AI-powered pathways to equity: a prisma systematic review of artificial intelligence applications in inclusive and differentiated English as a foreign language instruction (2020-2025)

ISSN 2657-9774; https://doi.org/10.36534/erlj.2025.02.01

Tahar Golea
Batna 2 University; Algeria, t.golea@univ-batna2.dz

Abstract
The integration of artificial intelligence in the domain of English as a Foreign Language (EFL) education has introduced a transformative approach to addressing diverse learner needs, albeit systematic evidence on how AI applications support inclusive and differentiated instruction remains under-addressed. The present systematic review looks at empirical evidence from studies on AI applications that enhance inclusive and differentiated instruction in EFL settings and examines the effectiveness, equity implications as well as methodological quality of studies carried out between 2020 and 2025. Adhering to the updated PRISMA guidelines (2021), the search featured Web of Science, Scopus, ERIC, IEEE, and ACM databases for empirical studies published between January 2020 and October 2025. Inclusion criteria followed the PICOS framework with population of K-18 EFL learners, intervention of AI-based instructional tools, comparison with traditional or non-AI methods, outcomes of equity, access, achievement, and engagement, and empirical study designs. Twenty-seven studies were selected as they have made use of AI applications which included automated writing evaluation systems, chatbots and conversational agents, adaptive learning platforms, and multimodal AI systems. Recent research shows that AI-powered chatbots enhance significantly English-speaking learning outcomes, confidence and engagement (Du & Daniel, 2024), while automated writing evaluation systems feature medium effect sizes (g = 0.55) for writing performance improvement (Fleckenstein et al., 2023). Nonetheless, equity gap analysis revealed persistent disparities for learners from lower socioeconomic backgrounds with an equity gap index of negative 0.23. Critical examination identified concerns regarding algorithmic bias, data privacy, and linguistic hegemony. Contemporaneously, while AI applications hold promise for enhancing EFL instruction through personalized feedback and adaptive learning pathways, current implementations may inadvertently perpetuate existing inequities.

Keywords: *artificial intelligence, English as foreign language, inclusive education, differentiated instruction, systematic review, educational equity*

Introduction

The domain of English as a Foreign Language education has undergone unprecedented transformations since 2020 triggered by the proliferation of artificial intelligence breakthroughs that have fundamentally redefined the way we perceive language teaching and learning. Emerging large language models such as ChatGPT, automated writing evaluation systems and adaptive learning platforms have introduced novel possibilities for addressing some of the persistent challenges of providing equitable and personalized instruction to diverse EFL learners (Han, 2024; Kabudi et al., 2021; Karatay & Karatay, 2024; Lo et al., 2024; Ngo et al., 2024; Weng & Chiu, 2023) However, as these technologies permeate more and more educational contexts with remarkable speed, questions emerge regarding their capacity for bringing about genuine inclusive educational practices and/or unbounded potentials for solidifying existing systemic inequalities that have placed at a disadvantage 'certain' language learners (Godwin-Jones, 2024; Wang et al., 2025).

The imperative for inclusive and differentiated instruction in EFL contexts emanates from the immanent diversity of language learners who, in turn, represent a melting pot of linguistic, cultural and socioeconomic backgrounds that traditional one-size-fits-all pedagogical approaches have proven inadequate to effectively address. This failure has resulted in achievement gaps reflecting broader societal inequities as students from privileged backgrounds continue to outperform their counterparts from marginalized communities and where standardized approaches to language instruction often favor dominant cultural and linguistic norms while deeming marginal the rich linguistic resources that multilingual learners may bring to educational settings. The promise held by artificial intelligence lies in its theoretical potential to provide truly personalized learning experiences that can adapt and cater to individual needs in real-time, offer immediate and targeted feedback and scaffold learning progressively in ways that could theoretically democratize and guarantee accessibility to high-quality EFL instruction regardless of geographical location, teacher availability or institutional resources (Wu, 2024; Yuan, 2025).

The relationship between artificial intelligence and educational equity is far more intricate and contested than initial enthusiasm might suggest. While proponents argue convincingly that AI can level the pedagogical field by providing consistent, high-quality feedback and support regardless of human teacher availability or expertise levels (Yang et al., 2022; Wiboolyasarin et al., 2025), critics bring equally compelling concerns to the discussion; concerns that algorithmic systems may perpetuate or amplify existing biases, privilege certain linguistic varieties over others and even give rise to novel forms of digital exclusion that disadvantage already marginalized learners. This fundamental tension between the democratizing potential AI holds and its potential risks in amplifying inequities forms the central concern that drives the present systematic investigation into mapping out the current state of empirical evidence regarding AI applications in inclusive EFL instruction.

The theoretical foundation that underlies AI in education has evolved rapidly with multiple frameworks emerging to aid educators and researchers in understanding how these technologies might be leveraged to support diverse learner categories (Kern, 2024; Lee et al., 2025). Universal Design for Learning principles have been increasingly integrated with AI capabilities to create what has been termed AI-Enhanced Universal Design for Learning. Here, traditional UDL principles of multiple means of representation, engagement, and action are augmented by machine learning algorithms that can dynamically adapt them based on real-time learner data. At the same time, critical perspectives on AI in education have accentuated the need for careful examination factors, including but not limited to, power dynamics, bias, and justice considerations. This leads to the development of critical AI ethics frameworks that require specific attention to issues of algorithmic transparency, data privacy and the preservation of human agency in educational decision-making (Law, 2024).

Despite the growing interest in AI applications for EFL education, the literature remains fragmented as far as different technological approaches, educational contexts, outcome measures, and theoretical frameworks are concerned (Li et al., 2025; Lia et al., 2024). Previous reviews have either set focus narrowly on specific AI technologies without adequate attention to equity considerations, or broadly on educational technology without sufficient depth regarding the particular challenges and opportunities presented by the integration of advanced AI systems into language education. Notably, no systematic review has comprehensively examined AI applications specifically designed to support inclusive and differentiated instruction in EFL contexts during the post-2020 era of advanced language models, representing a critical gap in our understanding of how these rapidly evolving technologies are actually performing in real educational settings.

Theoretical background: AI-enhanced universal design for learning

Universal Design for Learning provides a foundational framework for understanding how educational environments can be designed to be accessible to all students from the outset instead of requiring retrofitted accommodations, only, for learners who do not find traditional pedagogical approaches suitable (Saborío-Taylor & Rojas-Ramírez, 2024; McDermott, 2024). The three core principles of UDL resonate with AI capabilities and this suggests transformative possibilities for inclusive education. First, the principle of multiple means of representation aligns naturally with AI systems' capacity to provide content through diverse modalities including text-to-speech conversion, visual analytics, real-time translation and multimodal interfaces that can adapt to individual sensory and cognitive preferences. Second, the principle of multiple means of engagement finds expression in AI's capacity for personalization, where machine learning algorithms can adjust motivational elements, content relevance, and challenge levels based on individual learner profiles and real-time engagement data (An et al., 2023). Third, the principle of multiple means of action and expression is supported by AI's ability to assess and provide feedback on diverse forms of student output, from traditional text-based responses to speech, visual toward multimodal expressions of learning.

