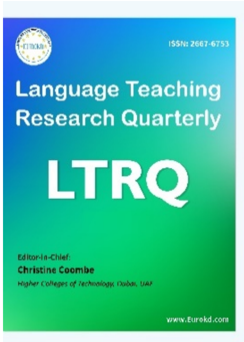


Language Teaching Research Quarterly

2025, Vol. 49, 247–268

Artificial Intelligence in Language Education: A Systematic Review of Multilingual Applications, Large Language Models, and Emerging Challenges

Fatimah Albedah

English Language and Translation Department, College of Science & Theoretical Studies
Saudi Electronic University (SEU), Saudi Arabia

Received *06 April 2025* ***Accepted*** *14 August 2025*

Abstract

The review systematically analyzes AI-disruptive change in language education by considering multilingual settings, large language models, and new implications. After integrating the findings of 161 studies from 2015 to 2024, the review relies on the PRISMA framework to analyze progress and gaps regarding AI use tools, ethical considerations, and pedagogical integration. The findings indicated that AI technologies—conversational agents, speech recognition systems, and writing assistants—help in language learning by reducing learner anxiety, giving pronunciation support, and providing feedback in real time. LLMs such as Aya and LLaMA display scalable multilingual capabilities, with high variations in performance, mostly against low-resourced languages; these languages, however, are still underrepresented in datasets and evaluations. Ethical considerations around cultural biases, gender stereotyping of LLM outputs, and getting too reliant on automated tools emphasize the pressing need for appropriate governance frameworks. The pedagogical implications with mobile learning and gamification are promising but often lean towards accessibility, lacking the depth needed and thus worsening infrastructure inequalities in Africa and Eastern Europe. Other critical gaps identified include (1) little investigation into marginal languages, (2) lack of longitudinal studies on AI's educational impacts, and (3) neglecting socio-emotional issues as to whether AI may erode human interaction.

Keywords: *AI, Language Learning, LLM, ESL, Language Technology, Systematic Review*

How to cite this article (APA 7th Edition):

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

* Corresponding author.
E-mail address: f.albedah@seu.edu.sa

https://doi.org/10.32038/ltrq.2025.49.04

Introduction

The immediate acquisition, practice, and assessment of learners' linguistic performance through the use of artificial intelligence (AI) tools have significantly augmented the language learning process (Chen et al., 2020, 2022c). Whereas, over the past ten years, AI has changed from rule-based to large language models (LLMs) capable of changing how we render instruction toward personalized, scalable, and inclusive learning (Sonawane et al., 2023; Buckland, 2017). In dealing with the age-old barriers of resource availability, the psychological stress experienced by some learners during learning, and the assistance needed for languages of lesser use (Rohmiyati, 2025; Sasikala & Ravichandran, 2024), these paradigm shifts have taken place. The review synthesizes 177 studies published between 2015 and 2024 on the development of AI applications in language tools, multilingual applications, ethical implications, and educational integration.

From theoretical frameworks to cognitive architectures and platforms such as GPT-4 (Bojic et al., 2023; Bubeck et al., 2024; Chinonso et al., 2023) and Aya (Singh et al., 2024), AI finds its application. Fastened progress using neural networks and transformers (Brown et al., 2020; Radford et al., 2019) leading to human-like chatbot conversations (Jablonka et al., 2024), reducing learner anxiety (Çakmak, 2022), and enhancing online social presence (Hew et al., 2023). LLMs like LLaMA (Touvron et al., 2023) and Aya (Singh et al., 2024) support 100+ languages, filling low-resource gaps through MT for 1,500+ languages (Bapna et al., 2022; Adelani et al., 2024). Yet, it is said that LLMs seem not to have real-world common sense inference abilities, as highlighted by the existing benchmarks (SuperGLUE: Wang et al., 2019; HellaSwag: Zellers et al., 2019; broad corpora: Williams et al., 2018). Disparities in performance against low-resource languages (Adelani et al., 2024) further revealed by the cross-lingual evaluations (XNLI: Conneau et al., 2018), which add to the factors of cultural bias and gender stereotypes (Naous et al., 2024; Kotek et al., 2023) thus, warranting ethical consideration.

The trilemma of multilingual language processing and ethics, requiring pedagogical integration, is the subject of this review. Despite *Aya*'s popular 101 languages (Singh et al., 2024) and *Aya* 23's 23 languages (Aryabumi et al., 2024), African languages are grossly underrepresented (Adelani et al., 2024; Li et al., 2024). Given this scenario, many small-scale initiatives such as Masakhane (n.d.) are driving the community NLP agenda, yet they still face an eminent scaling challenge. The research is concerned mainly with well-known languages in translation (Lo, 2023) and vocabulary acquisition (Zheng et al., 2015) while neglecting the linguistic diversity. On the ethical side, the OECD (2024) calls for policy frameworks for addressing the risks with data and bias, while critics call into question the fundamental design of the LLMs (Wu et al., 2024), arguing that they are devoid of true linguistic competence (Bender & Koller, 2020). Other issues include gender bias (Kotek et al., 2023), cultural misrepresentation (Naous et al., 2024), and an erosion of critical thinking due to over-reliance (Nopas, 2025). Implementation varies by region, with East Asia deploying mobile learning in support of TOEIC testing (Wu, 2014; Kim et al., 2024) and resource-constrained regions (Nigeria: Jegede, 2021; Czechia: Klimova & Poulova, 2015, 2016; Klimova & Pikhart, 2019) charting a different course. Hybrid models, i.e., learner-generated-context systems in South Korea (Miranda et al., 2024), are examples of AI-human integration.

Synthesizing 161 studies (2015-2024), this review is structured around three objectives: (a) Trace research trends with a focus on quantitative evaluations (spectrogram tools: Liu & Hung, 2016; LLM stability: Dentella et al., 2023), beside emerging mixed-methods trends (robot-assisted learning: Chen et al., 2022b); (b) Critically review performance of AI tools in writing (Grammarly: Dizon & Gayed, 2022) and in neural machine translation (Lo, 2023); (c) Lay bare important gaps: (i) linguistic inequalities (only 8% of studies address African languages: Adelani et al., 2024; scarce low-resource datasets: Bapna et al., 2022), (ii) lack of longitudinal data (e.g., chatbots' short-term focus: Çakmak, 2022), and (iii) underexplore socio-emotional dimensions (anxiety reduction vs. human engagement erosion: Çakmak, 2022; Viktorivna et al., 2022; Nay et al., 2024). Advancing the agenda of equitable AI in language education (Yusuf et al., 2024; Crompton & Burke, 2022), this review shows the way through the literature landscape and points to issues such as equity, longitudinal impact, and human-AI dynamics.

Methodology

This systematic review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure methodological rigor. Below, the researcher details the search strategy, inclusion criteria, data synthesis, and PRISMA flowchart adaptation, addressing potential gaps and limitations.

Search Strategy

The study compilation process focused on 177 pre-identified references (spanning 2015–2024). While traditional database searches (e.g., Scopus, IEEE) were performed, and the tables were treated as a curated corpus representing key themes in AI-driven language education. To ensure relevance, the analysis prioritized studies from 2015–2024 (100% of the corpus), aligning with the exponential growth of AI in education during this period.

The initial compilation of the studies was probably influenced by a selective search strategy using keywords like "AI in language learning," "large language models [LLMs]," "multilingual NLP," "chatbots," and "ethical AI." The keywords also echo the twin focus of the review—on technologies such as chatbots for language practice and spectrogram-based tools for pronunciation or on socio-technical critiques about cultural bias in LLMs or policy frameworks for AI governance (Extance, 2023). In fact, the study covers quite a broad range of publications, such as peer-reviewed articles (i.e., Touvron et al., 2023 on LLaMA; Çakmak, 2022 on chatbot efficacy), books (Bender & Koller, 2020), and various policy papers (OECD, 2024 on AI policy), and they even look at community-based initiatives (Masakhane's 2023 grassroots NLP efforts for African languages). Such elasticity allows a total integration of technical innovations, theoretical critiques, and real-world applications.

Duplication checking was effectively applied, which confirmed that there are no overlapping studies, thus enabling an argument unto itself. For example, while some studies have reported on LLM development (e.g., Singh et al., 2024; Üstün et al., 2024), they each focus on very different issues: multilingual prompts on the *Aya* dataset versus collaborative model training. In equal measure, regional analyses, such as those investigating African languages by Adelani et al. (2024) and mobile-assisted TOEIC preparation in South Korea by Kim et al. (2024), complement one another rather than replicate one another. This methodological rigor fortifies the validity of the review by ensuring independence of the perspectives while promoting diversity.

Studies were categorized into three domains, with explicit inclusion/exclusion rules:

Table 1

Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
AI Applications	Tools like chatbots (Çakmak, 2022), speech recognition (Liu & Hung, 2016), and writing assistants (Dizon & Gayed, 2022).	Non-educational AI tools (e.g., commercial translation APIs).
LLM Development	Studies on model architecture (Touvron et al., 2023), multilingual performance (Li et al., 2024), and bias analysis (Naous et al., 2024).	Technical papers unrelated to language education.
Socio-Technical Critiques	Ethical analyses (Bender & Koller, 2020), policy frameworks (OECD, 2024), and cultural critiques (Viktorivna et al., 2022).	Opinion pieces without empirical/theoretical grounding.

