI-supported practice. As a result, rather than existing independently of educational excellence, they are essential to it.

12.2 Core Pathway

The framework presents a central pathway connecting human-AI interaction, pedagogical design, AI capabilities, the learning process, and language and educational outcomes. This route demonstrates the need for appropriate pedagogical design to transform technological capability into educational value. Even the most sophisticated AI system cannot enhance learning on its own. The use of AI in education is shaped by prompt design, learning activities, feedback processes, instructor roles, learner responsibilities, and evaluation conditions. These factors influence how students and instructors engage with the technology and decide whether or not its capabilities result in meaningful learning outcomes. The learning process is then impacted by human-AI interaction. AI can be used by students for autonomous revision, critical assessment of feedback, and repeated practice. They might also start to rely on generated content. AI can be used by educators to monitor progress, scaffold learning, and verify outputs. But they can potentially give the system too much power. Whether AI promotes deeper learning or results in passive dependence, superficial correction, and unreliable assessment depends on these interaction patterns.

12.3 Moderating Conditions

The central pathway is shaped by learner, institutional, linguistic, and sociocultural conditions. Learner proficiency affects the ability to identify incorrect feedback and produce effective prompts. The degree of supervision needed and the appropriateness of open-ended generative technologies depend on educational attainment. Model accuracy, speech recognition, translation quality, and cultural representation are all impacted by the availability of language resources. Because effective usage necessitates both technical proficiency and evaluative judgment, teacher preparedness also influences pedagogical integration. Communication norms, views toward automation, authorship requirements, and appropriate teacher-student-AI relationships are all influenced by culture. Access to devices, connectivity, subscriptions, speech technology, and institutional support are all determined by infrastructure. Permitted uses, privacy protection, disclosure requirements, assessment design, and accountability are all influenced by policy and ethics. These factors contribute to the explanation of why the same system may provide various outcomes in different educational environments, learner groups, institutions, and languages.

12.4 Feedback and Redesign Loops

The framework is iterative rather than linear. Redesign should be prompted by weak learning outcomes, poor feedback uptake, learner reliance, bias, hallucinations, unreliable assessment, privacy problems, or poor learner fit. Redesign could entail altering the technology, reducing the scope of the work, updating prompts, boosting teacher mediation, enhancing learner preparation, adding human evaluation, or substituting automated procedures with non-AI alternatives. Because early involvement may represent novelty rather than long-term educational benefit, positive outcomes should also be tracked over time. Thus, ongoing evaluation is necessary for system performance, pedagogy, learner behaviour, assessment validity, and institutional governance. Three testable hypotheses are generated by the framework. First, the relationship between AI capabilities and language-learning outcomes will be strengthened by closer pedagogical alignment and stronger instructor mediation. Second, the link between system flexibility and instructional efficacy will be moderated by learner proficiency and verification competence. Third, as outputs become more consequential, unpredictable, and open-ended, the demand for governance will grow.

The integrated framework is shown in Figure 9 as a cyclical model where human contact, educational design, contextual factors, and ethical protections filter technical capabilities. Therefore, choosing and overseeing the most suitable system for a specific educational goal rather than using the most sophisticated system on the market is considered responsible use. From the synthesis, the framework produces five propositions. First, improved educational outcomes are not a direct result of increased system capabilities. Alignment with a well-defined linguistic task and a suitable instructional design determine its value. Second, while open-ended generative capability increases the flexibility of instruction, it also raises the necessity for human monitoring, verification, and transparent reporting. Third, whether AI promotes active learning or fosters linguistic and cognitive dependence depends on student agency and teacher mediation. Fourth, student proficiency, target language, infrastructure, institutional capability, cultural context, and language-resource availability all have an impact on how effective education is. Fifth, governance is not a supplement to implementation. The legitimacy and acceptance of AI-supported learning are directly impacted by privacy, fairness, integrity, openness, and accountability.

13. Research Gaps and Future Research Agenda

Four criteria were used to prioritise the research gaps: lack of evidence, educational significance, methodological flaws, and implications for equality or governance. Four questions were used to evaluate each gap. These examined whether the limitation decreased replication or generalizability, whether it presented a significant risk of exclusion or damage, whether it affected a significant language skill or learner group, and whether existing conclusions relied mostly on weak or indirect evidence. The resulting priorities are not based on a validated numerical scale and are therefore interpretive (Albedah, 2025; Bao *et al.,* 2025; Wu, 2024).

13.1 Longitudinal and Delayed Learning Outcomes

The most significant study gap is still the absence of longitudinal evidence. Most studies assessed performance immediately after an intervention, learner perceptions, or short-term engagement. These designs are unable to demonstrate if the improvements are maintained after AI support is removed, transferred to independent

tasks, or maintained. Delayed post-tests, repeated measures, and follow-up periods spanning semesters or academic years should all be included in future research. The question of whether long-term AI use increases learner autonomy and feedback literacy or increases reliance on automated support should receive special consideration.