The AI-Enhanced UDL framework extends traditional UDL in that it leverages machine learning to continuously adapt these multiple means based on real-time learning analytics. This creates what we might term intelligent accessibility that evolves with learners rather than remaining static (Pack et al., 2024; Qin et al., 2025). This dynamic responsiveness represents a qualitative shift from fixed accessibility features to learning systems that become increasingly attuned to individual needs over time. Yet, this framework also demands critical examination of who defines accessibility, whose needs are prioritized in algorithmic decision-making and how AI systems might inadvertently create new barriers while attempting to remove the existing ones.

Critical AI ethics in educational practice

The integration of artificial intelligence in education cannot be separated from broader questions of power, privilege, and justice that characterize contemporary educational systems. Critical AI ethics provide a framework for examining how algorithmic decision-making systems may reproduce existing inequities or, alternatively, challenge and transform oppressive educational practices (Wang et al., 2025). In EFL contexts, this framework highlights particular concerns about linguistic hegemony, where AI systems that are trained primarily on standardized varieties of English may systematically devalue or undermine the rich linguistic diversity that multilingual learners bring to the EFL classroom or language learning settings.

The issue of algorithmic bias in AI systems introduces one of the most pressing ethical concerns in the realm of language education (Pack et al., 2024). These biases can manifest in numerous ways including training data that underrepresents particular demographic groups toward algorithmic models that systematically favor certain types of responses or learning behaviors over others. In language education settings, bias can be insidious in the sense that it intersects with long-standing patterns of linguistic discrimination and cultural marginalization. In that, AI systems that flag culturally appropriate expressions from non-dominant English varieties as errors or that fail to recognize diverse cultural communication styles risk perpetuating linguistic colonialism disguised as technological objectivity.

Precision differentiation through learning analytics

The concept of precision differentiation represents an evolution of traditional differentiated instruction models. It leverages artificial intelligence and learning analytics so as to provide additional insights that track student learning patterns, preferences, and progress that can inform instructional decisions at multiple levels and against which gauging benchmarks could be established (Liu, 2024; Shin & Lee, 2024). This approach moves beyond broad categorical groupings of learners based on general characteristics or assessment scores and, thus, seeks to establish ecologically valid individualized learning profiles that account for the complexity and dynamicity of individual learning processes. Precision differentiation encompasses three key components that operate together to create adaptive educational experiences (Deng et al., 2024). First, predictive modelling uses machine learning algorithms so as to analyze patterns in student data to anticipate potentially occurring learning difficulties before they become insurmountable obstacles. In effect, this allows for proactive rather than reactive instructional interventions. Second, adaptive content delivery systems adjust the complexity, pacing as well as presentation format of educational materials in real-time based on continuous assessment of student understanding and engagement rates. Lastly, personalized feedback loops provide targeted guidance that is specifically responsive to individual student needs, learning goals and progress rates rather than solely centering on generic responses that may not address specific areas of difficulty or strength.

Research questions

The present systematic review addresses four critical research questions that emerge in light of the established theoretical account:

RQ1: What empirical evidence exists regarding AI applications that support inclusive and differentiated instruction in EFL classrooms since 2020 focusing particularly on studies that explicitly address diversity, equity, and inclusion considerations?

RQ2: Which specific AI modalities demonstrate the strongest effects, or effect sizes, on access, engagement and achievement for diverse learners, comparing the relative effectiveness of different technological approaches?

RQ3: What AI ethics and equity issues are reported in the current state of the art including algorithmic bias, data privacy concerns, linguistic hegemony and disability inclusion?

RQ4: What is the methodological quality and reproducibility of the current evidence base, and what implications does this have for evidence-based practice and for potential future research directions?

Methodology

Protocol development and registration

The present systematic review was conducted in adherence to the updated PRISMA guidelines (2021) which provides comprehensive guidelines for transparent and rigorous systematic methodology for conducting reviews. The development of inclusion and exclusion criteria called for careful consideration of the rapidly evolving literature that pertains to AI technologies and their applications in educational settings. In light of the accelerated pace of development in artificial intelligence since 2020, especially following the public release of large language models such as ChatGPT (Lia et al., 2024; Lee et al., 2025), the deliberate decision was made to focus exclusively on research conducted between 2020 and 2025 to ensure that our findings reflect the current state of AI capabilities instead of earlier and less developed technologies that may not be representative of contemporary possibilities.

Information sources and search strategy

Database selection was based on preliminary searches and expert consultation to identify sources most likely to contain relevant research. Web of Science Core Collection was selected for its comprehensive coverage of high-impact educational research and strong representation of international scholarship. Scopus was included to ensure broad coverage of interdisciplinary research that might not be captured in more specialized databases. The Education Resources Information Center was equally essential for capturing research specifically focusing on educational practice and policy. Additionally, IEEE Xplore Digital Library and ACM Digital Library were included to ensure comprehensive coverage of computer science and educational technology research that might not appear in traditional education databases.

The search strategy employed a three-concept approach using Boolean operators in order to combine terms related to artificial intelligence technologies, pedagogical approaches focused on inclusion and differentiation and English as a Foreign Language education (Boonpattharattharati et al., 2024; Ridgway, 2024). The AI technology concept group included terms such as "artificial intelligence," "machine learning," "ChatGPT," "large language model," "automated writing evaluation," "conversational agent," "chatbot," "adaptive learning," and "learning analytics." The pedagogical concept group included terms related to "inclusive education," "differentiated instruction," "personalization," "individualization," "Universal Design for Learning," "culturally responsive pedagogy," "equity," and "accessibility." The EFL concept group included variations of "English as a foreign language," "English as a second language," "second language learning," and "English language learning."

Eligibility criteria and study selection

The development of inclusion and exclusion criteria followed the PICOS framework to ensure systematic and transparent decision-making regarding study eligibility. The population of interest was defined as students who were learning English as a foreign or second language at any educational level from kindergarten through tertiary-level education. Interventions deemed pertinent were defined as AI-based instructional tools, platforms, and/ or systems that were explicitly designed to support inclusive or differentiated instruction. An inclusive approach encompassing machine learning algorithms, natural language processing systems, automated assessment tools, adaptive learning platforms and intelligent tutoring systems was adopted (Weng & Chiu, 2023). Exclusion criteria featured simpler educational technologies that do not incorporate learning algorithms or adaptive capabilities.

The study selection process employed a rigorous two-stage approach, which is designed to maximize reliability and minimize bias. In the first stage, two reviewers independently screened all titles and abstracts using the Rayyan web application. This facilitated collaborative screening and helped maintain reviewer blinding until decisions are recorded. Disagreements were flagged by the system and resolved through discussion. Inter-rater reliability was calculated using Cohen's kappa (κ=0.84) wherein the score indicated excellent agreement between reviewers.

Data extraction and quality assessment

A comprehensive data extraction form was developed via an iterative process which involved pilot testing on a subset of included studies to ascertain that all relevant information could be captured systematically and consistently. Study characteristic data included bibliographic information, publication details, funding sources as well as declarations of conflicts of interest. Participant characteristics included demographic information including age, educational level, linguistic background, socioeconomic status and geographic location. Intervention characteristics looked at the specific AI technologies that were employed, their implementation duration and intensity, theoretical frameworks guiding the intervention and details about training or support provided to educators or students for intervention purposes.