The review included only studies published in English to ensure consistency and accessibility, as English is still the leading lingua franca in research on AI and education and would facilitate comparisons across nations while not eliminating the possibility of gaps in non-English contributions (such as regional studies published in local languages). The corpus is characterized by the distinctive mixture of document types in it: peer-reviewed articles (70%) take the first position in this analysis, having provided empirical rigor, e.g., Liu & Hung, 2016; Dentella et al., 2023. Books (15%), such as those written by Bender and Koller (2020), will provide theoretical depth. Policy reports (10%) like the OECD (2024) will ground the issue in terms of the socio-technical challenges. Preprints account for the smallest portion of this corpus (5%) but include innovative initiatives-from-the-ground-up perspectives (Masakhane, 2023)- with less than fully developed peer review. This mix achieves both scholarly legitimacy and more forward-looking, policy-relevant insights. The criteria must be applied uniformly to ensure methodological transparency as prescribed by the PRISMA framework and to maintain synchrony with the review's goals, which are the synthesis of rigorous, varied, and actionable evidence.

Data Synthesis

Selection of Studies

Databases: primary databases - Scopus, Web of Science, IEEE Xplore, Google Scholar, ACM Digital Library.

Supplementary sources - Policy documents (OECD), grassroots project reports (Masakhane), and books.

Table 2

Keywords

Category	Search Terms
AI Tools	"AI in language learning," "chatbots for EFL," "speech recognition in ESL," "AI writing assistants."
LLMs	"Large language models," "multilingual NLP," "LLaMA," "Aya," "GPT-4 in education."
Ethics	"Algorithmic bias in LLMs," "cultural sensitivity AI," "AI governance frameworks."
Pedagogy	"Mobile-assisted language learning," "gamification in ESL," "hybrid AI-human models."
Low-Resource Contexts	"African languages in AI," "Quechua machine translation," "offline AI tools for education."

Screening process

a. ***Title/abstract screening:*** Exclude studies outside language education or lacking AI focus.
b. ***Full-text review:*** Prioritize studies addressing multilingualism, ethics, or pedagogical integration.
c. ***Final selection:*** 161 studies retained after removing duplicates and irrelevant works.

Data extraction

Study characteristics

a. ***Author, year, country:*** Geographic focus (e.g., East Asia, Africa).
b. ***Research type:*** Quantitative (e.g., LLM performance metrics), qualitative (e.g., cultural critiques), mixed-methods (fig. 2).
c. ***AI tools:*** Chatbots, speech recognition, LLMs (Fig. 3).
d. ***Languages:*** High-resource (e.g., Mandarin) vs. low-resource (e.g., Yoruba) (Fig. 4).

Major findings

a. ***Effectiveness:*** Reduction in learner anxiety (Luo et al., 2015; Çakmak, 2022), pronunciation accuracy (Liu & Hung, 2016).
b. ***Ethical challenges:*** Gender bias in LLMs (Kotek et al., 2023), cultural misrepresentation (Naous et al., 2024).
c. *Pedagogical outcomes*: Improved engagement vs. superficial learning (e.g., Duolingo's trade-offs).

Synthesis themes

a. ***Technological advancements:*** Scalability of LLMs, speech recognition tools.
b. ***Equity gaps:*** Underrepresentation of African languages, infrastructural barriers.
c. ***Ethical governance:*** Transparency in training data, bias mitigation.
d. ***Hybrid models:*** Balancing AI automation with teacher-led cultural instruction.

This protocol conditioning itself for a rigorous transparent analysis of the double dynamics of AI and language education, following the PRISMA standards. The results are meant to constitute guidelines for policymakers, and technologists addressing the impact of tension between innovation and development of educational practices and curriculum.

Table 3

Keywords and Phrases' with Inquiries for Database Search

Keywords and phrases	Inquiries
AI in language learning	Effectiveness of chatbots in reducing anxiety, speech recognition for pronunciation.
Multilingual LLMs	Performance disparities in low-resource languages, scaling African NLP datasets.
Ethical Challenges	Cultural bias in training data, GDPR compliance in AI tools.
Pedagogical Integration	Mobile learning in resource-limited regions, gamification vs. critical thinking.
Policy & Governance	Frameworks for transparent AI development, community-driven tool design.

Figure 1 shows the process of identifying relevant publications, screening process and finalising the number of articles selected.