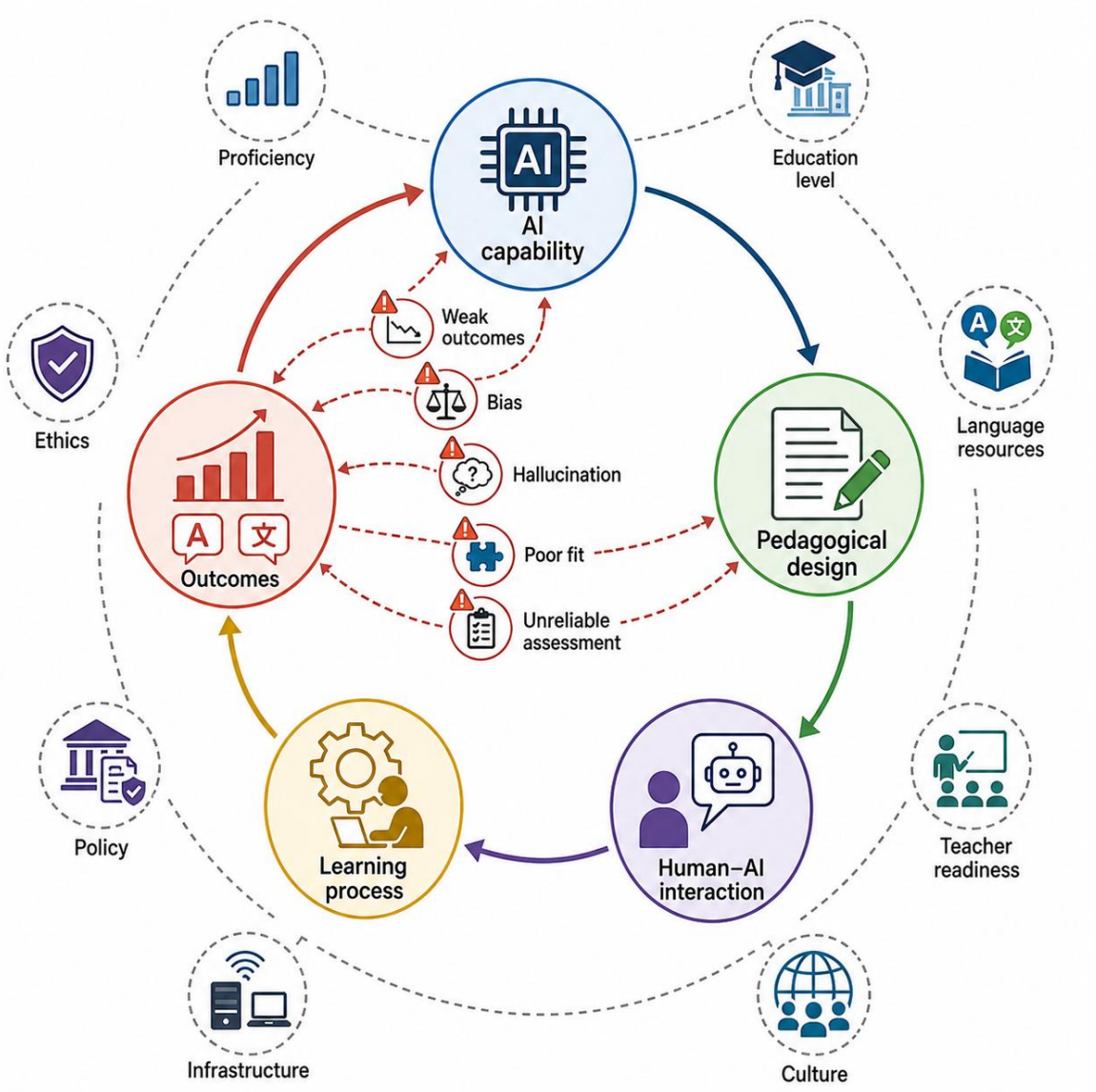

Figure 9. Integrative framework for responsible AI-supported language education.

13.2 Low-Resource and Multilingual Contexts

English and other high-resource languages continue to be the primary focus of research. This disparity raises the possibility of replicating current language disparities and restricts the generalizability of current findings. Low-resource languages, regional variations, dialects, code-switching, and multilingual teaching methods should all be investigated in future research. Local educators, students, linguists, and communities should be included in the development and assessment of AI systems. To ascertain whether multilingual model competency translates into equivalent instructional quality across languages, comparative research is also needed.

13.3 Reading and Listening

Reading and listening are still understudied, although writing and speaking predominate in the body of research. AI-supported inferential reading, critical comprehension, source evaluation, listening in real-world settings, accent variation, and multimodal comprehension should all be the focus of future research. The technical quality of produced materials and real learner gains should be distinguished in research. To determine if adaptive reading and

listening systems enhance strategic competency rather than just short-term task performance, longer interventions are required.