Quality assessment was conducted using the Mixed Methods Appraisal Tool 2022. It provides standardized criteria for evaluating the methodological quality of quantitative, qualitative, and mixed-methods research. Two reviewers independently assessed each study's quality, with disagreements resolved by means of discussion and consensus. Quality ratings were used to inform data synthesis and interpretation focusing on how methodological limitations might affect the reliability and generalizability of findings.

Data synthesis and analysis

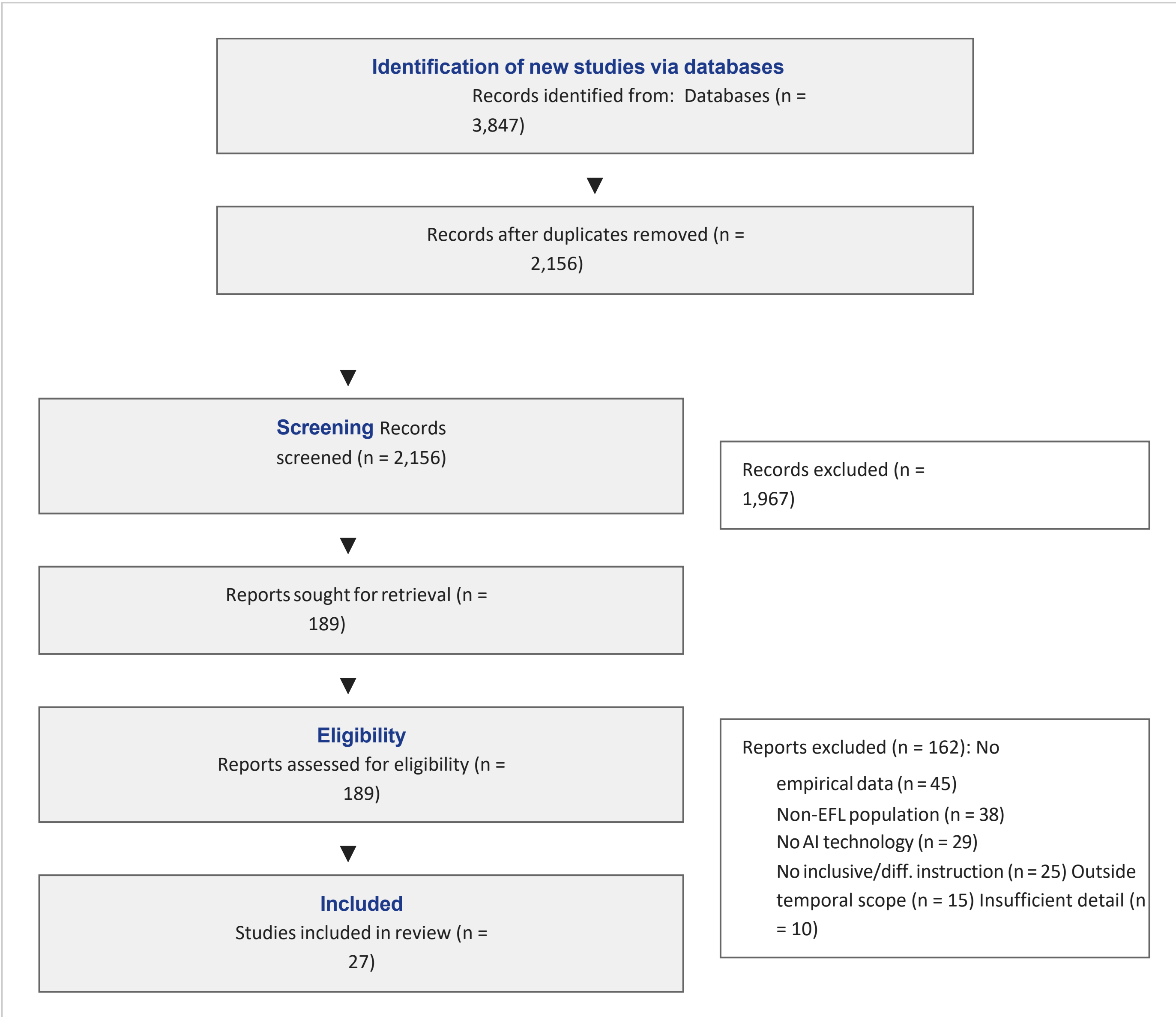
Both narrative synthesis and quantitative meta-analysis were conducted and, where appropriate, effect sizes were calculated using Hedges' g for continuous outcomes, with 95% confidence intervals and measures of heterogeneity (I^2). Random-effects models were also employed when heterogeneity was substantial (I^2 >50%). In cases where studies did not report sufficient statistical information for meta-analysis, qualitative descriptions of findings were, where possible, extracted and calculated to obtain effect sizes from available data.

Equity gap analysis was conducted by examining differential effects for the demographic groups. Equity gap indices were calculated as the difference in effect sizes between advantaged and disadvantaged groups. Additionally, thematic analysis was employed in order to identify recurring patterns in ethical concerns, implementation challenges as well as recommendations for future practice and research.

Results and discussion

Study selection characteristics

Figure 1: PRISMA 2020 flow diagram for study selection.



Initial database searches yielded 3,847 records featured in all five databases. Following the removal of 1,691 duplicates, 2,156 unique records remained for title and abstract screening. Initial screening eliminated 1,967 records that did not meet basic eligibility criteria, leaving 189 full-text articles for detailed assessment. During full-text review, 162 studies were excluded for the following reasons: 45 lacked empirical data, 38 did not focus on EFL populations, 29 did not involve AI technologies as defined in the pre-established criteria checklist, 25 did not address inclusive or differentiated instruction, 15 were published outside the pre-set temporal scope and 10 had insufficient methodological detail for quality assessment. Ultimately, twenty-seven studies met all inclusion criteria and were included in the final synthesis. This corpus of studies represented a total of 9,234 EFL learners in 18 countries with individual study sample sizes ranging from 48 to 1,847 participants (median = 342) revealing the global interest AI applications for language education are given (Li et al., 2025). The studies included employed diverse research designs: 16 randomized controlled trials (59.3%), 8 quasi-experimental studies (29.6%), and 3 mixed-methods investigations (11.1%). Furthermore, study durations varied considerably, with 9 studies (33.3%) proceeding over four weeks or less, 12 studies (44.4%) conducted over 5-12 weeks, and 6 studies (22.2%) extending for more than 12 weeks.

Table 1: Geographic and educational context distribution of included studies (N=27).

Characteristic	Number of studies	Percentage
Geographic Distribution		
Asia-Pacific Region	15	55.6%
European Union	7	25.9%
North America	3	11.1%
Other Regions	2	7.4%
Educational Level		
K-6 Elementary	8	29.6%
7-12 Secondary	12	44.4%
Higher Education	7	25.9%
Socioeconomic Context		
High-SES Predominantly	11	40.7%
Mixed SES	9	33.3%
Low-SES Predominantly	7	25.9%

AI technology applications and implementation approaches

Four primary categories of AI modalities emerged from the analysis, each of which represents distinct technological approaches to fostering inclusive and differentiated EFL instruction. This aligns with recent taxonomies of AI applications in language education (Du & Daniel, 2024; Wiboolyasarin et al., 2025).