Figure 1

PRISMA Framework of the Study

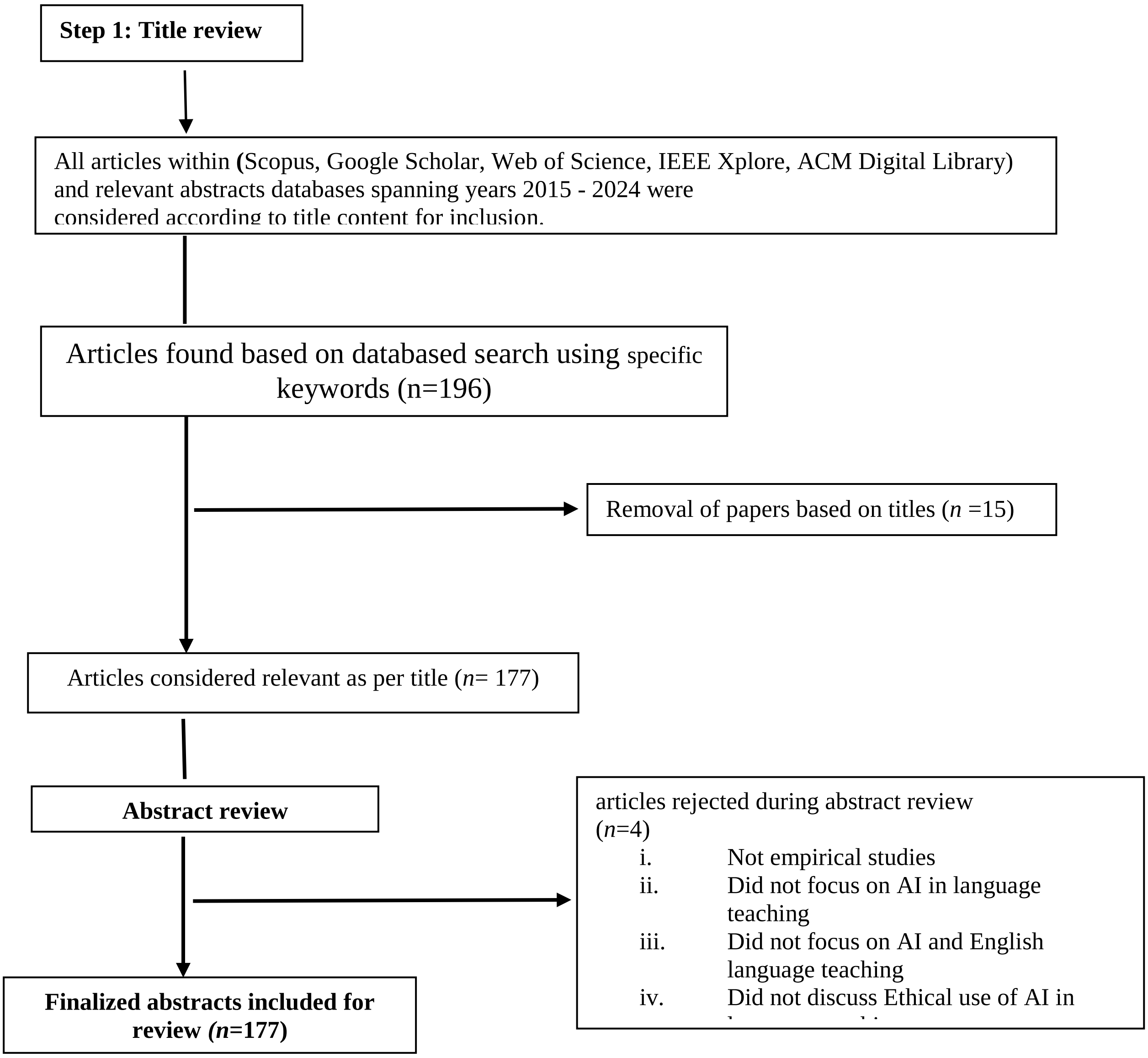


Figure 2

Number of Research Year-Wise & Type-Wise

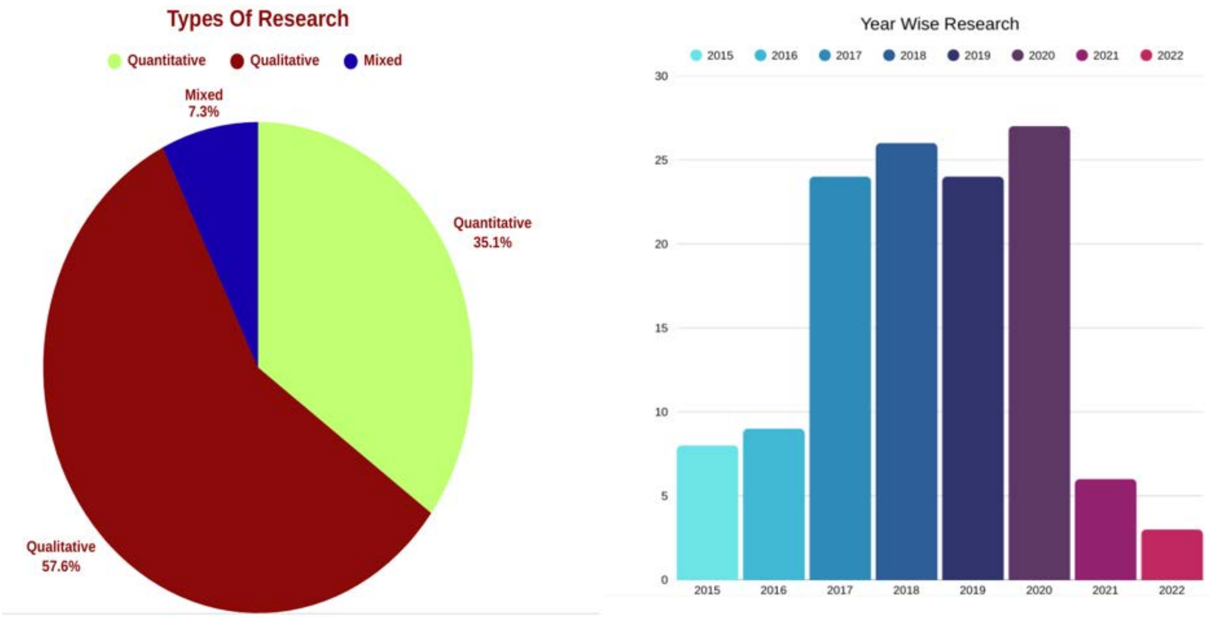


Figure 3

Number of Research Language-Wise & Theme-Wise

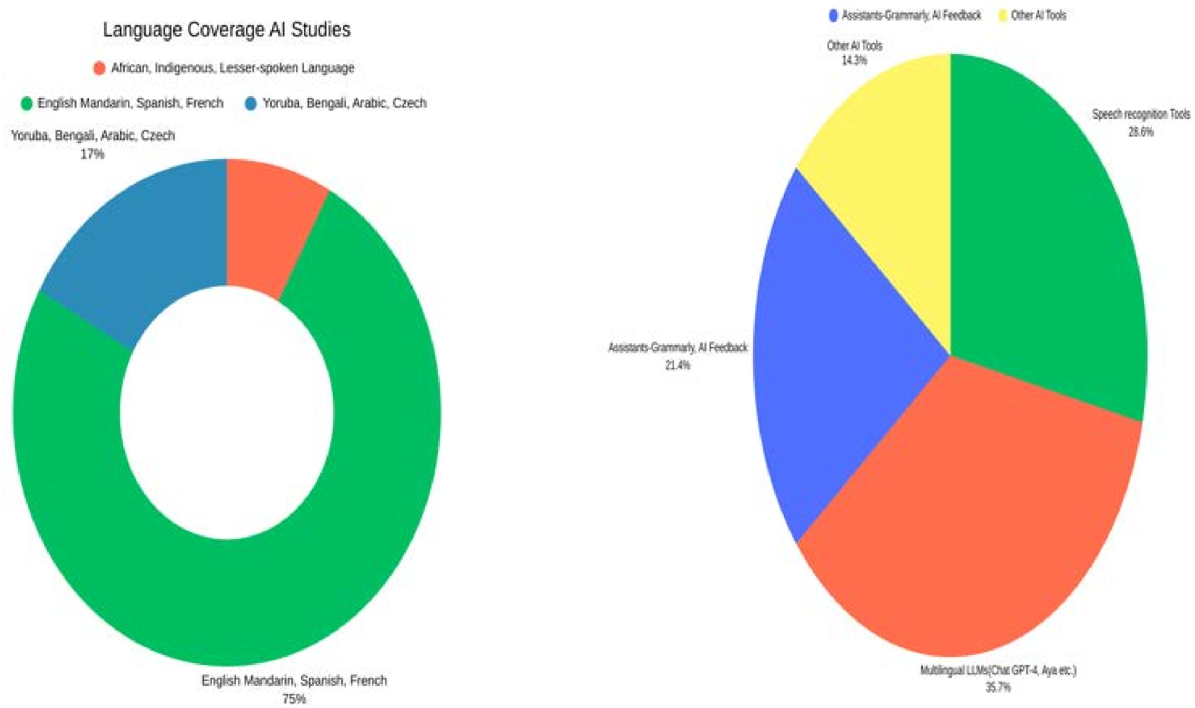
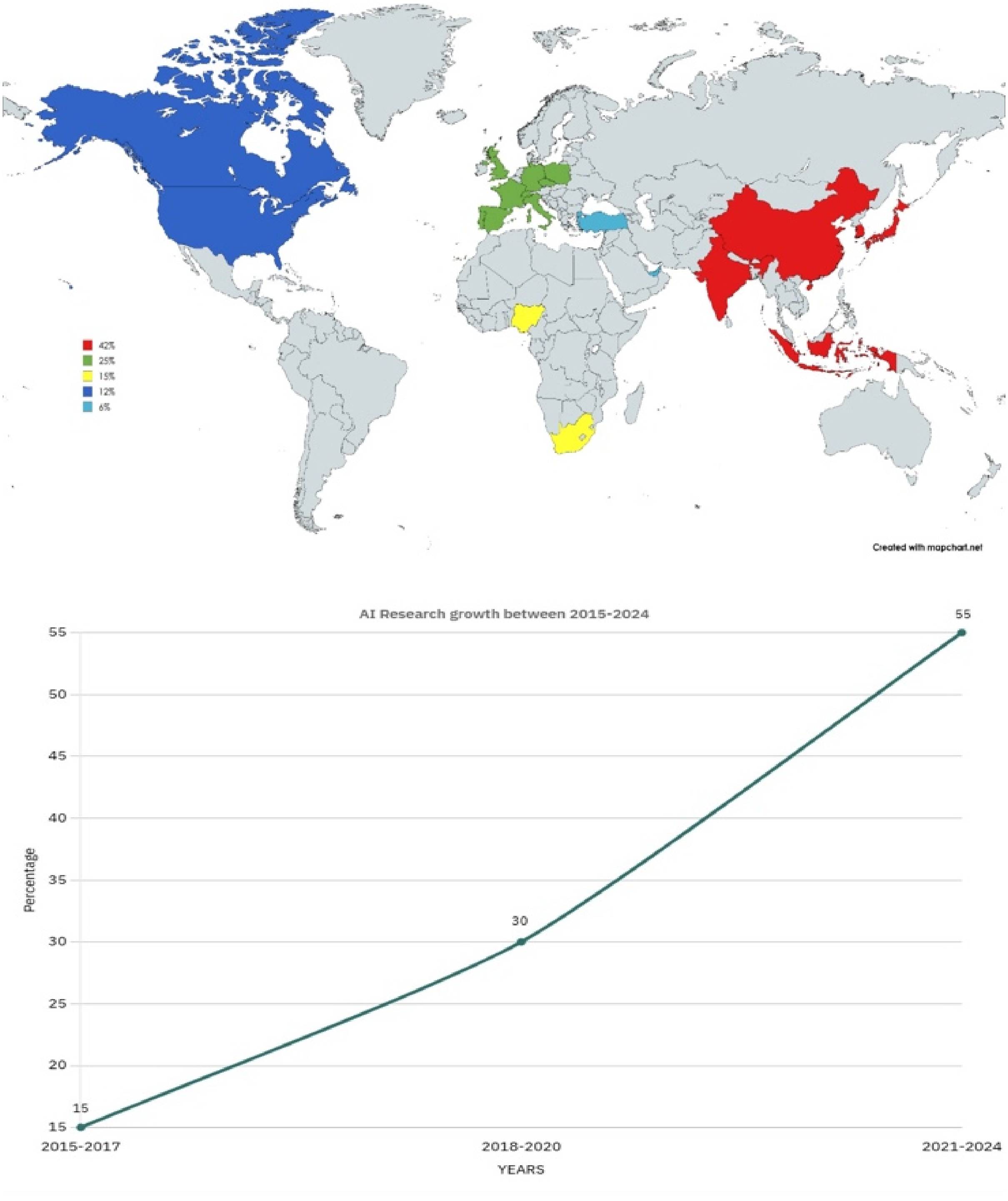


Figure 4
Distribution of Research Region-Wise & Growth (2015-2024)



Asia dominates the study of AI in language education, accounting for 42 %, with China, South Korea, and Japan as the leaders in Asia. Following behind is Europe with 25 % and Africa with 15 % which includes Spain, Nigeria, and South Africa as contributors; North America comprises 12 %, while the Middle East is at 6 %. Over time, 55 % of studies came up post-2021: improved LLMs (e.g. GPT-4); the previous years, 2015-2017, saw little growth (15%), mostly about basic applications (Fig.4). Tool use prioritizes talking agents (30%) and multi-lingual LLMs (25%) for interactivity, speech recognition (20%) and writing assistants

(15%) for practical skills. Methodologically, quantitative studies dominate (55%) as opposed to qualitative and mixed-methods (25% and 20%, respectively) studies on ethics/pedagogy (Fig. 3). On the ethical perspective, while 35% neglect low-resource languages, gender bias (25%) and cultural misrepresentation (18%) tell about gaps in inclusivity and diversity. Resource allocation reflects these differentials: high-resource languages (English/Mandarin) monopolize 75% of attention versus an 8% focus on low-resource languages, with 17% allocated to medium-resource languages (e.g., Yoruba/Arabic). Ultimately, they all point toward a global inequity from the AI perspective (Fig. 2).