13.4 Younger Learners

The majority of research was done in higher education. Despite the fact that younger learners may differ significantly in digital literacy, self-regulation, vulnerability to false content, and need for supervision, evidence involving primary- and secondary-school learners was limited. Future studies should look into developmental suitability, child-data protection, parental and teacher mediation, age-appropriate interfaces, and the impact of AI on fundamental language abilities. Because open-ended generative systems may expose children to inappropriate content or promote unthinking dependence, studies with younger learners should implement tighter precautions.

13.5 Human-AI Comparative Studies

The educational value of AI is frequently measured against pre-intervention performance rather than strong human or non-AI alternatives. AI-only, teacher-only, peer-only, and blended human-AI conditions should all be compared in future studies. Equivalent instructional duration, task difficulty, and feedback intensity should be used in these comparisons. Such designs would make it clear whether AI is solely useful as an auxiliary tool, offers a true educational edge, or is as effective at a lower cost. Additionally, rather than being viewed as a secondary variation of technology use, human-AI collaboration should be investigated as a separate condition.

13.6 Teacher Orchestration

Although teacher mediation became a key factor in determining efficacy, it was rarely thoroughly investigated. Future research should look into how educators choose resources, create prompts, organize tasks, check results, track student dependency, and integrate AI-generated feedback into lessons. Research should determine which professional skills are necessary and which orchestration techniques yield better results. Comparative research between inexperienced and seasoned educators may shed light on how pedagogical proficiency influences the effectiveness of AI-assisted instruction.

13.7 Explainable and Transparent Assessment

Automated assessments necessitate stronger evidence, fairness, and explainability. Future research should look at how AI-generated scores connect to intended constructs, how teachers and students can understand scoring judgments, and whether models perform consistently across demographic groups, languages, and competence levels. Human moderation, appeal processes, subgroup analysis, and repeated-run dependability should all be included in high-stakes evaluation studies. Explainability should go beyond the significance of technical features to provide pedagogically meaningful explanations for scores and comments.

13.8 Prompt and Model Reporting

Reproducibility is currently limited by inadequate reporting of model versions, prompts, access dates, and generation parameters. The precise model, version, date of access, system prompt, user prompt, follow-up instructions, decoding settings, number of generations, and response-selection process should all be recorded in future research. When feasible, prompt libraries and assessment procedures should be made available. Because proprietary models are subject to change over time, researchers should archive representative outputs and make it apparent whether the results were acquired prior to or following significant model revisions.

13.9 Replication and External Validation

More replication is needed in the field. Numerous conclusions are based on a particular learner group, a single program, or a single institution. Promising interventions should be replicated in future research across nations, educational levels, proficiency groups, and languages. Automated scoring, speech recognition, and adaptive learning systems should all require external validation. To ascertain whether reported effects are robust or context-specific, multicenter trials and shared benchmark tasks might be helpful.

13.10 Privacy and Governance

There is still a lack of research on institutional responsibility, data retention, consent, and privacy. Future research should look at the storage, processing, and reuse of learner writing, voice recordings, assessment data, and behavioural traces. Evidence-based governance frameworks that cover authorized tools, data minimization, transparency, human oversight, risk classification, and incident reporting protocols are also necessary for institutions. It may be possible to ascertain whether governance structures are both practical and protective for teaching through comparative examinations of institutional policies.