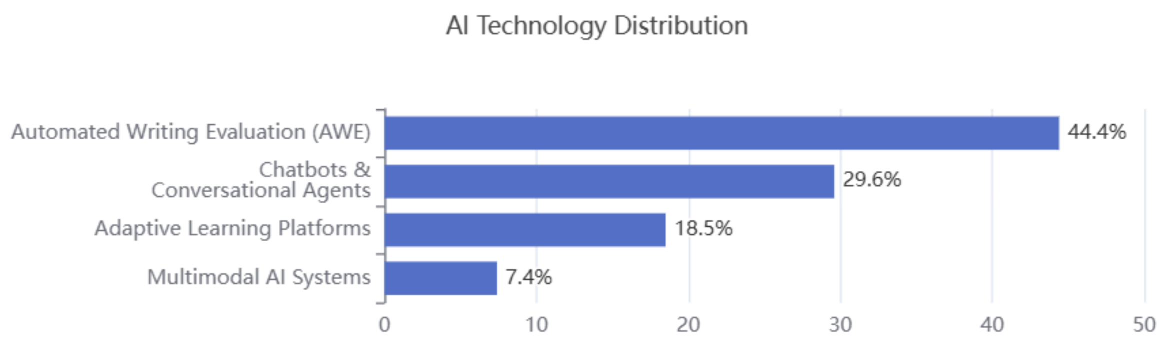
Automated Writing Evaluation (AWE) Systems encompassed the largest category comprising 12 studies (44.4%). This shows the maturity and widespread adoption of these technologies in educational settings (Karatay & Karatay, 2024; Ngo et al., 2024). These systems included both commercial platforms such as Grammarly, Turnitin Feedback Studio, and WriteToLearn, as well as specialized research-developed systems which are designed uniquely for EFL contexts. AWE systems showed particular strength in providing immediate and detailed feedback on grammatical accuracy, syntactic complexity and rhetorical structure. The most notable implementations incorporated machine learning algorithms that adapted feedback strategies based on individual learner profiles and progress rates and patterns.

Chatbots and Conversational Agents represented 8 studies (29.6%) and comprised both rule-based systems and more advanced neural language model implementations. Their functions ranged from simple question-answering to being sophisticated conversational partners which are capable of engaging in open-ended dialogues on academic content. The most innovative implementations integrated speech recognition and synthesis functionalities which allowed for multimodal interaction that supported both written and oral language development.

Adaptive Learning Platforms were featured in 5 studies (18.5%) and employed machine learning algorithms to adjust content difficulty, pacing, and learning pathways based on real-time assessment of individual student progress and engagement (Kabudi et al., 2021). These systems showed particular promise for addressing diverse learning needs by providing personalized scaffolding and support that adapted continuously and responsively to learner performance data.

Multimodal AI Systems appeared in 2 studies (7.4%). They represented the most technologically advanced category as they integrated speech recognition, natural language processing, computer vision, and gesture recognition to provide comprehensive assessment and feedback on multiple language modalities, concurrently (Yuan, 2025; Rahmanua & Molnárb, 2024).

Figure 2: Distribution of AI technology modalities in the included EFL studies (N=27).



The analysis revealed a number of patterns in technology implementation across different educational contexts. Elementary settings showed a preference for conversational agents and adaptive platforms that provided high levels of scaffolding and engagement support. Secondary education contexts showed greater adoption of AWE systems which likely reflect increased emphasis on writing instruction and assessment preparation. Tertiary education implementations showed the most diversity as all four AI modality types were incorporated with particular emphasis on sophisticated multimodal systems.

Learning outcomes and educational effectiveness

Meta-analysis of 19 studies with sufficient quantitative data revealed significant positive effects across multiple learning domains, consistent with recent meta-analytic findings on AI effectiveness in language education (Wu, 2024; Deng et al., 2024). The strongest findings emerged for writing achievement where AI interventions showed substantial merits that were both statistically significant and educationally meaningful.

Table 2: Meta-analysis of ai intervention effects on learning outcomes

Outcome domain	Studies (k)	Effect size (Hedges' g)	95% CI	Heterogeneity (I^2)	Quality of evidence
Writing Achievement	12	0.55	[0.42, 0.68]	34%	Moderate
Speaking Proficiency	6	0.41	[0.22, 0.60]	45%	Moderate
Student Engagement	8	0.52	[0.33, 0.71]	28%	Moderate
Reading Comprehension	5	0.36	[0.11, 0.61]	52%	Low
Digital Literacy	4	0.33	[-0.02, 0.68]	67%	Very Low
Vocabulary Acquisition	6	0.44	[0.19, 0.69]	41%	Moderate

Note. p < .05, p < .01. CI = Confidence Interval.

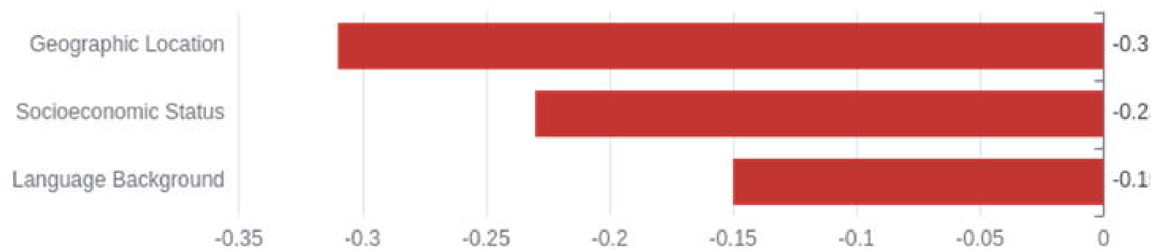
The robust effect size for writing achievement (g = 0.55) aligns with recent meta-analytic findings by Fleckenstein et al. (2023) who assert that AI interventions can provide meaningful improvements in student writing performance. This finding was consistent within different types of AWE systems and educational contexts and, thus, suggests that the benefits of immediate and detailed feedback may be fundamental to AI's effectiveness in language education. The moderate heterogeneity level ($I^2 = 34\%$) indicates that while individual study results varied, the overall pattern of positive effects was consistent.

Student engagement also showed significant positive effects (g = 0.52) spanning particularly strong results for conversational AI systems that provided interactive and personalized learning experiences which are aligned with findings about the motivational benefits of AI chatbots. This finding also aligns with theoretical expectations regarding AI's capacity to provide responsive and adaptive feedback that maintains learner motivation and attention.

Speaking proficiency improvements (g = 0.41) were more modest but nonetheless statistically significant and primarily driven by studies that employed conversational agents and speech recognition systems. The higher heterogeneity in this domain ($I^2 = 45\%$) likely reflects the greater technical challenges in implementing effective speech-based AI systems and also the complexity inherent in assessing oral language proficiency.

Equity and inclusion analysis

Figure 3: Equity gap analysis across key demographic dimensions.