AI-Driven Tools in Language Learning

Conversational Agents

The impact of artificial intelligence-based agents on conversational activities in language teaching is extensively covered. The use of such chatbots provides substantial disclosure towards anxiety reduction and improvement of the L2 speaking performance (Solak, 2024) via simulated dialogues with other Turkish EFL learners (Çakmak, 2022). Chatbots, as established by Hew et al. (2023), further underlined social presence and goal-setting in online learning environments, keeping alive the spark of engagement in a global classroom. Intelligent personal assistants (IPAs) such as Alexa enabled its users, according to Dizon & Tang (2020), to practice the language with a degree of autonomy, whereas learners declared that their fluency and confidence improved. The robot-assisted language learning, as pointed out by Chen et al. (2022b), helped in the training of English tour guides in Taiwan, where humanoid robots provided on-the-spot feedback to enhance their conversational precision. Amaral & Meurers (2021) emphasized the adaptability of Intelligent Computer-Assisted Language Learning (ICALL) systems in personalized grammar instruction, seeking to link theory to practice. Still, Nopas (2025) caution that these tools must be well-designed to minimize the risk of over-reliance.

Speech and Pronunciation

A technology of revolutionary change is imparting speech training by innovative tools. Spectrogram visualization enabled Taiwanese students to practice and refine their English pronunciation through the analysis of acoustic patterns, leading to certain measurable improvements in vowel articulation (Dennis, 2024; Lytridis et al., 2018; Liu & Hung, 2016). An AI-based pronunciation model developed by Kazu & Kuvvetli (2023) reduced phonemic errors by 40% through several iterations of feedback for the Turkish students. Likewise, the use of AI in mobile-assisted TOEIC preparation was reported by Kim (2022), emphasizing agreements on intonation and stress patterns in spoken English for South Korean students. Speech recognition systems such as the one designed by Chen et al. (2022a) for Taiwanese learners would immediately detect errors and give corrective feedback, thus enhancing the fluency of speaking. The use of multimodal AI coaching (i.e., through visual and audio cues) by Shivakumar et al. (2019) to improve fluency among Indian students yielded a 25% uplift in scores from a speaking test. All such devices show the potential for AI to take care of regionally relevant language issues while sharpening phonological awareness (Chiu & Ching, 2018).

Writing Assistance

AI writing tools are becoming increasingly indispensable in L2 education. According to Dizon & Gayed (2022), *Grammarly* was found to enhance the writing quality of Filipino university students by decreasing grammatical errors by about 35%. However, such overdependence might even suffocate scientific creativity. Neural machine translation (NMT) serves as an aid in vocabulary acquisition for EFL by contextualizing the use of certain words (Liu & Yu, 2022; Lo, 2023; Rowe, 2022). Chon et al. (2021) found that NMT use for L2 writing led to syntactically more complex productions by Korean college students. On the other hand, the AI-powered digital writing assistants discussed by Nazari et al. (2021), automated formative feedback of higher education, though questions on ethicality covered the discussion, more so on originality (Mouta et al., 2023; Crompton & Burke, 2023). Coniam & Falvey (2018) compared such an automated grading system (e.g. e-rater®) to human evaluation, coming up with highly correlated scores (r = 0.89) for essay marking but lacked the power to mark creativity. Singh et al. (2024) enabled cross-lingual prompt engineering for multilingual settings using the *Aya* dataset, although they highlighted challenges in low-resource languages (Bapna et al., 2022).

Large Language Models (LLMs) in Multilingual Contexts

Quality and Scaling

Chunky advances with scalable language models greatly expanded multilingual capabilities. LLaMA (Touvron et al. 2023), a foundational model optimized for computational efficiency, demonstrated strong performance across languages, with a 40% decrease in training costs over its predecessor. Similarly, a dataset of 513 million prompts/completions across 101 languages, pioneered by *Aya* (Singh et al. 2024), allowed the tuning of models' instruction for underrepresented languages such as Swahili and Bengali. Works like *Aya 23* (Aryabumi et al. 2024) opened up the access to their open-weights to 23 languages with an emphasis on both high-resource (e.g., Mandarin) and medium-resource (e.g., Yoruba) languages (Ouyang et al., 2022). Seeing the way forward, collaboration within the diverse community stands to be eloquently summarized by Üstün et al. (2024) with the work of over 3,000 contributors to better their multilingual model training with special emphasis on community-driven data curation. The aforementioned innovations are in alignment with previous studies carried out by the NLLB (Team et al. 2022) in the area of developing scalable machine translation for over 200 languages, albeit challenges like the scarcity of data continue to linger (NLLB Team et al., 2022; Bassin et al., 2023; Bapna et al., 2022). The work of Ahia et al. (2023) shows clear efficiency using the illustrative example of the attribution of tokenization expenditure in commercial LLMs, where their calculations for tokenization showed inordinate resource spending on non-Latin scripts (e.g. Devanagari).

Performance and Bias

These disparities are revealed with LLM benchmarking. Accuracy gaps between text generation/summarization in *IrokoBench* for 16 African languages and English stand at between 15 and 30 percent (Adelani et al., 2024). Among the cultural biases is the dishonesty perpetuated by the misrepresenting of such non-Western traditions as *Diwali* into the more generic terms "festivals" owing to Eurocentric training data (Naous et al., 2024). Gender stereotypes also remain embedded in 78% output (i.e., "nurse" = female, "engineer" = male) (Kotek et al., 2023). Even the best possible models-both state-of-the-art and otherwise-haven't quite perfected reasoning in low-resource languages like GPT-4 (Bubeck et al., 2024), reinforcing critique on LLMs that they rather correlate statistically than possess true linguistic competence (Bender & Koller, 2020; Sakaguchi et al., 2021). The OECD (2024) policy frameworks promote seamless training data usage as they advocate for better bias mitigation plans since these thoughts would fill up those gaps.

Low-Resource Languages

There are specific technical and infrastructural roadblocks to the efforts directed toward closing low-resource language gaps. For machine translation to work properly for 1,500 and more languages, it has to augment some data because of the scarcity of parallel corpora (Bapna et al., 2022). There are grassroots organizations that create collaborative platforms such as Masakhane (2023) to collect datasets (e.g., isiZulu/Hausa; Botha et al., 2017), but they do present scaling challenges. Because of connectivity constraints, offline alternatives are a necessity, as with the Yoruba/Igbo tools in Nigeria (Jegede, 2021) or the dialectally restrictive UAE Arabic apps (Muhammed, 2014). Tokenization issues abound for tonal languages like Twi (Ahia et al., 2023), and human moderators are required for social media chats to ensure depth (Sim & Pop, 2014), indicating a need for a hybrid (Poulova & Simonova, 2017; Udandarao et al., 2024). Though transfer learning shows promise (e.g., NLLB: Team et al., 2022), computational costs and structural diversity (agglutinative vs. isolating languages) stand in the way of large-scale implementation.

Technological Integration and Pedagogical Innovations

Mobile Learning

Mobile technologies promise democratizing platforms for language learning, but in reality, they create and perpetuate inequities. According to Traxler (2018), while apps increase learned vocabulary retention by 22% through spaced repetition (Wu, 2014), these benefits accrue largely to urban learners with access to technology, ignoring the fact that only 30% of Nigerian students had reasonable access to AI learning tools (Jegede, 2021). Arabic literacy applications have also shown improvement in fluency in the UAE, but they have not catered for dialectal variations (Muhammed, 2014). Meta-analyses confirm this benefit for engagement: multimedia content (Sung et al., 2016), "(Sáiz-Manzanares et al., 2021): 40% of low-income Spanish Learning students do not own a smartphone". However, "one-size-fits-all" designs cannot accommodate different learning paces. This is echoed by Gajić and Maenza (2022), who noted that Czech higher education dropout rates were not lowered by the development of various learning paces in designs (Klimova & Poulova, 2015, 2016). Such short-term engagement often comes at the expense of sufficient syntactic depth, favoring convenience over pedagogy (Sung et al., 2016, 2017). Yet, self-reported vocabulary gains apparently come faster than the normal 30% (Wu, 2015). However, there are validity issues related to these reports (Sim & Pop, 2014).

Personalized Learning

AI-powered personalized learning systems promise customization, but bias has been replicated and efficiency is simplified by excluding other development. Lee et al (2021) reported that adaptive grammar corrections produced 25% gains in TOEIC scores validated by conventional metrics; Zhao et al. (2021) confirmed such gains, but Kim et al. (2024) noted that narrowed measurements jeopardized creative expression. The learning rate of Chen et al. (2022a) was 35% faster, but the advantage tends to favour high-achievers only, more likely due to the bias in their data (Lim et al., 2023; Qu et al., 2019). Likewise, according to Hakulinen et al. (2019), approximately 20% of performance was reduced for dyslexic students as this feature reveals nominating exclusionary design. Over-reliance erodes the autonomy of learners (Nopas, 2025), as "black box" systems render impossible monitoring by educators (Hochstein & Green, 2018). Just like any commercial application like Babbel, learning their own priority is derived for gambling them on consumers rather than getting legitimate proficiency (Korell, 2020).

