English Language Learning with AI: Proficiency Gains and Learner Experience

Arwa Althobaiti[1]

[1] Assistant Professor of English and Applied Linguistics, Jouf University, KSA

Correspondence: Arwa Althobaiti, Assistant Professor of English and Applied Linguistics, Jouf University, KSA

Received: May 28, 2025 Accepted: July 21, 2025 Online Published: August 25, 2025

doi:10.5430/wjel.v15n8p228 URL: https://doi.org/10.5430/wjel.v15n8p228

Abstract

This study investigates the effectiveness of an AI-assisted language learning platform compared to traditional instruction among first-year Saudi university students. Using a quasi-experimental mixed-methods design, the study assigned 147 students to either an AI-assisted group or a traditional classroom group over a six-week period. Quantitative data from pre- and post-tests (TOEFL ITP) revealed significantly greater gains in the AI group ($d = 0.85$), even after controlling for baseline proficiency. Regression analysis showed that time spent on the AI platform was a strong predictor of learning gains, with each additional 30 minutes of usage correlating with a 1.8-point improvement. Qualitative data from 30 post-intervention interviews highlighted the perceived benefits of immediate feedback, gamified motivation, and self-paced learning, alongside reported challenges such as streak-related anxiety, accent misclassification, and technical issues. Findings support a blended learning model that uses AI to give students consistent feedback, while keeping teachers focused on guiding learning and supporting students emotionally and socially. The study contributes practical and theoretical insights relevant to policymakers, curriculum designers, and educators seeking to implement AI in EFL contexts.

Keywords: AI-assisted language learning; English in Saudi Arabia; second language acquisition; speech recognition; gamification

1. Introduction

Artificial intelligence (AI) is moving rapidly and is fast becoming an essential component of contemporary education. Recent market estimates place the value of AI-based learning products at approximately US $5.88 billion in 2024, with projections of US $32 billion by 2030—an average annual growth rate exceeding 31 percent (Grand View Research, 2024). A complementary industry brief anticipates an even faster short-term rise as it forecasts US $7.57 billion for 2025 and a five-year compound annual growth rate of 41 percent (The Business Research Company, 2025). These figures reflect rapid advances in cloud infrastructure, real-time learning analytics, and large language models (LLMs) that generate feedback that fits the learner's needs and appears almost instantly. The policy environment mirrors this trend: UNESCO now classifies AI as a "general-purpose technology" capable of promoting inclusive and high-quality learning (UNESCO, 2024), while the European Commission's Digital Education Action Plan 2021-2027 allocates roughly €700 million to AI-driven teacher-support initiatives.

Within this broader context, English-language learning (ELL) has consistently acted both as a catalyst for technical innovation and as a primary beneficiary of AI-enhanced tools. The discipline's technological development covers early computer-assisted language learning (CALL) in the 1960s, multimedia CALL in the 1990s, and mobile-assisted language learning (MALL) in the 2010s. The most recent stage, AI-assisted language learning (AIALL)**,** integrates several distinct technologies: adaptive item-sequencing engines (e.g., Duolingo's Birdbrain**),** speech-recognition tutors (e.g., ELSA Speak, Google Read Along), and LLM-based conversational agents (e.g., ChatGPT, Gemini). These systems provide real-time feedback on vocabulary, syntax, pronunciation, and pragmatic choices, and are widely used: Duolingo alone reported 16.3 billion exercises completed in 2024, with 61 percent of sessions guided by AI-generated sequences (Duolingo Research Team, 2024).

Empirical findings generally corroborate the benefits of these tools. A meta-analysis of 2 156 participants across 15 controlled studies identified a large aggregate effect ($d = 1.17$) for AI-enhanced interventions on overall L2 achievement (Xu, Yu, & Liu, 2025). A separate synthesis focused on AI-mediated assessments reported a medium effect ($g = 0.39$) for language-learning outcomes. This underscores the formative value of automated quizzes (Chen et al., 2025). A systematic review that comprised 37 chatbot-based studies noted consistent improvements in vocabulary, fluency, and engagement. It also emphasized the continued importance of teacher facilitation to maintain substantive interaction (Li et al., 2025).