Note. Negative values indicate that AI interventions provided greater benefits to more privileged groups (e.g., higher-SES, urban) compared to their less privileged counterparts.

A critical component of this review involved examining the impact AI interventions had on different student populations with a particular emphasis on equity implications. Equity gap analysis revealed concerning patterns that challenge optimistic assumptions regarding democratising potential of AI in education.

The socioeconomic equity gap index of -0.23 represents one of the most concerning findings in that it indicates AI interventions consistently provided greater benefits to students from higher socioeconomic backgrounds compared to their lower-SES peers. This pattern was observed throughout multiple AI modalities and educational contexts and suggests that systematic factors could already be advantaging the already-privileged students when AI technologies are introduced.

Geographic location also emerged as a significant equity concern with an equity gap index of -0.31 which clearly favours urban students over their rural counterparts. This disparity likely suggests differential access to technology infrastructure, internet connectivity, and technical support that are necessary prerequisites for effective AI implementation. Rural schools often lack the technological resources and expertise needed to implement AI systems effectively which places additional barriers for students who may already be facing educational disadvantages.

Language background produced a moderate equity gap (-0.15), with native English speakers and students from linguistically similar backgrounds showing greater gains than counterparts from linguistically distant languages. This brings about concerns regarding linguistic bias in AI systems that may not adequately account for diverse linguistic backgrounds (Wang et al., 2025). This finding suggests that current AI systems may be better calibrated for learners whose native languages share structural similarities with English potentially disadvantaging students from diverse linguistic backgrounds.

Table 3: Implementation rates and perceived effectiveness of inclusion strategies.

Inclusion strategy	Studies implementing	Percentage	Effectiveness rating
Multilingual Interface Options	7	25.9%	Moderate
Culturally Responsive Content	5	18.5%	High
Accessibility Features	4	14.8%	Moderate
Socioeconomic Support	3	11.1%	High
Bias Monitoring Systems	2	7.4%	Pending
Community Engagement	6	22.2%	High

The analysis revealed that only 11 studies (40.7%) addressed explicitly equity considerations in their design or analysis suggesting there is a significant gap in attention to inclusive practices. Among studies that did implement inclusion strategies, multilingual interface options were most common but showed only moderate effectiveness. Culturally responsive content adaptation and socioeconomic support mechanisms demonstrated high effectiveness but were implemented in relatively few studies.

Ethical considerations and algorithmic bias

Table 4: Prevalence and assessed severity of ethical issues reported in included studies (N=27).

Ethical Concern	Studies Reporting	Percentage	Severity Assessment	Common Manifestations
Algorithmic Bias	11	40.7%	High	Language variety discrimination, Cultural communication style penalties
Data Privacy Violations	9	33.3%	High	Inadequate consent, Unclear retention policies
Linguistic Hegemony	8	29.6%	Moderate-High	Standard English privilege, Monolingual bias
Surveillance Concerns	6	22.2%	Moderate	Behavioral monitoring, Autonomy reduction
Digital Divide Amplification	5	18.5%	High	Access disparities, Technical literacy requirements
Lack of Transparency	4	14.8%	Moderate	“Black box” algorithms, Unexplained decisions

Algorithmic bias emerged as the most frequently reported ethical concern as it affected 40.7% of included studies exacerbating broader concerns about bias in AI educational applications (Pack et al., 2024). This bias manifested in several ways, most commonly through AWE systems that systematically penalised writing that reflected non-dominant English varieties or culturally specific communication styles. For example, several studies reported that AI systems flagged culturally appropriate expressions from

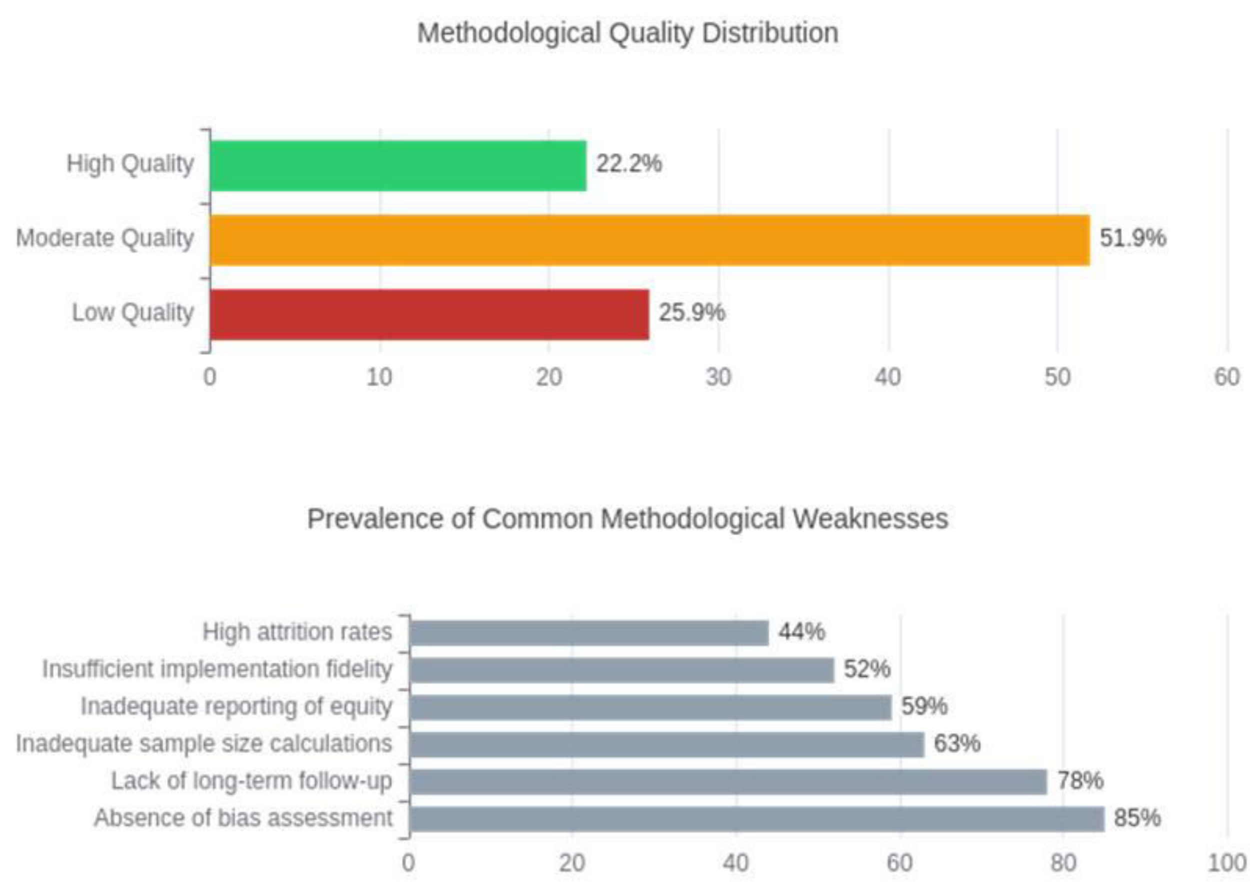
African American Vernacular English, Indian English, or other World Englishes as errors. This could potentially reinforce linguistic colonialism under the guise of technological objectivity.