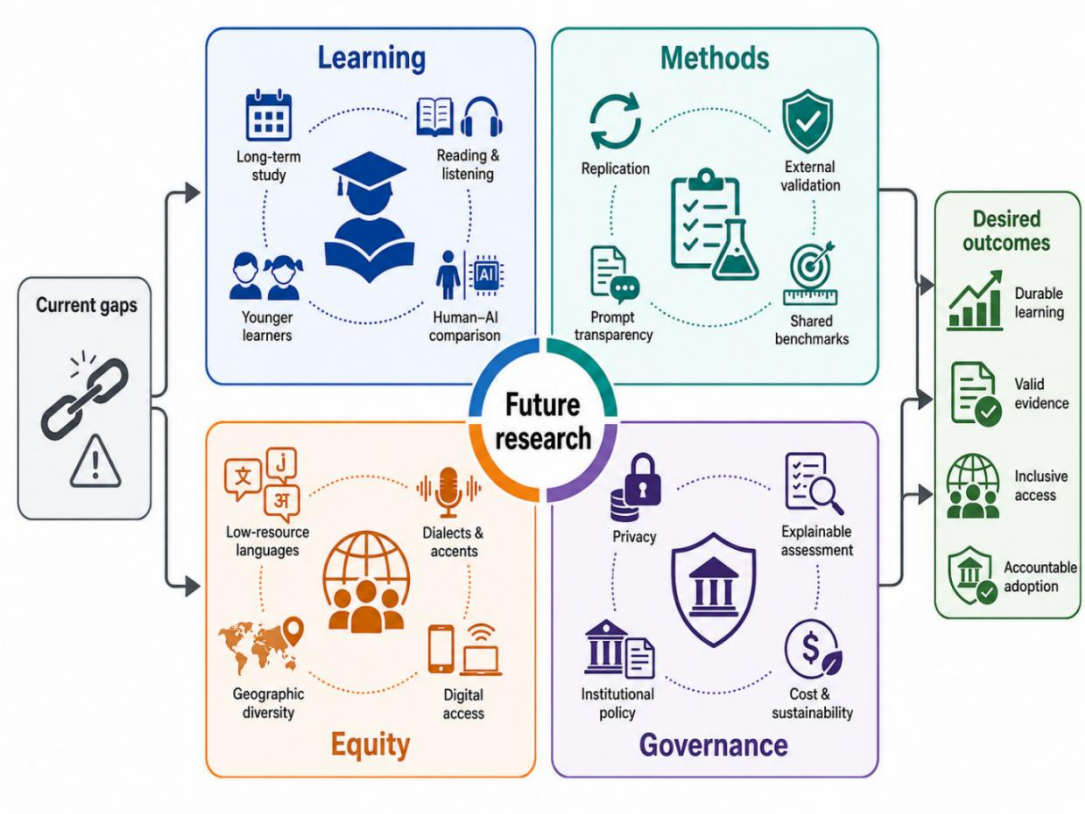

Figure 10. Future research agenda.

Table 10. Prioritised research-gap matrix

Research gap	Evidence deficiency	Educational importance	Methodological urgency	Priority level	Recommended research direction
Longitudinal and delayed outcomes	Very limited follow-up evidence.	High: retention, transfer and dependency are unknown.	High	Critical	Semester/year-long studies with delayed tests and AI-withdrawal conditions.
Low-resource and multilingual contexts	English/high-resource dominance.	High: equity and linguistic validity.	High	Critical	Community-led multilingual validation and cross-language comparisons.
Reading and listening	Sparse compared with writing and speaking.	High: major language competencies remain underexamined.	High	Critical	Controlled, skill-specific and multimodal interventions.
Younger learners	Limited primary/secondary-school evidence.	High: developmental suitability and child protection.	High	High	Age-appropriate, safeguarded school-based research.

Human–AI comparison	Few equivalent teacher-only, AI-only and blended designs.	High: added pedagogical value remains unclear.	High	High	Matched-comparison and factorial studies.
Teacher orchestration	Teacher mediation rarely isolated as a mechanism.	High: implementation quality depends on teachers.	High	High	Observe and test task design, verification and monitoring practices.
Explainable assessment	Validity and fairness exceed available explanation.	High: consequential decisions.	High	High	Construct-based explanation, subgroup fairness and appeal research.
Prompt/model reporting	Inconsistent technical description.	Moderate to high: reproducibility.	High	High	Standardised LLM reporting checklists and shared prompt libraries.
Replication/external validation	Predominantly single-site and single-model evidence.	High: generalisability.	High	High	Multicentre replication across models, languages and proficiency levels.
Privacy/governance	Limited empirical evaluation of institutional safeguards.	High: learner rights and accountability.	High	High	Policy comparison, privacy impact and governance implementation studies.
Environmental/financial cost	Rarely measured.	Moderate: sustainability and institutional feasibility.	Moderate	Moderate	Cost-effectiveness, infrastructure and energy-use analysis.
Open-source versus commercial systems	Few direct educational comparisons.	Moderate to high: control, cost and privacy.	Moderate	Moderate	Comparative trials of effectiveness, localisation, privacy and total cost.

13.11 Environmental and Financial Costs

The current body of evidence pays little attention to the financial and environmental effects of AI adoption. Subscription fees, infrastructure needs, teacher preparation, technical assistance, device access, energy usage, and the cost of maintaining commercial systems should all be investigated in future studies. In institutions with limited resources, these concerns are particularly crucial. Rather than assuming that automation will automatically lower costs, cost-effectiveness studies should compare AI-supported approaches with current pedagogical alternatives.

13.12 Open-Source versus Commercial Systems

It is necessary to compare open-source and commercial systems directly. Commercial models frequently offer robust performance and user friendly access, but they have lower levels of reproducibility, data control, and transparency.

Although they need the right infrastructure and technical know-how, open-source solutions may provide more customization, local hosting, and multilingual adaptability. The educational efficacy, cost, privacy, language coverage, stability, teacher usability, and institutional control of these systems should all be compared in future research. In general, durability, transfer, equity, openness, and governance should be prioritized over technological innovation in the future research agenda. An author-proposed research agenda employing qualitative priority levels based on methodological urgency, ethical significance, educational importance, and evidence scarcity is shown in Table 10. These classifications are not a proven numerical ranking; rather, they are interpretive. The future agenda at the learner, teacher, institutional, technology, and policy levels is summarized in Figure 10.

14. Implications

The findings show that the value of AI in language instruction depends more on the quality of pedagogical integration, human oversight, contextual suitability, and governance than on technological novelty. Although AI can provide access to practice, feedback, personalization, and instructional support, its advantages are still contingent. Therefore, the ramifications go beyond adoption in the classroom to include system design, learner preparation, institutional policy, public regulation, and scholarly quality control (Albedah, 2025; Wu, 2024; Bao *et al.,* 2025).