Besides proficiency measures, AI appears to influence learner motivation and self-regulation, both of which are essential for sustained language development. Duolingo telemetry indicates that its adaptive algorithm reduces dropout by 14 percent and extends average study streaks by approximately seven days (Duolingo Research Team, 2024). Survey evidence corroborates these behavioural patterns. For instance, in a U.S. study involving 436 undergraduates, 72 percent regarded AI tools as beneficial, although 48 percent expressed concerns regarding academic integrity and data privacy (Breese, Rebman, & Levkoff, 2024). Similarly, an international survey of 3 839

students across 16 countries reported widespread enthusiasm, but indicated calls for clearer institutional guidelines on plagiarism, privacy, and algorithmic transparency (Digital Education Council, 2024).

The Saudi Arabian context similarly highlights both the promise and the constraints of bringing AI into language education. Under Vision 2030, the Kingdom launched the National Strategy for Data and AI (NSDAI) and set up a dedicated authority, SDAIA. This signals that AI is a cornerstone of economic diversification, and media estimates place public funding for AI and digital-learning initiatives at US $40 billion across the current decade (Telecom Review, 2025). In the 2024 budget, the Ministry of Education set aside SR 201 billion (≈ US $53 billion) for sector reforms, a portion of which targets AI-enabled curricula and teacher training (Arab News, 2024). On paper, the country has a robust digital infrastructure: The Communications, Space & Technology Commission (CST) reports 99 percent internet penetration nationwide, with average mobile speeds exceeding 200 Mbps (CST, 2023). Yet regional disparities persist; stakeholder interviews in the Najran and Al-Jawf governorates cite patchy 5G coverage and higher latency during peak hours, conditions that could hamper real-time speech-recognition feedback in remote schools.

Linguistic diversity adds another challenge. Classrooms operate in an environment that blends Gulf Arabic varieties (Najdi, Hijazi, Eastern), Modern Standard Arabic, and English as the primary foreign language. Because mainstream speech-recognition engines are trained mainly on North-American and British English phonologies, they are likely to mis-score Gulf-Arabic–accented vowels and the emphatic consonant set / ṣ ḍ ṭ /. This reinforces deficit views of local pronunciation norms. Applied-linguistics scholars in KSA and the Middle East therefore advocate a "glocal" paradigm (Ahmad, 2023), which, in the case of AI, would argue that Global AI systems should be adapted to local speech patterns and include culturally familiar content, such as Hijazi idioms and NEOM-related STEM vocabulary (Al-Zahrani, 2025). Such localisation efforts would be in sync with the government's target of training 20 000 AI and data specialists by 2030 (Telecom Review, 2025) and ensure that the pedagogical benefits of AI are equitably distributed across Saudi Arabia's diverse educational landscape.

Despite the accelerated implementation, three main knowledge gaps continue to hamper evidence-based decision-making. **First,** comparative effectiveness remains under-specified. Many studies rely on unmatched control groups or limit their scope to single micro-skills, which complicates causal inference (Xu et al., 2025). Second, long-term learning paths are largely unexplored as most investigations span no more than eight weeks, a fact which leaves retention and transfer unclear (Chen et al., 2025). Third, learner experience and equity have not been adequately discriminated by variables such as gender, socioeconomic status, or digital access, considerations that are crucial in settings characterized by a marked digital divide (Li et al., 2025).

Against this backdrop, the present study pursues two objectives. Objective 1 is to compare English-language proficiency gains among Saudi students who follow a traditional instructor-centred curriculum, while studying the same content and covering similar instructional hours. Objective 2 is to evaluate learners' perceptions of feedback quality, motivational impact, and cognitive load in both instructional modes. By integrating quantitative outcome measures with qualitative data on learner experience, the research aims to provide insights for teachers, curriculum designers, and policy-makers regarding appropriate and equitable integration of AI into ELL programs.