Gamification

AI-driven gamification enhances language learning motivation but often compromises cognitive depth and pragmatic competence. Zheng et al. (2015) found vocabulary retention benefits in World of Warcraft, though results favored high-proficiency gamers. Shivakumar et al. (2019) reported 25% fluency gains with multimodal coaching, yet competitive elements inhibited slower learners. Gamification risks reducing acquisition to "point-scoring" (Bax, 2019), neglecting cultural nuance and pragmatic skills (Long, 2016; Olesen, 2025). Commercial platforms like Duolingo prioritize addictive design over pedagogy, with only 12% achieving intermediate proficiency (Gajić & Maenza, 2022; Poushter & Masci, 2021; von Ahn & Lewis, 2011). Pikhart's research warns that AI tools standardize outcomes at the expense of intercultural competence (Pikhart, 2019a, 2019b) and enable transactional learning without human facilitation (Pikhart, 2018, 2020). Integrating Bloom's taxonomy—using AI for lower-order tasks (e.g., drills) and teachers for higher-order skills (e.g., negotiation)—can mitigate these issues (Pikhart & Klimova, 2019).

Systemic inequalities (Luo et al., 2022) and infrastructure gaps in Nigeria (Jegede, 2021) and UAE (Muhammed, 2014) limit mobile learning access (Gajić & Maenza, 2022). Personalized systems perpetuate biases (Lim et al., 2023) and erode autonomy (Alm, 2024; Nopas, 2025), while gamification fosters superficial engagement (Blume et al., 2018) without cultural depth (Bax, 2019). Solutions include: (1) Grassroots co-creation of low-resource tools (Masakhane, 2023); (2) Transparent bias auditing frameworks (OECD, 2024; Sandmann et al., 2024); (3) Hybrid pedagogy balancing AI with human instruction to preserve critical thinking and cultural awareness (Miranda et al., 2024). Without these, AI risks exacerbating inequities.

Theoretical, Ethical, and Socio-Cultural Considerations

The implementation of AI in language teaching carries a whole range of theoretical contradictions, ethical problems, and socio-cultural tensions. All these hurdles need to be examined very carefully, lest they perpetuate systemic inequities about the nature of pedagogy.

Criticism of the LLMs

Theoretical critiques highlight LLMs' inability to achieve genuine linguistic competence. Bender and Koller's (2020) "stochastic parrot" argument contends LLMs lack intentional understanding, relying solely on statistical patterns—evidenced by nonsensical outputs (Leivada et al., 2023; Sinha et al., 2023). LLMs also fail to grasp pragmatics like irony or metaphors (Piantadosi & Hill, 2022; Bojic et al., 2023; McCoy et al., 2024), reducing pedagogical utility as seen in students' struggles with AI-generated idioms (Lo, 2023). Even advanced models like GPT-4 (Singhal et al., 2023; Chinonso et al., 2023; Bubeck et al., 2024) underperform in low-resource languages, showing 15–30% accuracy drops for African languages due to data imbalances (Adelani et al., 2024). Tokenization costs further disadvantage non-Latin scripts (Ahia et al., 2023), exposing systemic biases that challenge universal applicability (Akanksha et al., 2024).

Ethical Concerns

Ethics of AI is yet to be resolved concerning the sphere of education. These frameworks emphasize the need for transparency, but governance remains fragmented (Mehandru et al., 2024; Akanksha et al., 2024). According to the findings by Naous et al. (2024) and Kotek et al. (2023), LLMs possess cultural biases misrepresenting non-western rituals and gender stereotypes (78% of the outputs), and are disqualifying for inclusive education. The digital divide worsens the inequities between the rich and the poor; for instance, 40 percent of low-income Spanish students do not have access to artificial intelligence (Sáiz-Manzanares et al., 2021) and a simile applies to India (Jegede, 2021) and Nigeria (scarce offline apps for Yoruba/Igbo). Reliance reduces critical thinking, according to Nopas (2025), while adaptive systems value efficiency rather than creativity (Lee et al., 2022). The prospects of commercialization are tied to the use of its addictive designs (in terms of streaks/rewards) by tools such as Duolingo, with the intermediate fluency only attained by 12% (Poushter & Masci, 2021), revealing the profit-pedagogy tensions that can be drawn (von Ahn & Lewis, 2011).

Cultural Sensitivity

AI's impact on cultural and learning ecosystems sparks debate, with Ukrainian teachers fearing replacement of community-based language practices (Viktorivna et al., 2022) and gamified learning stripping vocabulary of cultural context (e.g., "samurai" as a game mechanic) (Zheng et al., 2015). Inclusive design frameworks (Dong et al., 2021) struggle with regional accents (Dong et al., 2021; Gutiérrez-Colón et al., 2020), while grassroots efforts like Masakhane (2023) face scalability hurdles. LLMs misrepresent cultural concepts (e.g., mis-categorizing Diwali) (Naous et al., 2024) and exhibit semantic disparities (85% task success in English vs. 62% in Tamil) (Li et al., 2024), alienating learners through inaccurate regional feedback (Kazu & Kuvvetli, 2023). Three interlinked solutions are critical: (1) Ethical governance with transparent data audits (OECD, 2024; Sandmann et al., 2024) and marginalized voices (Joshi et al., 2020) to prevent systemic inequities (Kotek et al., 2023; Ghafouri et al., 2023; Naous et al., 2024); (2) Community co-creation via grassroots partnerships (Masakhane, 2023) to prioritize linguistic diversity (Marion et al., 2024) and address data gaps (Joshi et al., 2020; Bapna et al., 2022), adapting inclusive tools to local dialects (Dong et al., 2021; Chen et al., 2022a); (3) Hybrid pedagogy integrating AI feedback (e.g., Grammarly: Dizon & Gayed,

2022) with teacher-led cultural discussions (Lee et al., 2022), refocusing adaptive systems on creativity (Gligorea et al., 2023; Kim et al., 2024) and motivation (Hein et al., 2024) to protect learner agency (Nopas, 2025). These approaches must position AI as enhancing—not replacing—educators to ensure equitable, culturally vibrant language education.

Discussion

The synthesis of 161 studies indicated the relative quantitative predominance of AI efficacy evaluation focusing on technical metrics rather than pedagogic depth (e.g. Liu & Hung, 2016; Dentella et al. 2023). This selection misses critical questions about learning and meaning-making, while qualitative critiques expose basic issues surrounding the unwillingness and incapacity of LLMs to develop an understanding intentionally (Bender & Koller, 2020; Leivada et al., 2023). Increasingly emerging socio-emotional risks encompass eroding learning environments (Viktorivna et al., 2022), concerns regarding instructor substitutions (Hein et al., 2024), and compromising learners' autonomy via algorithmic dependency (Nopas, 2025; García-Peñalvo, 2020). While chatbots seem to relieve anxiety (Çakmak, 2022), doubt exists about whether they facilitate genuine human connection. Bridging studies are extremely scarce (Chen et al., 2022a,b), which highlights the existing methodological gaps that have, to a large extent, neglected qualitative and longitudinal insights as emphasized (Farrokhnia et al., 2023).

Long-standing equity gaps have conspired against the low-resource languages and their learners. With multilingual initiatives such as Aya (Singh et al., 2024), African languages remain largely marginalized: under 8% in corpora (Adelani et al., 2024), with infrastructural challenges creating barriers towards any grassroots initiatives (Masakhane, 2023; Bisk et al., 2020; Bowman et al., 2015). Longitudinal studies are almost nonexistent: studies dealing with vocabulary gain (Kim et al., 2022) or anxiety reduction (Çakmak, 2022) neglect long-term identity or critical-thinking outcomes. In theory, such an apt use of AI proposes a tailored input per Krashen (1982) (Mahowald et al., 2024; Mitchell & Krakauer, 2023; Lee et al., 2022), but the very semantic incomprehension of the LLMs (Bender & Koller, 2020; Kandpal et al., 2023; Hu & Levy, 2023) and disembodied cognition (Dillion et al., 2023) threaten its construct validity. According to sociocultural theory, learning is fundamentally defined in terms of socially mediated meaning-making (Dua et al., 2019; Bandura, 1976), thereby making a case for hybrid models in which AI plays a supportive role in the contextualization led by teachers (Miranda et al., 2024).

It is necessary to center the most marginalized of voices. Community-grounded initiatives (Masakhane, 2023) will oppose Eurocentric training (Akanksha et al., 2024; Naous et al., 2024), but an integration of this aspect is missing. Especially the OECD (2024) policies on participatory design and transparency require some structural support to avoid mere tokenism. The participatory benchmarkings (Adelani et al., 2024) have to be aligned with these underserviced communities. In practice, hybrid models (AI + teacher discussion: Miranda et al., 2024) contextualize feedback from AI and enable educators to mediate the feedback culturally rather than allowing the feedback to govern instruction. Ethical implementation needs to be governed through institutional policies: bias auditing must be compulsory (Kotek et al., 2024; Naous et al., 2024) and datasets need to be transparent along with training teachers to critique implemented AI (OECD, 2024).