14.1 Implications for Language Teachers

AI tools should be evaluated by language teachers based on their educational goals rather than their level of technological complexity. A system that supports a well-defined learning aim with acceptable levels of dependability, transparency, and learner control is the most suitable. Therefore, educators should differentiate between tools that are appropriate for low-stakes practice and those that are suitable for grading, feedback, or other important decisions. Before using AI-generated feedback, explanations, examples, and evaluation results, they should be examined. Instead of allowing students to passively accept AI results, teachers should create activities that force them to interpret, defend, edit, and reflect on them. Evidence of the writing process can be seen in drafts, revision histories, and reflective commentary. These documents aid in the preservation of authorship and the visibility of learning. AI should enhance human contact rather than replace it in speaking and conversational practice, particularly in areas like pragmatic skill, precise pronunciation, and culturally appropriate communication. Therefore, prompt design, output verification, bias identification, privacy protection, assessment redesign, and learner dependence monitoring should all be included in professional development.

14.2 Implications for Learners

Students should be aware that fluent AI output isn't necessarily suitable, correct, or helpful for learning. Critical analysis, source verification, comparison of alternative answers, and a firm grasp of the distinction between aid and substitution are necessary for effective utilization. When students use AI to ask questions, practice frequently, test concepts, or edit their own work, it can promote autonomy. However, over-reliance might impair one's ability to write independently, solve problems, plan, and take linguistic risks. As a result, students should document how AI was used, note which recommendations were accepted or rejected, and provide justification for significant changes. Disclosure, authorship, privacy, and appropriate usage all require clear expectations. Students should be aware that there may be dangers to data security when they post private, sensitive, or unpublished content to commercial sites.

14.3 Implications for Institutions

Coordinated governance should take the place of dispersed individual adoption in institutions. Institutional policies should specify approved tools, acceptable and unacceptable applications, disclosure obligations, data security guidelines, evaluation procedures, and human review roles. Additionally, policies should make a distinction between high-risk applications like automated grading, placement, or disciplinary decisions and low-risk uses like brainstorming or language practice. Where AI use is required, educational institutions should ensure that all students have equal access to it and refrain from using assessment methods that favour those who can buy more sophisticated technologies. Teacher preparation, technical assistance, accessibility, multilingual services, and privacy assessment all require funding. Adopted systems should be routinely assessed to make sure they still adhere to legal, ethical, and educational criteria.

14.4 Implications for Developers

Language-education systems should be designed for instructional validity rather than technical performance alone. Systems should enable instructor control, convey uncertainty, offer interpretable feedback, and permit modification to curriculum, learner proficiency, and linguistic environment. Multiple languages, dialects, accents, educational levels, and demographic groups should all be considered in the evaluation. Model limits, subgroup performance, data provenance, version changes, and known failure scenarios should all be reported by developers.

Audit records, explanation methods, repeated-run testing, and human override choices should all be included in tools designed for assessment. Accessibility, low-resource language support, local data control, and privacy-preserving design have to be considered essential characteristics rather than optional ones.

14.5 Implications for Policymakers

Legislators should create criteria for educational AI that are both adaptable and enforceable. Protection of student data, accountability for detrimental or discriminatory results, automated evaluation, commercial procurement, transparency, and accessibility should all be covered by regulations. Both unfettered adoption and outright ban should be avoided in policy. Stronger controls should be implemented to high-stakes assessments, vulnerable learner groups, sensitive data, and opaque proprietary systems in a risk-based approach. Public funding should also support multilingual datasets, open educational models, teacher preparation, and infrastructure for under-resourced institutions.

14.6 Implications for Journal Editors and Assessment Bodies

Transparent reporting of model versions, prompts, access dates, generating settings, datasets, comparison circumstances, and human-verification processes should be mandated by journal editors. Studies that focus mostly on learner views shouldn't be presented as proof of the efficacy of education unless they are accompanied by performance metrics. Technical accuracy should also not be used as evidence of improved learning. Stricter guidelines for automated scoring and feedback systems should be implemented by assessment organizations. Construct alignment, human-rater comparison, repeated-run consistency, subgroup fairness, external validation, and appeal processes should all be part of the validation process. Accountable human review should continue to be applied to high-stakes decisions. Alignment of technological capabilities, pedagogical intent, evidential strength, equity, and governance is necessary for responsible adoption across all stakeholder groups.AI should therefore be considered as a conditional educational resource, the worth of which is determined by how well it is chosen, implemented, monitored, and improved.

15. Limitations of the Review

Several limitations should be considered when interpreting this review. First, the final corpus was constructed to provide analytical coverage of technological, pedagogical, methodological, multilingual, and governance issues. It should not be interpreted as a complete bibliometric census of global research production. Publication frequencies, journal distributions, regional patterns, and technological categories describe the included evidence base rather than the entire field. Second, the corpus was temporally uneven. Most publications appeared after 2022, and no primary studies from 2018–2021 were included. The review therefore represents the generative-AI era more strongly than earlier stages of AI-supported language education. Earlier technologies were used mainly as historical and conceptual comparators. The cross-technology analysis should therefore not be interpreted as an evenly balanced longitudinal comparison.