Therefore, the investigation addresses two research questions:

1. Does AI-assisted instruction lead to significantly greater gains in English-language proficiency than traditional instruction delivered under similar conditions?
2. How do learners perceive and experience AI-assisted language learning compared with traditional methods in terms of the usefulness of feedback, motivation, and cognitive load?

Triangulating standardized test scores with psychometric surveys and semi-structured interviews will clarify not only whether AI enhances learning but also how and for whom it does so. The findings are expected to inform a framework for AI adoption that balances technological capabilities with pedagogical best practice and KSA's multilingual realities.

2. Literature Review

2.1 Theoretical Framework: An Integrated SLA Perspective

This study draws on a multi-theoretical foundation that includes behaviourist, cognitive, sociocultural, interactionist, and cognitive load theories, each offering a distinct lens on second language acquisition (SLA) and informing both the analysis and design implications of AI-based language learning tools.

Early studies of second-language instruction were grounded in behaviourism, which viewed learning as the formation of stimulus–response relationships (Skinner, 1957). Audiolingual drills and rote substitution exercises are illustrations of this paradigm, as they emphasized habit formation through repetition and corrective feedback. In reaction, cognitive theories reframed language learning as an information-processing activity in which learners construct, store, and retrieve mental representations (Anderson, 1983; Chomsky, 1965). Within cognitive SLA, input-based models (Krashen, 1985) highlight the role of comprehensible input in developing implicit linguistic knowledge, whereas output-oriented hypotheses (Swain, 2005) argue that it is production that prompts learners to notice gaps in their interlanguage.

Sociocultural and constructivist perspectives subsequently shifted attention from individual cognition to socially mediated meaning-making. Drawing on Vygotsky's (1978) Zone of Proximal Development (ZPD), these frameworks contend that learning is maximized when novices interact with more knowledgeable peers or scaffolding tools. This principle was later operationalised in

computer-supported collaborative learning environments. Interactionist theory further maintains that negotiated exchanges, such as recasts, clarification requests, and confirmation checks, provide "developmentally optimal" conditions for internalising form–meaning mappings (Long, 1996; Gass & Mackey, 2006). Contemporary AI platforms embody these strands: adaptive algorithms deliver incremental input in line with cognitive models, elicit output with real-time feedback aligned with interactionist claims, and frequently embed social or game-based features that give support to constructivist peer mediation.

Recent work has added cognitive-load theory (CLT) to the theoretical underpinnings. CLT distinguishes intrinsic, extraneous, and germane load, which emphasizes the need to minimize unnecessary interface complexity (Sweller, 2019). Empirical eye-tracking studies show that learners may spend up to one-third of on-screen time focusing on non-instructional badges and animations (Sharma, Giannakos, & Dillenbourg, 2020). This suggests that poorly designed gamification can undermine the cognitive benefits of adaptive sequencing. On the other hand, explainable AI tools—like color-coded visuals that show which sounds were mispronounced—can help students focus on what matters by turning confusing feedback into something clear and useful (Górriz et al., 2023).

In a nutshell, these frameworks shape how we understand the potential and limitations of AI-driven language learning: behaviourism informs feedback loops, cognitive theories support sequencing and input/output tasks, sociocultural perspectives highlight collaborative scaffolding, interactionism guides task design, and CLT offers insight into interface optimization.

2.2 Traditional Language-Learning Methods

Grammar-Translation (GT): Originating in eighteenth-century classics instruction, GT gives priority to written accuracy and explicit metalinguistic analysis. Although it is often criticized for limited communicative transfer, recent corpus-based studies indicate that GT can raise reading proficiency and academic vocabulary when there is sufficient exposure to the lessons or rules (Richards & Rogers, 2014).