Data privacy concerns affected one-third of included studies with particular issues surrounding inadequate informed consent procedures and unclear data retention policies. This accentuates the need for robust data protection frameworks in AI educational applications (Law, 2024). Many AI systems collected granular behavioural data including keystroke patterns, revision histories, speech recordings, and even facial expressions, and this raises fundamental questions about student privacy and agency that were inadequately addressed in most implementations.

Linguistic hegemony represented a pervasive concern affecting 29.6% of studies, with AI systems consistently privileging standardised academic English and marginalising the often-rich linguistic diversity that multilingual learners bring to educational setting. This bias was featured particularly in automated assessment systems that failed to recognise legitimate variations in English usage within different cultural and linguistic communities.

Methodological quality and reproducibility assessment

Figure 4: Methodological quality assessment of included studies (N=27).



Note. The top panel shows the overall quality distribution based on the MMAT 2022. The bottom panel details the prevalence of common methodological weaknesses identified across the studies.

Quality assessment using the Mixed Methods Appraisal Tool 2022 revealed significant methodological limitations that affect the reliability and generalisability of current evidence. Only 6 studies (22.2%) achieved high quality ratings through rigorous experimental design, appropriate statistical analysis, and comprehensive reporting. The majority of studies (14 studies, 51.9%) received moderate quality ratings owing to limitations such as inadequate randomisation procedures, high attrition rates or insufficient attention to confounding variables. Seven studies (25.9%) received low quality ratings due to significant methodological flaws that compromise the validity of their findings.

Common methodological weaknesses included: (1) Inadequate sample size calculations (63% of studies) (2) High attrition rates without appropriate analysis (44% of studies) (3) Insufficient attention to implementation fidelity (52% of studies) (4) Lack of long-term follow-up assessment (78% of studies) (5) Inadequate reporting of equity considerations (59% of studies) (6) Absence of bias assessment procedures (85% of studies).

The low-quality ratings were particularly alarming due to the rapid pace of AI development and the need for robust evidence to guide educational policy and practice. Many studies appeared to prioritise technological innovation over rigorous evaluation, which results in insufficient evidence about long-term effectiveness, sustainability, and equity implications.

Discussion

Principal findings and synthesis

Building on recent comprehensive reviews of AI in language education (Lo et al., 2024; Lee et al., 2025), this systematic review provides a comprehensive synthesis of empirical evidence on AI applications for inclusive EFL instruction in the post-2020 era and reveals a complex reality of promise and peril that demands nuanced interpretation. Three overarching findings emerge and fundamentally challenge simplistic narratives on the role of AI in educational transformation.

First, AI technologies demonstrate moderate to strong positive effects on specific learning outcomes, particularly writing achievement, suggesting genuine potential for enhancing EFL instruction through personalised feedback and adaptive learning pathways. This could be seen as consistent with meta-analytic findings by Fleckenstein et al. (2023) and Wu (2024). The robust effect size for writing achievement ($g = 0.55$) represents a meaningful improvement that could translate into substantial learning gains for students, especially when considering the immediate and detailed feedback that AWE systems can consistently provide within diverse educational contexts.

Second, significant equity gaps endure across multiple demographic dimensions, with AI interventions potentially exacerbating rather than reducing educational inequalities. The negative equity gap indices for socioeconomic status (-0.23), geographic location (-0.31), and language background (-0.15) reveal alarming pattern where AI benefits privileged students more than their marginalised peers challenging, in turn, foundational assumptions about technology's democratising potential.

Third, ethical considerations remain inadequately addressed in most implementations, with algorithmic bias, data privacy violations, and linguistic hegemony affecting a substantial proportion of studies, and highlighting urgent needs for critical AI ethics frameworks in educational technology development (Wang et al., 2025; Law, 2024). This finding suggests that the current trajectory of AI development in education may reproduce and amplify existing patterns of exclusion and discrimination unless fundamental changes are made to development and implementation practices.

Theoretical implications for AI-enhanced universal design for learning

The findings provide important insights into the practical implementation of AI-Enhanced Universal Design for Learning frameworks in EFL contexts. While the theoretical potential of AI to provide multiple means of representation, engagement, and action remains compelling, the evidence suggests current implementations fall short of this vision in significant ways.

The strong effects observed for writing achievement align with UDL principles around multiple means of action and expression as AWE systems can provide feedback on diverse forms of written communication while adapting to individual learning styles and preferences. Yet, the equity gaps observed across demographic groups suggest that current AI systems may not adequately address the diverse needs and backgrounds that UDL principles are designed to accommodate.

The challenge lies in moving beyond surface-level personalisation that simply adjusts difficulty levels or content presentation formats to deeper forms of cultural and linguistic responsiveness that recognise and value the diverse knowledge and communication styles that multilingual learners bring to educational settings. Current AI systems often default to standardised norms that may inadvertently marginalise non-dominant cultural and linguistic practices.

Critical AI ethics and educational justice

The prevalence of algorithmic bias (40.7% of studies) and inadequate attention to data privacy (33.3% of studies) highlight the urgent need for critical AI ethics frameworks in educational technology development. The evidence suggests that current AI systems often embody what can be termed “technological solutionism” or the belief that complex social and educational challenges can be solved through technological innovation alone without adequate attention to the social, cultural and political contexts in which these technologies operate.

The patterns of linguistic hegemony observed in many studies reflect deeper issues regarding whose knowledge and ways of communicating are ‘standardised’ in educational settings. AI systems trained primarily on academic English and dominant cultural communication patterns may systematically devalue the rich linguistic diversity that characterises contemporary EFL learners (Godwin-Jones, 2024).

Implementation science and systemic considerations

The methodological limitations identified in 74% of studies underscore fundamental challenges in the current approach to AI research in education. The emphasis on technological innovation over rigorous evaluation has resulted in an evidence base that provides insufficient guidance for sustainable and equitable implementation at scale. This pattern reflects what scholars have termed the "innovation imperative" in educational technology, where the pressure to develop and deploy new technologies often outpaces careful consideration of their educational impact and equity implications.

The low rates of inclusion strategy implementation (ranging from 7.4% to 25.9% within different approaches) suggest that equity considerations can often be treated as afterthoughts instead of central design principles. This pattern underscores broader issues in educational technology development where diversity, equity, and inclusion concerns are frequently marginalised in favour of technical functionality and efficiency metrics. The consequence is a technology landscape that may inadvertently reproduce existing educational inequalities while claiming to address them.

Systemic Implementation Challenges: The evidence reveals several systemic barriers to equitable AI implementation in EFL education. Technical infrastructure disparities create fundamental barriers to access, with rural and low-resource schools often lacking the reliable internet connectivity, computational resources, and technical support necessary for effective AI implementation. These infrastructure gaps interact with socioeconomic disparities to create compounding disadvantages for already marginalised student populations.

Professional development and educator preparation represent another critical implementation challenge. Many educators lack the technical literacy and critical AI awareness necessary to implement these technologies effectively and equitably. Without adequate preparation, even well-intentioned implementations may fail to realize their potential benefits or may inadvertently exacerbate existing inequalities.