Third, only English-language publications were included in the review. The seeming dominance of English and other high-resource language environments may have been reinforced by this restriction. It might have also overlooked pertinent research that was written in a different language. The evidence was subsequently focused on Asian EFL environments and universities. Low-resource languages, reading, listening, younger students, Africa, Latin America, and a number of multilingual areas continued to be underrepresented. Fourth, the research design, sample characteristics, intervention length, AI system, language proficiency, educational context, comparison condition, and outcome measure varied greatly among the included studies. Formal meta-analytic aggregation over the entire corpus was made impossible by this heterogeneity. Thus, several syntheses were made of technical performance, learner perceptions, affective results, and objectively measurable learning outcomes. Fifth, the quality of reports varied. Model versions, prompts, system settings, intervention fidelity, delayed testing, external validation, and subgroup performance were all inadequately covered in a number of studies. Because interfaces, costs, access requirements, and system behaviour may alter after publishing, rapid changes in proprietary AI platforms can lower reproducibility.

Sixth, some studies could not be thoroughly analyzed because complete methodological or outcome information was not always accessible. The conclusions of these research were limited to details that could be validated. They should not be utilized to establish detailed causal or quality assertions unless the entire text is available. Finally, publication bias cannot be eliminated. Studies reporting positive, new, or statistically significant findings may be published more frequently than studies revealing null effects, implementation failure, or negative consequences. The conclusions should therefore be regarded as a critical synthesis of the available evidence rather than a definitive estimate of the overall effectiveness of AI in language education. The evidence base varied by learner group, geography, and level of language proficiency. Reading, listening, younger learners, low-resource languages, and other global regions were significantly underrepresented compared to writing, speaking, postsecondary education, and Asian EFL contexts. These disparities make it more difficult to draw broad conclusions and highlight the need for longer-term, more varied studies.

16. Conclusion

This systematic integrative review examined 40 primary empirical or technical studies and used 15 secondary evidence syntheses to contextualise AI-supported language education. The corpus was strongly concentrated after 2022 and therefore represents the generative-AI era more fully than earlier technological periods. Conventional machine-learning systems, automated assessment, deep-learning applications, and task-oriented chatbots were consequently used mainly as historical and conceptual comparators. The evidence indicates that generative AI and LLM-based systems have expanded the range of language-education activities that can be supported through one interface. These systems can assist writing, feedback, speaking practice, tutoring, assessment, translation, content development, and teacher preparation. However, broader capability does not establish stronger educational effectiveness. Many positive findings were based on short interventions, small samples, single institutions, technical performance, learner perceptions, or self-report measures. Evidence for retention, transfer, longitudinal development, younger learners, reading, listening, low-resource languages, and multilingual educational validity remained limited.

The synthesis also identified a capability-control relationship. Task-specific systems generally operated within clearer boundaries, whereas LLM-based systems offered greater adaptability and interaction but produced more variable and difficult-to-reproduce outputs. Prompt sensitivity, changing model versions, hallucination, authorship ambiguity, privacy, fairness, and assessment reliability create additional requirements for verification and human oversight. The proposed framework therefore treats responsible adoption as a conditional relationship among AI-system characteristics, educational task, pedagogical integration, human–AI interaction, contextual moderators, educational outcomes, and governance conditions. The framework is not a validated causal model. It provides a synthesis-derived structure for future research and institutional decision-making. The central conclusion is that technological novelty should not be used as a substitute for educational evidence. AI usage is best justified when it serves a clear language-learning goal, retains learner agency, incorporates instructor mediation, respects linguistic and social diversity, and follows transparent institutional norms. Future research should prioritize longitudinal studies, direct comparison designs, multilingual validation, consistent reporting standards, external replication, subgroup fairness, and detailed documentation of models, prompts, and system settings.

References

Alam, S., Usama, M., Alam, M.M., Jabeen, I., Ahmad, F. (2023). Artificial Intelligence in global world: A case study of Grammarly as e-Tool on ESL learners' writing. *International Journal of Information and Education Technology*, 13(11), 1747-1747. https://doi.org/10.18178/ijiet.2023.13.11.1984

Albedah, F. (2025). Artificial Intelligence in Language Education: A Systematic Review of Multilingual Applications, Large Language Models, and Emerging Challenges. *Language Teaching Research Quarterly*, 49, 247-268. https://doi.org/10.32038/ltrq.2025.49.13

Al-khresheh, M.H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence,* 6, 100218. https://doi.org/10.1016/j.caeai.2024.100218

Alotaibi, H.M., Sonbul, S.S., El-Dakhs, D.A. (2025). Factors influencing the acceptance and use of ChatGPT among English as a foreign language learners in Saudi Arabia. *Humanities and Social Sciences Communications, 12, 628*. https://doi.org/10.1057/s41599-025-04945-2