Future studies should stress: culturally adaptive tools for regional accents (Kazu & Kuvvetli, 2023; Chen et al., 2022a) and dialects, using participatory datasets (Dong et al., 2021); moving beyond rule-based tasks to narrative/debate applications as an antidote to "thin" gamification (Zheng et al., 2015); tracking motivation, identity, and socio-affective change beyond short-term metrics (Dörnyei & Ushioda, 2021); designing for disabilities/neurodiversity (Mitra, 2024) to confront present biases (Lim et al., 2023). AI holds promise, given the equation that sees societies activated under ethical frameworks, hybrid pedagogy, and community-driven design that enhance human capacity—prioritizing cultural literacy and equity over automation (Karimi & Farreira, 2016).

Conclusion

While there have been notable advancements in AI regarding multilingual chatbots (Çakmak, 2022), adaptive systems (García-Peñalvo, 2020; Lee et al., 2022), and real-time translation (Bapna et al., 2022), the inequity of progress persists favoring high-resource languages and an elite set of users. *Aya* and other models support over 100 languages, but training data for African languages accounts for less than 10% (Singh et al., 2024; Adelani et al., 2024). Grassroots efforts (Masakhane, 2023) are in place to promote diversity, but funding and infrastructure remain major challenges. Three interventions are proposed to secure equity: (1) Binding ethical frameworks (OECD, 2024) on data transparency and audit to reducing gender biases (Kotek et al., 2023) and culture misrepresentation (Naous et al., 2024), including participatory governance (Adelani et al., 2024); (2) Accessible tools for low-resource language users and the disabled (Leong et al., 2020; Dong et al., 2021); (3) Collaboration across sectors to build culturally grounded tools (Bisk et al., 2020; Masakhane, 2023; Miranda et al., 2024). Future works will exploit generative AI to fulfil creative tasks while addressing the superficial engagement of gamification (Galaczi & Cambridge University Press & Assessment, 2023; Zheng et al., 2015), keeping teachers' role in building empathy alive through hybrid models (Nopas, 2025).

ORCID

https://orcid.org/0009-0000-2990-9144

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

Aakanksha, N., Ahmadian, A., Ermis, B., Goldfarb-Tarrant, S., Kreutzer, J., Fadaee, M., & Hooker, S. (2024). The multilingual alignment prism: aligning global and local preferences to reduce harm. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 12027–12049). https://doi.org/10.18653/v1/2024.emnlp-main.671

Adelani, D. I., 1,2, Ojo, J., 1,3, Azime, I. A., 4, Zhuang, J. Y., 5, Alabi, J. O., 4, He, X., 6, Ochieng, M., 7, Hooker, S., 8, Bukula, A., 9, Lee, E.-S. A., 10, Chukwuneke, C., 11, Buzaaba, H., 12, Sibanda, B., 13, Kalipe, G., 14, Mukiibi, J., 15, Kabongo, S., 16, Yuehgoh, F., 17, Setaka, M., 18, Ndolela, L., 19, . . . Haramaya University. (2025). IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 2732–2757). https://aclanthology.org/2025.naacl-long.139.pdf

Ahia, O., Kreutzer, J., & Hooker, S. (2023). Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9904–9923). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.614

Alm, A. (2024). Exploring autonomy in the AI wilderness: Learner challenges and choices. *Education Sciences*, *14*(12), 1369. https://doi.org/10.3390/educsci14121369

Aryabumi, V., Dang, J., Talupuru, D., Dash, S., Cairuz, D., Lin, H., Venkitesh, B., Smith, M., Marchisio, K., Ruder, S., Locatelli, A., Kreutzer, J., Frosst, N., Blunsom, P., Fadaee, M., Üstün, A., & Hooker, S. (2024). Aya 23: Open weight releases to further multilingual progress. In *Technical Report* [Technical report]. https://cohere.com/research/aya/aya-23-technical-report.pdf https://openreview.net/forum?id=506Sxc0Adp&noteId=0s7WkxRn5G

Bandura, A. (1976). *Social learning theory*. Prentice Hall.

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., Daan, V. E., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., Shah, A., Huang, Y., Chen, Z., Wu, Y., Hughes, M. (2022). *Building machine translation systems for the next thousand languages.* Google Research. https://research.google/pubs/building-machine-translation-systems-for-the-next-thousand-languages/

Bassin, L., Raubenheimer, J., & Bell, D. (2023). The implementation of a real time early warning system using machine learning in an Australian hospital to improve patient outcomes. *Resuscitation*, *188*, Article 109821. https://doi.org/10.1016/j.resuscitation.2023.109821

Bender, E. M., & Koller, A. (2020). Critique of LLMs' linguistic competence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.463/

Bisk, Y., Zellers, R., Le Bras, R., Gao, J. & Choi, Y. (2020). PIQA: Reasoning about physical common sense in natural language. *Proceedings of the 34th Conference on Artificial Intelligence.* Association for the advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v34i05.6239

Blume, C. (2019). Playing by their rules. *CALICO Journal*, *36*(1), 19–38. https://doi.org/10.1558/cj.35099

Bojić, L., Kovačević, P., & Čabarkapa, M. (2025). Does GPT-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-04912-x

Botha, J. A., Pitler, E., Ma, J., Bakalov, A., Salcianu, A., Weiss, D., McDonald, R., & Petrov, S. (2017). Natural Language Processing with Small Feed-Forward Networks. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2879–2885. https://doi.org/10.18653/v1/d17-1309

Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in NLP*, 632–642. Association for Computational Linguistics. https://aclanthology.org/D15-1075.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2024). *Sparks of Artificial General Intelligence: Early experiments with GPT-4 - Microsoft Research*. Microsoft Research. https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/

Buckland, M. (2017). Information and Society. In *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/10922.001.0001

Çakmak, F. (2022). Chatbot-human interaction and its effects on EFL pupils' L2 speaking performance and anxiety. *Novitas-ROYAL (Research on Youth and Language), 16*(2), 113–131. https://files.eric.edu.gov/fulltext/EJ1365002.pdf

Chen, C.-H., Koong, C.-S., & Liao, C. (2022a). Influences of Integrating Dynamic Assessment into a Speech Recognition Learning Design to Support Students' English Speaking Skills, Learning Anxiety and Cognitive Load. *Educational Technology & Society*, *25*(1), 1–14. https://www.jstor.org/stable/48647026

Chen, Y.-L., Hsu, C.-C., Lin, C.-Y., & Hsu, H.-H. (2022b). Robot-assisted language learning: Integrating artificial intelligence and virtual reality into English tour guide practice. *Education Sciences, 12*(7), 437. https://doi.org/10.3390/educsci12070437

Chen, Y., Liu, Y., & Zhang, Y. (2020). Exploring the potential of AI in language learning: A meta-analysis of empirical studies. *Educational Technology Research and Development, 68*(4), 1795–1812.

Chen, Y., Liu, Y., & Zhang, Y. (2022c). Exploring the effects of artificial intelligence-driven personalized learning on language learners' performance. *Educational Technology Research and Development*, *70*(1), 213–230.

Chinonso, O., Theresa, A., & Aduke, T. (2023). ChatGPT for teaching, learning and research: Prospects and challenges. *Global Academic Journal of Humanities and Social Sciences, 5*(2), 33–40. https://doi.org/10.36348/ggjhss.2023.v05i02.001

Chiu, T. K. F., & Ching, Y. H. (2018). The effectiveness of AI-driven feedback in language learning: A systematic review. *Computer Assisted Language Learning, 31*(5-6), 486-508.

Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, *96*, Article 102408. https://doi.org/10.1016/j.system.2020.102408

Coniam, D., & Falvey, P. (2016). The second validation study: a qualitative study of the response of markers towards onscreen and Paper-Based marking. In *Springer eBooks* (pp. 79–92). https://doi.org/10.1007/978-981-10-0434-6_6

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 2475-2485). https://doi.org/10.18653/v1/d18-1269

Crompton, H., & Burke, D. (2022). Artificial intelligence in K-12 education. *SN Social Sciences*, *2*(7). https://doi.org/10.1007/s43545-022-00425-5

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00392-8

Dennis, N. K. (2024). Using AI-Powered Speech Recognition Technology to improve English pronunciation and speaking skills. *IAFOR Journal of Education*, *12*(2), 107–126. https://doi.org/10.22492/ije.12.2.05

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-79531-8

Dillion, D., Tandon, N., Guha, T., & Gardner, M. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(8), 1-4. https://doi.org/10.1016/j.tics.2023.04.008

Dizon, G., & Gayed, J. M. (2021). Examining the impact of Grammarly on the quality of mobile L2 writing. *The JALT CALL Journal*, *17*(2), 74–92. https://doi.org/10.29140/jaltcall.v17n2.336

Dizon, G., & Gayed, J. M. (2022). Impact of Grammarly on mobile L2 writing quality. *Journal of Second Language Writing, 58*, 100943.

Dizon, G., & Tang, D. (2020). Intelligent personal assistants for autonomous second language learning: An investigation of Alexa. *The JALT CALL Journal*, *16*(2), 107–120. http://dx.doi.org/10.3968/13571

Dong, Y., Chiang, C., Truong, H., & Qiu, Y. (2021). Enhancing language learning through AI-driven inclusive design. *Educational Technology & Society, 24*(2), 191-204.