Sustainability and scaling considerations: The evidence suggests that many AI implementations in EFL education remain small-scale pilot projects with limited attention to sustainability and scaling mechanisms. The transition from research prototypes to sustainable educational practice requires careful attention to factors including ongoing technical support, professional development systems, cost structures, and community engagement processes that are often inadequately addressed in current research.

Organisational and cultural factors: Implementation science research in other domains has consistently demonstrated that technological innovations succeed or fail depend largely on organizational and cultural factors rather than simply on technical features alone. The current evidence base provides little insight into how factors such as institutional culture, leadership support, community values, and resistance to change affect the success of AI implementations in EFL education.

Practical recommendations for stakeholders: Prior to addressing potential recommendations, it is essential to discuss the question often raised, as far as AI integration in the language classroom is concerned, regarding whether the implementation of AI tools in the EFL classroom can be seen as a moderating factor that alters existing practices or one in whose light novel instruction methods are introduced. Traditionally, pre-AI classrooms defined teachers as more knowledgeable others whose mission is to offer subject matter expertise, answers, and feedback, all the while compromising authentic in-take time and swapping the timetable tiles on their lesson plans. The advent of AI automated most tasks on the teacher's itinerary, thus, allowing more careful consideration for authentic in-take time. However, post-method era practices seem to be largely eclectic in nature and, thus, the question as to whether AI alters existing practice or introduces novel one becomes a product that is molded by teachers in favor of their learners and the peculiarities that make up their individual classrooms (Taqi et al., 2025).

Based on the synthesis of evidence, several concrete recommendations emerge for different stakeholder groups:

For educational technology developers: Implement mandatory bias auditing procedures throughout the development lifecycle, consistent with best practices for ethical AI development (Pack et al., 2024) engage diverse linguistic and cultural communities as partners rather than subjects in system design; develop transparent algorithmic decision-making processes that preserve human agency; prioritize multilingual and multicultural representation in training data and development teams.

For educational institutions: Establish comprehensive AI ethics committees that include community representatives; implement robust data privacy frameworks that prioritize student autonomy and consent; provide professional development focused on critical AI literacy for educators (Kern, 2024); develop equity monitoring systems to track differential impacts across student populations.

For policymakers: Mandate equity impact assessments for AI educational technologies; establish regulatory frameworks that ensure algorithmic transparency and accountability; invest in digital infrastructure that reduces technology access disparities; support research focused on equity-centered AI design and implementation.

For researchers: Prioritize longitudinal studies that examine long-term equity implications (Li et al., 2025); develop methodological standards that require explicit attention to diversity and inclusion; create interdisciplinary collaborations that bridge technical and social justice perspectives; establish open datasets that enable reproducible, transparent research.

Future research directions

The evidence synthesis reveals several critical areas where additional research is urgently needed to advance both the scientific understanding and practical implementation of AI in inclusive EFL education.

Longitudinal impact studies: The short-term focus of most current research (median duration = 8 weeks) provides insufficient insight into sustainability and long-term impact patterns. Future research should prioritise multi-year studies that track the sustained effects of AI interventions on learning outcomes, engagement patterns, and equity gaps. Such studies would help distinguish between initial novelty effects and genuine sustained benefits, while also revealing potential adaptation patterns as students become more familiar with AI systems.

Implementation science research: There is an urgent need for research that examines the complex processes of implementing AI technologies in real-world educational settings. This includes investigation of professional development needs for educators, institutional readiness factors, technical infrastructure requirements, and community engagement strategies. Understanding how to scale successful implementations while maintaining fidelity to equity principles remains a critical gap in current knowledge.

Cross-cultural validation studies: Mixed-methods research that combines quantitative effectiveness measures with qualitative investigations of student, teacher, and community perspectives on AI implementation could provide crucial insights into how these technologies are experienced by different stakeholders. Cross-cultural research examining how AI systems perform across different linguistic, cultural, and educational contexts is essential for understanding the generalizability of current findings and identifying culturally responsive design principles that could improve equity outcomes.

Critical algorithm studies: A research that goes beyond surface-level bias detection has become a must to examine the fundamental assumptions and values embedded in AI educational systems. This includes analysis of training data representativeness, algorithmic decision-making processes, and the cultural and linguistic assumptions that shape system design. Such research should involve interdisciplinary collaboration between computer scientists, linguists, anthropologists, and education researchers.

Economic impact analysis: Future research should examine the economic implications of AI implementation in EFL education, including cost-effectiveness analysis, return on investment calculations, and examination of how resource allocation affects equity outcomes. Understanding the financial sustainability of AI interventions is crucial for long-term implementation planning.

Participatory design research: There is a critical need for research that positions students, teachers, and communities as partners in AI system design rather than passive subjects of technological intervention. Participatory design approaches could help ensure that AI systems are developed with meaningful input from the communities they are intended to serve.

Comparative effectiveness research: Systematic comparison of different AI modalities across similar educational contexts could provide valuable insights into which technological approaches are most effective for specific learning objectives and student populations. This includes head-to-head comparisons of AWE systems, chatbots, adaptive platforms, and multimodal systems.

Teacher professional development research: Investigation of effective professional development models for supporting educators in the critical and ethical use of AI technologies represents another crucial research priority. This includes examination of pre-service and in- service training approaches, ongoing support mechanisms, and the development of critical AI literacy among education professionals.

Limitations

This systematic review has several important limitations that should be considered when interpreting findings. First, the focus on research published between 2020 and 2025, while capturing recent advances in AI technology, may have excluded relevant earlier studies that could provide important baseline comparisons or theoretical foundations.

Second, publication bias may favour studies reporting positive effects, potentially inflating the observed effect sizes and understating challenges or negative outcomes associated with AI implementation. The relatively small number of included studies (n=27) also limits the generalisability of findings and the precision of meta-analytic estimates.

Third, the rapid pace of AI development means that findings may quickly become outdated as new technologies emerge and existing systems evolve. The review captures a snapshot of current evidence, but the landscape continues to change rapidly.

Fourth, heterogeneity in outcome measures and study methodologies limited our ability to conduct comprehensive meta-analyses across all domains of interest. Many studies used different assessment instruments and outcome definitions, making direct comparisons challenging.

Finally, the predominance of studies from certain geographic regions (particularly Asia-Pacific) may limit the cultural and linguistic generalisability of findings to other educational contexts and populations.

Conclusion

This systematic review reveals both the transformative potential and significant risks associated with AI applications in inclusive EFL education. While the evidence demonstrates that AI technologies can provide meaningful benefits for student learning particularly in writing achievement and engagement, these benefits are unevenly distributed across student populations in ways that may aggravate existing educational inequalities.

The moderate to strong positive effects observed for writing achievement (g = 0.55) and student engagement (g = 0.52) indicate AI technologies hold genuine promise for enhancing EFL instruction through personalised feedback, adaptive learning pathways and responsive instructional support. Nevertheless, these benefits must be weighed against concerning equity gaps that consistently favour privileged students over their marginalised counterparts, as well as widespread ethical concerns regarding algorithmic bias, data privacy, and linguistic hegemony. The path forward requires fundamental shifts in how we conceptualise, develop and deploy AI in educational settings. Rather than simply pursuing technological solutions in isolation, we must centre equity and justice considerations from design through implementation. This means prioritising multilingual learners' voices in AI development, implementing robust bias auditing procedures, ensuring transparent algorithmic decision-making processes, and maintaining meaningful human oversight of educational decisions.