Annamalai, N., Rashid, R.A., Hashmi, U.M., Mohamed, M., Alqaryouti, M.H., Sadeq, A.E. (2023). Using chatbots for English language learning in higher education. Computers and Education: *Artificial Intelligence,* 5, 100513. https://doi.org/10.1016/j.caeai.2023.100153

Bai, L., Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology,* 37(1), 67-81. https://doi.org/10.1080/01443410.2016.1223275

Bao, W., Wang, T., Zhang, L., Yusop, F.D., Ruan, X. (2025). A systematic review of AI in second language acquisition using the expanded SAMR model (2015–2024). *Discover Computing,* 28(1), 292. https://doi.org/10.1007/s10791-025-09833-6

Biju, N., Abdelrasheed, N.S.G., Bakiyeva, K., Prasad, K.D.V., Jember, B. (2024). Which one? AI-assisted language assessment or paper format: An exploration of the impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. *Language Testing in Asia*, 14(1), 45. https://doi.org/10.1186/s40468-024-00322-z

Cao, S., Phongsatha, S. (2025). An empirical study of the AI-driven platform in blended learning for Business English performance and student engagement. *Language Testing in Asia*, 15(1), 39. https://doi.org/10.1186/s40468-025-00376-7

Cong-Lem, N., Tran, T.N., Nguyen, T.T. (2024). Academic Integrity in the Age of Generative AI: Perceptions and Responses of Vietnamese EFL Teachers. *Teaching English with Technology*, 24(1), 28-47. https://doi.org/10.56297/FSYB3031/MXNB7567

Du, J., Daniel, B.K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence,* 6, 100230. https://doi.org/10.1016/j.caeai.2024.100230

Ebadi, S., Amini, A. (2024). Examining the roles of social presence and humanlikeness on Iranian EFL learners' motivation using artificial intelligence technology: A case of CSIEC chatbot. *Interactive Learning Environments,* 32(2), 655-673. https://doi.org/10.1080/10494820.2022.2096638

Ericsson, E., Johansson, S. (2023). English speaking practice with conversational AI: Lower secondary students' educational experiences over time. *Computers and Education: Artificial Intelligence*, 5, 100164. https://doi.org/10.1016/j.caeai.2023.100164

Escalante, J., Pack, A., Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Fathi, J., Rahimi, M., Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, 122, 103254. https://doi.org/10.1016/j.system.2024.103254

Ghafouri, M. (2024). ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System*, 120, 103209. https://doi.org/10.1016/j.system.2023.103209

Guo, K., Li, Y., Li, Y., Chu, S.K.W. (2024a). Understanding EFL students' chatbot-assisted argumentative writing: An activity theory perspective. *Education and Information Technologies,* 29(1), 1-20. https://doi.org/10.1007/s10639-023-12230-5

Guo, K., Pan, M., Li, Y., Lai, C. (2024b). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, 63, 100962. https://doi.org/10.1016/j.iheduc.2024.100962

Guo, K., Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Hong, Q.N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4), 285-291. https://doi.org/10.3233/EFI-180221

Hsu, M.-H., Chen, P.-S., Yu, C.-S. (2023). Proposing a task-oriented chatbot system for EFL learners' speaking practice. *Interactive Learning Environments,* 31(7), 4297-4308. https://doi.org/10.1080/10494820.2021.1960864

Huang, W., Jia, C., Hew, K.F., Guo, J. (2024). Using chatbots to support EFL listening decoding skills in a fully online environment. *Language Learning & Technology*, 28(2), 62-90. https://doi.org/10.64152/10125/73572

Huang, W.J., Hew, K.F., Fryer, L.K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237-257. https://doi.org/10.1111/jcal.12610

Huang, X., Zou, D., Cheng, G., Chen, X., Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1), 112-131. https://doi.org/10.30191/ETS.202301_26(1).0009

Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 36(7), 1338-1364. https://doi.org/10.1080/09588221.2021.1987272

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, 37(1-2), 1-26. https://doi.org/10.1080/09588221.2021.2021241

Jeon, J., Lee, S., Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 206, 104898. https://doi.org/10.1016/j.compedu.2023.104898

Ji, H., Han, I., Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1), 48-63. https://doi.org/10.1080/15391523.2022.2142873

Ji, H., Han, I., Park, S. (2024). Teaching foreign language with conversational AI: Teacher-student-AI interaction. *Language Learning & Technology*, 28(2), 91-108. https://doi.org/10.64152/10125/73573

Karataş, F., Abedi, F.Y., Gunyel, F.O., Karadeniz, D., Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, 29(15), 19343-19366. https://doi.org/10.1007/s10639-024-12574-6

Kwon, S.K., Shin, D., Lee, Y. (2023). The application of chatbot as an L2 writing practice tool. *Language Learning & Technology*, 27(1), 1-19. https://doi.org/10.64152/10125/73541

Li, B., Bonk, C.J., Kou, X. (2023). Exploring the multilingual applications of ChatGPT: Uncovering language learning affordances in YouTuber videos. *International Journal of Computer-Assisted Language Learning and Teaching*, 13(1), 1-22. https://doi.org/10.4018/IJCALLT.326135