Communicative Language Teaching (CLT): Building on Hymes's (1972) notion of communicative competence, CLT puts emphasis on authentic, meaning-centred interaction. Classroom experiments link CLT to measurable gains in fluency, strategic competence, and willingness to communicate (Canale & Swain, 1980; Ellis, 2012). Nonetheless, critics note inconsistent grammatical accuracy when input is inadequate, which highlights the need to make compromises between fluency and form.

Teacher-centred instruction: Lecture-plus-exercise formats remain widespread, especially in high-enrolment contexts. Observational research suggests that teacher-fronted lessons efficiently convey rules and illustrate target forms, though they often limit opportunities for negotiation of meaning or individualized feedback (Brown, 2014). These documented gains and constraints serve as a baseline against which the gains of AI must be evaluated.

Task-Based Language Teaching (TBLT): Although it is sometimes subsumed under CLT, TBLT deserves attention because its focus on real-world tasks provides an "ecological benchmark" for evaluating the authenticity of AI-generated activities (Ellis et al., 2018). Studies comparing AI task prompts with teacher-designed tasks report comparable lexical variety, but they also note fewer pragmatic moves in AI-generated dialogues (Chen, Li, & Ye, 2024).

2.3 AI-Assisted Language Learning (Rosetta Stone as an Example)

AI systems extend CALL by combining natural-language processing (NLP) with adaptive scheduling to personalize feedback. Duolingo's *Birdbrain* algorithm, for instance, selects items predicted to maximize retention based on Bayesian knowledge-tracing (Duolingo Research Team, 2024). Pronunciation-centred applications such as ELSA Speak employ deep neural acoustic models to generate phoneme-level diagnostics.

Rosetta Stone represents a mature instantiation of AIALL. Its TruAccent® speech-recognition engine aligns user utterances with native-speaker acoustic models and provides segmental and suprasegmental feedback. A quasi-experimental study involving **120** Indian engineering undergraduates reported significant gains in accuracy, fluency, vocabulary, and pronunciation when compared to a traditional control group (Dandu, Charyulu, & Kumari, 2024). An earlier U.S. randomized controlled trial (RCT) with 221 middle-school English learners found that Rosetta Stone users outperformed peers on listening and speaking after one semester (Harper et al., 2021). Although these studies are promising, they often confound Rosetta Stone usage with additional teacher guidance, a fact which complicates claims of generalizability across proficiency levels and age groups.

A systematic review of 37 chatbot-mediated interventions concluded that dialogue agents promote vocabulary growth, fluency, and engagement, although sustained success depends on explicit teacher–AI orchestration (Li et al., 2025). Similarly, a meta-analysis of 2 156 participants across 15 trials produced a large aggregated effect ($d = 1.17$) for AI interventions but revealed substantial heterogeneity ($Q = 148.2$, $p < .001$). This indicates the presence of unexamined moderators, such as interface design, session length, and learner age (Xu, Yu, & Liu, 2025).

2.4 Comparative Studies: AI vs Traditional Methods

Quantitative syntheses now lead to the following conclusion: AI-assisted modalities generally outperform traditional instruction when standard proficiency measures are considered (Xu et al., 2025). A complementary meta-analysis of AI-enabled assessments reported a medium pooled effect ($g = 0.39$) and stressed efficiency gains in formative feedback cycles (Chen et al., 2025). Classroom-level RCTs echo these findings, yet they illuminate boundary conditions. Harper et al.'s (2021) study revealed that proficiency gains were highly correlated with usage time, which implies that time and engagement mediate outcomes. Other trials report diminished returns beyond

approximately eight hours of weekly exposure, suggesting a non-linear "sweet spot" for AI use (Xiu-Yi, 2024).

Strengths of AI approaches include:

- *Personalization*: adaptive sequencing aligns with spacing and retrieval theories, therefore maximizing retention.
- *Multimodal input*: speech-recognition and NLP modules supply immediate feedback, which is by no means available in large classes.
- *Data analytics*: Logs enable instructors to diagnose persistent error patterns and tailor follow-up instruction.