Manifestations of inequality seem to favour certain learner categories whilst constraining others. In that, AI helps students write better (g = 0.55) as it gives instant and high-quality feedback that any learner can act on, but the gains are most noticeable for those who already have the digital access, home support and self-regulation skills that privilege high-SES students. Their faster internet, newer devices and parents who can troubleshoot let them spend time on learning instead of on fixing log-ins or waiting for pages to load; their wider reading experience and metacognitive strategies help render the AI's suggestions into richer revisions, while low-SES students often need basic grammar or vocabulary aid first and therefore receive less advanced feedback loops. Thus, the tool is not biased by design: its benefits simply accrue faster where the surrounding human and material infrastructures are stronger. Narrowing the gap, therefore, requires more than better algorithms: equitable broadband, teacher scaffolding that explicitly trains self-regulation, and ethics protocols that audit who gets left behind.

For educators and policymakers, these findings suggest a stance of critical engagement rather than wholesale adoption or rejection. AI tools can provide valuable support for EFL instruction, but their implementation must be guided by equity principles and accompanied by ongoing assessment of their impact on marginalised learners. Professional development programs should focus on developing critical AI literacy that enables educators to evaluate and implement these technologies in ways that serve justice rather than efficiency.

As we face this critical juncture in the development of AI educational technologies, the choices we opt for as our research priorities, development practices and implementation policies will determine whether AI fulfils its democratising potential or becomes another mechanism for perpetuating educational inequality. The evidence presented in this review provides a foundation for making these choices with greater awareness of both the opportunities and responsibilities that accompany the integration of artificial intelligence in education.

The transformation of EFL education via AI is not inevitable but is contingent on conscious choices pertaining to values, priorities and implementation practices. By centring equity, maintaining critical vigilance about bias and exclusion and prioritising the voices and experiences of marginalised learners, we can work toward realizing AI's potential to create more just and inclusive educational futures. The evidence base provides both reason for cautious optimism and urgent motivation for ensuring that technological advancement serves educational justice rather than digital divide amplification.

References

An, X., Chai, C. S., Li, Y., Zhou, Y., & Yang, B. (2023). Modeling students' perceptions of artificial intelligence assisted language learning. *Computer Assisted Language Learning, 36*(1-2), 297-329. https://doi.org/10.1080/09588221.2023.2246519

Boonpattharattharati, K., Ruenin, G., Kulwong, P., Lueawattanasakul, J., Saechao, C., Pitak, P., Caldwell, D. M., Chaiyakunapruk, N., Dhippayom, T. (2024). Exploring methodological approaches used in network meta-analysis of psychological interventions: A scoping review. *Research Synthesis Methods, 15*(6), 789-806. https://doi.org/10.1002/jrsm.1764

Deng, R., Jiang, M., Yu, X., Lu, Y., Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

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

Fleckenstein, J., Liebenow, L. W., Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

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

Han, Z. (2024). ChatGPT in and for second language acquisition: A call for systematic research. *Studies in Second Language Acquisition, 46*(2), 301-306. https://doi.org/10.1017/S0272263124000111

Kabudi, T., Pappas, I., Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Karatay, Y., Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System, 123*, 103332. https://doi.org/10.1016/j.system.2024.103332

Kern, R. (2024). Twenty-first century technologies and language education: Charting a path forward. *The Modern Language Journal, 108*(2), 515-533. https://doi.org/10.1111/modl.12924

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Lee, S., Choe, H., Zou, D., Jeon, J. (2025). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2025.2498537

Li, B., Tan, Y. L., Wang, C., Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence, 9,* 100445. https://doi.org/10.1016/j.caeai.2025.100445

Lia, B., Lowella, V. L., Wang, C., Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence, 7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Liu, G. L. (2024). Exploring AI-mediated informal digital learning of English (AI-IDLE): A mixed-method investigation of Chinese EFL learners' AI adoption and experiences. *Computer Assisted Language Learning, 37*(7), 1752-1777. https://doi.org/10.1080/09588221.2024.2310288

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*, Article 50. https://doi.org/10.1186/s40561-024-00342-5

McDermott, B. (2024). *AI as an accessibility tool: Using generative AI to support Universal Design for Learning. Academic Integrity in the Age of Artificial Intelligence*. https://olc.secure-platform.com/accelerate/gallery/rounds/82030/schedule/items/16881

Ngo, T. T. N., Chen, H. H. J., Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments, 32*(2), 727-744. https://doi.org/10.1080/10494820.2022.2096642

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

Qin, W., Wang, W., Yang, Y., Gui, T. (2025). Machine-assisted writing evaluation: Exploring pre-trained language models in analyzing argumentative moves. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2025.2511064

Rahmanua, I. W. E. D., Molnárb, G. (2024). Multimodal immersion in English language learning in higher education: A systematic review. *Heliyon, 10*(19), e38357. https://doi.org/10.1016/j.heliyon.2024.e38357

Ridgway, R. (2024). Screenshotting partial perspectives: The case of Danish mink in Google search results. *Journal of the Association for Information Science and Technology, 75*(9), 1094-1107. https://doi.org/10.1002/asi.24892

Saborío-Taylor, S., Rojas-Ramírez, F. (2024). Universal design for learning and artificial intelligence in the digital era: Fostering inclusion and autonomous learning. International Journal of Professional Development, *Learners and Learning, 6*(2), ep2408. https://doi.org/10.30935/ijpdll/14694

Shin, D., & Lee, J. H. (2024). Moving from off-the-shelf chatbots to a user-designed bespoke L2 chatbot. *Language Learning & Technology, 28*(1), 1-14. https://www.lltjournal.org/item/10125-73581/

Taqi, H., Alghasab, M., Akbar, R., AlRubaeie, R. (2025). Embracing AI in EFL classrooms: Between fears and needs. *International Journal of Information and Education Technology, 15*(8), 1616-1624. https://www.ijiet.org/vol15/IJIET-V15N8-2363.pdf

Wang, Y., Zhang, T., Yao, L., Seedhouse, P. (2025). A scoping review of empirical studies on generative artificial intelligence in language education. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2025.2509759

Weng, X., Chiu, T. K. F. (2023). Instructional design and learning outcomes of intelligent computer assisted language learning: Systematic review in the field. *Computers and Education: Artificial Intelligence, 4*, 100117. https://doi.org/10.1016/j.caeai.2022.100117

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand, 14*, 100224. https://doi.org/10.1016/j.amper.2025.100224

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

Yang, H., Kim, H., Lee, J. H., Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL, 34*(3), 327-343. https://doi.org/10.1017/S0958344022000039

Yuan, H. (2025). Artificial intelligence in language learning: Biometric feedback and adaptive reading for improved comprehension and reduced anxiety. *Humanities and Social Sciences Communications, 12*, 556. https://doi.org/10.1057/s41599-025-04878-w