Li, J., Huang, J., Wu, W., Whipple, P.B. (2024a). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, 11(1), 1268. https://doi.org/10.1057/s41599-024-03755-2

Li, Y., Zhou, X., Chiu, T.K. (2024). Systematic review on artificial intelligence chatbots and ChatGPT for language learning and research from self-determination theory: What are the roles of teachers? *Interactive Learning Environments*, 33(3), 1850-1864. https://doi.org/10.1080/10494820.2024.2400090

Liang, J.C., Hwang, G.J., Chen, M.R.A., Darmawansah, D. (2024). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, 31(7), 4270-4296. https://doi.org/10.1080/10494820.2021.1958348

Lin, G.Y., Jhang, C.C., Wang, Y.S. (2024). Factors affecting parental intention to use AI-based social robots for children's ESL learning. *Education and Information Technologies,* 29(5), 6059-6086. https://doi.org/10.1007/s10639-023-12023-w

Lo, C.K., Yu, P.L.H., Xu, S., Ng, D.T.K., Jong, M.S.Y. (2024). Exploring the Application of ChatGPT in ESL/EFL Education and Related Research Issues: A Systematic Review of Empirical Studies. *Smart Learning Environments*, 11(1), 50. https://doi.org/10.1186/s40561-024-00342-5

Ma, Y., Chen, M. (2024). AI-empowered applications' effects on EFL learners' engagement in the classroom and academic procrastination. *BMC Psychology*, 12(1), 739. https://doi.org/10.1186/s40359-024-02248-w

Mingyan, M., Noordin, N., Razali, A.B. (2025). Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: An empirical investigation. *Humanities and Social Sciences Communications,* 12(1), 370. https://doi.org/10.1057/s41599-025-04688-0

Mohamed, A.M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a Foreign Language teaching: Perceptions of EFL faculty members. *Education and Information Technologies*, 29(3), 3195-3217. https://doi.org/10.1007/s10639-023-11917-z

Moorhouse, B.L. (2024). Beginning and first-year language teachers' readiness for the generative AI age. *Computers and Education: Artificial Intelligence*, 6, 100201. https://doi.org/10.1016/j.caeai.2024.100201

Moorhouse, B.L., Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System*, 122, 103290. https://doi.org/10.1016/j.system.2024.103290

Nugroho, A., Putro, N.H.P.S., Syamsi, K., Mutiaraningrum, I., Wulandari, F.D. (2024). Teacher's experience using ChatGPT in language teaching: An exploratory study. *Computers in the Schools*, 1-20. https://doi.org/10.1080/07380569.2024.2441161

Pack, A., Barrett, A., Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. https://doi.org/10.1136/bmj.n71

Rahman, A., Raj, A., Tomy, P., Hameed, M.S. (2024). A comprehensive bibliometric and content analysis of artificial intelligence in language learning: Tracing between the years 2017 and 2023. *Artificial Intelligence Review*, 57(4), 107. https://doi.org/10.1007/s10462-023-10643-9

Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S., Ayala, A.P., Moher, D., Page, M.J., Koffel, J.B. (2021). PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, 10(1), 39. https://doi.org/10.1186/s13643-020-01542-z

Sharadgah, T.A., Sa'di, R.A. (2022). A Systematic Review of Research on the Use of Artificial Intelligence in English Language Teaching and Learning (2015–2021): What Are the Current Effects? *Journal of Information Technology Education: Research*, 21, 337. https://doi.org/10.28945/4999

Shen, Y., Chen, L. (2025). 'Critical chatting' or 'casual cheating'? How graduate EFL students utilize ChatGPT for academic writing. *Computer Assisted Language Learning*, 1-29. https://doi.org/10.1080/09588221.2025.2479141

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., Olson, C.B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Tai, T.Y., Chen, H.H.J. (2024). Improving elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions. *Computers & Education*, 220, 105112. https://doi.org/10.1016/j.compedu.2024.105112

Teng, M.F. (2024). ChatGPT is the companion, not enemies: EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. Computers and Education: *Artificial Intelligence*, 7, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Uygun, E., Demir, B. (2026). A PRISMA-based systematic review of artificial intelligence in English as a foreign and second language education (2023–2025). *Discover Education,* 5, 478. https://doi.org/10.1007/s44217-026-01504-y

Wang, X., Zhong, Y., Huang, C., Huang, X. (2024). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies*, 17, 1722-1736. https://doi.org/10.1109/TLT.2024.3405747

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, 14, 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

Wu, X.-Y. (2024). Artificial Intelligence in L2 learning: A meta-analysis of contextual, instructional, and social-emotional moderators. *System,* 126, 103498. https://doi.org/10.1016/j.system.2024.103498

Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: A randomized controlled trial. *Humanities and Social Sciences Communications*, 12(1), 425. https://doi.org/10.1057/s41599-025-04672-8

Yang, H., Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology*, 38(5), 180-210. https://doi.org/10.14742/ajet.7492

Has this article been screened for Similarity?

Yes

Conflict of interest

The Author's declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

About The License