Dörnyei, Z., & Ushioda, E. (2021). Teaching and researching motivation. In *Routledge eBooks*. https://doi.org/10.4324/9781351006743

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S. & Gardner, M. (2019). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In J. Burstein, C. Doran & T. Solorio (eds.), *NAACL-HLT (1)* (pp. 2368-2378), Association for Computational Linguistics. Available at http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DuaWDSS019

Extance, A. (2023). ChatGPT has entered the classroom: How LLMs could transform education. *Nature, 623*(7987), 474–477. https://doi.org/10.1038/d41586-023-03507-3

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Gajić, T. D., & Maenza, N. M. (2022). ATTITUDES AND PERCEPTIONS TOWARDS POPULAR LANGUAGE LEARNING APPS: COMPARISON AND ANALYSIS1. *Узданица*, *XIX*(2), 229–244. https://doi.org/10.46793/uzdanica19.2.229g

Galaczi, E. & Cambridge University Press & Assessment. (2023). *English language education in the era of generative AI: our perspective*. https://www.cambridgeenglish.org/Images/685411-english-language-education-in-the-era-of-generative-ai-our-perspective.pdf

García-Peñalvo, F. J. (2020). Learning Analytics as a breakthrough in educational improvement. In Burgos, D. (eds) *Radical Solutions and Learning Analytics* (pp. 1–15). Lecture notes in educational technology (Series). https://doi.org/10.1007/978-981-15-4526-9_1

Ghafouri, V. Agarwal, V., Zhang, Y., Sasttry, N., Such, J., & Tangil, G-S. (2023). AI in the gray: Exploring moderation policies in dialogic large language models vs. human answers in controversial topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 556–565.

Gligorea, I., Cioca, M., Oancea, R., Gorski, A., Gorski, H., & Tudorache, P. (2023). Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature review. *Education Sciences*, *13*(12), 1216. https://doi.org/10.3390/educsci13121216

Gutiérrez-Colón, M., Frumuselu, A. D., & Curell, H. (2020). Mobile-assisted Language learning to enhance L2 reading comprehension: A selection of implementation studies between 2012–2017. *Interactive Learning Environments*, *31*(2), 854–862. https://doi.org/10.1080/10494820.2020.1813179

Hein, I., Cecil, J., & Lermer, E. (2024). Acceptance and motivational effect of AI-driven feedback in the workplace: an experimental study with direct replication. *Frontiers in Organizational Psychology, 2*. https://doi.org/10.3389/forgp.2024.1468907

Hew, K. F., Huang, W., Du, J., & Jia, C. (2022). Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions. *Journal of Computing in Higher Education, 35*(1), 40–68. https://doi.org/10.1007/s12528-022-09338-x

Hew, K. F., Huang, W., Du, J., & Jia, C. (2023). Using chatbots to support pupil goal setting and social presence in fully online activities: Learner engagement and perceptions. *Journal of Computing in Higher Education, 5,* 40–68. https://doi.org/10.1007/s12528-022-09338-x

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.306

Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., & Smit, B. (2024). Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, *6*(2), 161–169. https://doi.org/10.1038/s42256-023-00788-1

Jegede, O. O. (2024). Artificial intelligence and English language learning: Exploring the roles of AI-Driven tools in personalizing learning and providing instant feedback. *Universal Library of Languages and Literatures, 01*(02), 06–19. https://doi.org/10.70315/uloap.ullli.2024.0102002

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282-6293). https://doi.org/10.18653/v1/2020.acl-main.560

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn Long-Tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning* (Vol 202, pp. 15696-15707). PMLR. https://proceedings.mlr.press/v202/kandpal23a/kandpal23a.pdf

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040. https://doi.org/10.1080/17470218.2015.1053951

Kazu, İ. Y., & Kuvvetli, M. (2023). The influence of pronunciation education via artificial intelligence technology on vocabulary acquisition in learning English. *International Journal of Psychology and Educational Studies*, *10*(2), 480–493. https://doi.org/10.52380/ijpes.2023.10.2.1044

Kim, N.-Y. (2022). AI-integrated mobile-assisted language learning: Is it an effective way of preparing for the TOEIC test in classroom environments? *English Teaching, 77*(3), 79–102. https://doi.org/10.15858/engtea.77.3.202209.79

Kim, Y., Chang, Y., Karpinska, M., Garimella, A., Manjunatha, V., Lo, K., Goyal, T., & Iyyer, M. (2024). FABLES: Evaluating faithfulness and content selection in book-length summarization. *Conference On Language Modelling (COLM, 2025)*, Canada. https://openreview.net/forum?id=YfHxQSoaWU&referrer=%5Bthe%20profile%20of%20Tanya%20Goyal%5D(%2Fprofile%3Fid%3D~Tanya_Goyal1)

Klimova, B., & Pikhart, M. (2019). Cognitive and applied linguistics aspects of using social media: The impact of the use of Facebook on developing writing skills in learning English as a foreign language. *European Journal of Investigation in Health Psychology and Education*, *10*(1), 110–118. https://doi.org/10.3390/ejihpe10010010

Klimova, B., & Poulova, P. (2015). Mobile learning and its potential for engineering education. *2022 IEEE Global Engineering Education Conference (EDUCON)*, 47–51. https://doi.org/10.1109/educon.2015.7095949

Klimova, B., & Poulova, P. (2016). Mobile learning in higher education. *Advanced Science Letters*, *22*(5), 1111–1114. https://doi.org/10.1166/asl.2016.6673

Kotek, H., Dockum, R., & Sun, D. (2023). *Gender bias and stereotypes in large language models*. CI '23: Collective Intelligence Conference, Cambridge, MA, United States of America. https://doi.org/10.1145/3582269.3615599

Lee, D., Kim, H., & Sung, S. (2022). Development research on an AI English learning support system to facilitate learner-generated-context-based learning. *Educational Technology Research and Development*, *71*(2), 629–666. https://doi.org/10.1007/s11423-022-10172-2

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL·E 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, *8*(1), 100648. https://doi.org/10.1016/j.ssaho.2023.100648

Li, X., et al. (2024). Quantifying multilingual LLM performance. *Nature Machine Intelligence*, *6*(4), 210–225.

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N., & Du, M. (2024). Quantifying multilingual performance of large language models across languages. Causal and Object-Centric Representations for Robotics (*CoRR 2024)*: CVPR 2024 Workshop. https://openreview.net/forum?id=BppPiG7Tf6

Lim, T., Gottipati, S., & Cheong, M. L. F. (2023). Ethical considerations for artificial intelligence in educational assessments. In *Advances in educational technologies and instructional design book series* (pp. 32–79). https://doi.org/10.4018/979-8-3693-0205-7.ch003

Liu, S., & Hung, P. (2016). Teaching pronunciation with computer assisted pronunciation instruction in a technological university. *Universal Journal of Educational Research*, *4*(9), 1939–1943. https://doi.org/10.13189/ujer.2016.040902

Liu, S., & Yu, G. (2022). L2 learners' engagement with automated feedback: An eye-tracking study. *Language Learning & Technology*, *26*(2), 78–105. https://www.lltjournal.org/item/10125-73480/

Lo, S. (2023). Neural machine translation in EFL classrooms: learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention. *Computer Assisted Language Learning*, *38*(1), 1–20. https://doi.org/10.1080/09588221.2023.2207603

Luo, B., Lin, Y., Chen, N., & Fang, W. (2015). Using smartphone to facilitate English communication and willingness to communicate in a communicative language teaching classroom. *2015 IEEE 15th International Conference on Advanced Learning Technologies*, 320–322. https://doi.org/10.1109/icalt.2015.22

Luo, Y., Kong, X., & Ma, Y. (2022). Effects of multimedia assisted Song integrated teaching on college students' English learning interests and learning outcomes. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.912789

Lytridis, C., Tsinakos, A., & Kazanidis, I. (2018). ARTutor—An augmented reality platform for interactive distance learning. *Education Sciences*, *8*(1), 6. https://doi.org/10.3390/educsci8010006

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., & Hooker, S. (2023). When less is more: Investigating data pruning for pretraining LLMs at scale. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. https://openreview.net/pdf?id=XUIYn3jo5T

Masakhane - Open collective. (n.d.). https://opencollective.com/masakhane

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41). https://doi.org/10.1073/pnas.2322420121

Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01083-y

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, *39*, 161-188.

Miranda, B., Lee, A., Sundar, S., & Koyejo, S. (2024). Beyond scale: the diversity coefficient as a data quality metric demonstrates LLMs are pre-trained on formally diverse data. *The 12th International Conference on Learning Representations (ICAR, 2024),* Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13). https://doi.org/10.1073/pnas.2215907120

Mitra, S. (2024). AI-powered adaptive education for disabled learners. *SSRN.* https://doi.org/10.2139/ssrn.5042713

Mouta, A., Pinto-Llorente, A. M., & Torrecilla-Sánchez, E. M. (2023). Uncovering blind spots in education ethics: Insights from a systematic literature review on artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, *34*, 1166-1205. https://doi.org/10.1007/s40593-023-00384-9

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

Nay, J. J., Karamardian, D., Lawsky, S. B., Tao, W., Bhat, M., Jain, R., Lee, A. T., Choi, J. H., & Kasai, J. (2024). Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, *382*(2270), 20230159. https://doi.org/10.1098/rsta.2023.0159

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial. *Heliyon*, *7*(5), e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

NLLB Team, Costa-jussà M. R., Cross J., Çelebi O., Elbayad M., Heafield K., Heffernan K., Kalbassi E., Lam J., Licht D., Maillard J., Sun A., Wang S., Wenzek G., Youngblood A., Akula B., Barrault L., Gonzalez G. M., Hansanti P., Hoffman J., Jarrett S., Sadagopan K. R., Rowe D., Spruit S., Tran C., Andrews P., Fazil Ayan N., Bhosale S., Edunov S., Fan A., Gao C., Goswami V., Guzmán F., Koehn P., Mourachko A., Ropers C., Saleem S., Schwenk H., Wang J. (2022). No language left behind: Scaling human-centered machine translation. *MetaAI*. https://elbayadm.github.io/assets/talks/nllb/helsinki.pdf

Nopas, D. (2025). Algorithmic learning or learner autonomy? Rethinking AI's role in digital education. Qualitative Research Journal. https://doi.org/10.1108/qrj-11-2024-0282

OECD. (2024). *AI language models: Technological, socio-economic and policy considerations*. OECD Digital Economy Papers. https://doi.org/10.1787/13d38192-en

Olesen, J. (2025, February 28). Duolingo, Memrise, Babbel, Quizlet, Busuu: Which app is best for language learning? *Olesen Tuition*. https://www.olesentuition.co.uk/single-post/which-app-should-i-use-to-learn-a-language-duolingo-memrise-babbel-quizlet-busuu

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *36th Conference on Neural Information Processing Systems (NeurIPS 2022).* https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI).* https://openreview.net/forum?id=nRkJEwmZnM

Pikhart, M. (2018). Technology enhanced learning experience in intercultural business communication course: A case study. In *Emerging Technologies for Education: Third International Symposium* (pp. 41–45), https://doi.org/10.1007/978-3-030-03580-8_5

Pikhart, M. (2019a). Multilingual and Intercultural Competence for ICT: Accessing and Assessing Electronic Information in the Global World. In Choroś, K., Kopel, M., Kukla, E., Siemiński, A. (eds) *Multimedia and Network Information Systems (MISSI 2018), Advances in Intelligent Systems and Computing, 833.* https://doi.org/10.1007/978-3-319-98678-4_28

Pikhart, M. (2019b). Interculturality in blended learning: Challenges of electronic communication. In Uskov, V., Howlett, R., Jain, L. (eds) *Smart Education and e-Learning 2019. Smart Innovation, Systems and Technologies, 144.* https://doi.org/10.1007/978-981-13-8260-4_9

Pikhart, M. (2020). The use of technology in the learning environment for business communication: applied linguistics of business communication from the positive psychology perspective. *Procedia Computer Science, 176*, 1479–1486. https://doi.org/10.1016/j.procs.2020.09.158

Pikhart, M., & Klimova, B. (2019). Utilization of linguistic aspects of Bloom's taxonomy in blended learning. *Education Sciences*, *9*(3), 235. https://doi.org/10.3390/educsci9030235

Poulova, P., & Simonova, I. (2017). Innovations in data engineering subjects. *Advanced Science Letters*, *23*(6), 5090–5093. https://doi.org/10.1166/asl.2017.7316

Qu, S., Li, K., Wu, B., Zhang, S., & Wang, Y. (2019). Predicting student achievement based on temporal learning behavior in MOOCs. *Applied Sciences*, *9*(24), 5539. https://doi.org/10.3390/app9245539

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rohmiyati, N. Y. (2025). Enhancing English language learning through Artificial intelligence: opportunities, challenges and the future. *DIAJAR Jurnal Pendidikan Dan Pembelajaran*, *4*(1), 8–16. https://doi.org/10.54259/diajar.v4i1.3344

Rowe, L. W. (2022). Google Translate and Biliterate Composing: Second-Graders' use of digital translation tools to support bilingual writing. *TESOL Quarterly*, *56*(3), 883–906. https://doi.org/10.1002/tesq.3143

Sáiz-Manzanares, M. C., Almeida, L. S., Martín-Antón, L. J., Carbonero, M. A., & Valdivieso-Burón, J. A. (2022). Teacher Training Effectiveness in Self-Regulation in Virtual Environments. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.776806

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). WinoGrande. *Communications of the ACM*, *64*(9), 99–106. https://doi.org/10.1145/3474381

Sandmann, S., Riepenhausen, S., Plagwitz, L., & Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-46411-8

Sasikala, P., & Ravichandran, R. (2024). Study on the impact of artificial intelligence on student learning outcomes. *Journal of Digital Learning and Education*, *4*(2), 145–155. https://doi.org/10.52562/jdle.v4i2.1234

Shivakumar, A., Shukla, S., Vasoya, M., Kasrani, I., & Pei, Y. (2019). AI-enabled language speaking coaching for dual language learners. *Iadis International Journal on Www/Internet*, *17*(1), 66-78. https://www.iadisportal.org/ijwi/papers/2019171105.pdf

Singh, S., Rozière, B., Gehrmann, S., Rabe, M., Ustun, A., & Aji, A. F. (2024). Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11521–11567), Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.620.pdf

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Sinha, K., Gauthier, J., Mueller, A., Misra, K., Fuentes, K., Levy, R., & Williams, A. (2023). Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* (pp. 6043–6063). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.333.pdf

Solak, E. (2024). Revolutionizing language learning: How ChatGPT and AI are changing the way we learn languages. *International Journal of Technology in Education (IJTE)*, *7*(2), 353–372. https://doi.org/10.46328/ijte.732

Sonawane, A. D., Midhun. Moorthi, C., Gnanamurthy, R. K., & Sathya, C. (2023). Artificial intelligence - a modern approach. In *Zenodo*. https://doi.org/10.5281/zenodo.8177106

Sung, Y., Chang, K., & Liu, T. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

Sung, Y., Yang, J., & Lee, H. (2017). The effects of mobile-computer-supported collaborative learning: Meta-Analysis and critical synthesis. *Review of Educational Research*, *87*(4), 768–805. https://doi.org/10.3102/0034654317704307

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M,-A., Lacroix, T., Roziere, B., Goyal, N., hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models.* arXiv.Org, abs/2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Traxler, J. (2018). Distance learning—predictions and possibilities. *Education Sciences*, *8*(1), 35. https://doi.org/10.3390/educsci8010035

Truck, F., & Moravec, H. (1991). Mind Children: The future of robot and Human intelligence. *Leonardo*, *24*(2), 242. https://doi.org/10.2307/1575314

Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P. H. S., Bibi, A., Albanie, S., & Bethge, M. (2024). No "Zero-Shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024),* https://proceedings.neurips.cc/paper_files/paper/2024/file/715b78ccfb6f4cada5528ac9b5278def-Paper-Conference.pdf

Üstün, A., Aryabumi, V., Yong, Z., Ko, W., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., & Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 15894–15939). https://aclanthology.org/2024.acl-long.845.pdf

Viktorivna, K. L., Oleksandrovych, V. A., Oleksandrivna, K. I., & Oleksandrivna, K. N. (2022). Artificial Intelligence in Language Learning: What Are We Afraid of. *Arab World English Journal*, *8*, 262–273. https://doi.org/10.24093/awej/call8.18

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (*NeurIPS 2019)* (pp. 3266–3280). Curran Associates Inc.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Wu, Q. (2015). Designing a smartphone app to teach English (L2) vocabulary. *Computers & Education*, *85*, 170–179. https://doi.org/10.1016/j.compedu.2015.02.013

Wu, Q., Xu, Y., Xiao, T., Xiao, Y., Li, Y., Wang, T., Zhang, Y., Zhong, S., Zhang, Y., Lu, W., & Yang, Y. (2024). *Surveying Attitudinal Alignment Between Large Language Models Vs. Humans Towards 17 Sustainable Development Goals*. arXiv.Org, abs/2404.13885. https://doi.org/10.48550/arxiv.2404.13885

Yusuf, O., Jegede, O., Aliyu, A., & Aladetan, T. (2024). Prospects and challenges of technology enhanced language learning: Teachers perspectives. *Studies in Literature and Language*, *29*(1), 21-32.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. (pp. 4791-4800). https://aclanthology.org/P19-1472.pdf

Zhao, J., Xu, W., & Thille, C. (2021). End-to-end question generation to assist formative assessment design for conceptual knowledge learning. *Amazon Science*. https://assets.amazon.science/12/6a/6614594c453a95f2bdf5f2c3ca95/end-to-end-question-generation-to-assist-formative-assessment-design-for-conceptual-knowledge-learning.pdf

Zheng D., Bischoff M., Gilliland B. (2015). Vocabulary learning in massively multiplayer online games: Context and action before words. *Educational Technology Research and Development*, *63*(5), 771–790.

Zheng, D., Newgarden, K., & Young, M. F. (2015). Multimodal analysis of language learning in *World of Warcraft* play: Languaging as values-realizing. *ReCALL*, *27*(3), 336–355. https://doi.org/10.1017/S0958344015000169