Limitations remain substantial:

- *Equity and access*: device availability and broadband quality negatively affect rural and low-income learners (Li et al., 2025).
- *Cognitive load*: complex dashboards may overwhelm beginners, particularly when multiple gamified elements compete for attention (Sweller, 2019).
- *Pedagogical alignment*: many AI tools focus on receptive drills, while they offer only limited support for extended discourse or intercultural pragmatics.

Research gaps persist in at least four domains:

1. *Longitudinal efficacy*: few studies track retention or transfer beyond eight weeks.
2. *Interaction effects*: systematic variation in teacher scaffolding is rare, which does not shed light on best scaffolding models.
3. *Affective variables*: motivation, anxiety, and learner autonomy remain under-studied despite their centrality to SLA.
4. *Population diversity*: the evidence base is heavily focused on tertiary and K-12 learners in high-resource settings, while heritage speakers, younger children, and low-connectivity contexts remain understudied.

Addressing these gaps will require multi-site, mixed-methods research that manipulates exposure time, interface design, and teacher involvement while capturing both cognitive and affective outcomes over extended periods.

3. Materials and Methods

3.1 Participants and Design

This study employed a quasi-experimental, mixed-methods design involving 147 first-year university students in Saudi Arabia. All participants were enrolled in a mandatory freshman English course and were native Arabic speakers with intermediate English proficiency. They were divided into two instructional conditions: an AI-assisted learning group (n = 74) and a traditional classroom group (n = 73). Group assignment was determined by intact class sections rather than randomization. A pre-intervention TOEFL ITP test confirmed no significant difference in English proficiency between the two groups ($t(145) = 0.19$, $p = .85$). This ensured a fair comparison of the instructional methods.

Students in the AI-assisted group learned English using the Rosetta Stone™ online platform. The platform features an AI-driven curriculum with interactive lessons covering vocabulary, grammar, reading, and listening, complemented by speaking exercises using the TruAccent® speech-recognition engine. TruAccent compares learners' pronunciation to native speaker models and provides instant, segmental feedback to help students refine their accent and fluency. Learners progressed through units matching the course syllabus (e.g., basic conversational topics and academic vocabulary); they received immediate corrective feedback on their responses. The platform also incorporated gamified elements (e.g. daily practice streaks and achievement badges) to motivate regular engagement. For the study, students in this group accessed Rosetta Stone during scheduled class lab sessions and were encouraged to practice independently between classes. All platform usage (time on task and performance data) was logged automatically for analysis. An experienced facilitator was present during lab sessions to assist with technical issues but did not provide direct language instruction, so that feedback and practice largely came from the AI system.

Students in the traditional instruction group received face-to-face teaching in a classroom setting. A qualified English instructor led these sessions using a communicative, textbook-based curriculum aligned to the same content units covered by the AI platform. This ensured both groups studied equivalent material (topics, vocabulary, and grammar structures) over the six weeks. Lessons typically included brief lectures or explanations of new language points, whole-class and pair-work activities (such as dialogues, grammar exercises, and listening tasks) using print and audio materials, and periodic reviews. The instructor provided feedback in the form of corrections, explanations, and encouragement during activities, but unlike the AI group, feedback was not instantaneous for every attempt. No AI or speech-recognition tools were used in this condition; however, students had access to standard course resources (e.g. worksheets and audio recordings) for practice. The traditional group's instructional approach reflected common EFL teaching methods in the region, which emphasized teacher guidance and scheduled assessments.

Both groups followed the intervention over a six-week period. Classes met twice per week for approximately 90 minutes per session (roughly 3 hours of instruction weekly). In total, each participant received about 18 hours of in-class instruction. The scheduling and contact hours were identical for the two groups to ensure a comparable "dosage" of English exposure. All participants completed the same homework assignments or practice exercises corresponding to the week's content; for the AI group these were fulfilled through the Rosetta Stone exercises, whereas the traditional group used written assignments and practice from their textbook. At the end of the six

weeks, both groups had covered the same curriculum topics and skill areas, which permitted for a direct comparison of learning outcomes.

3.2 Instruments (Quantitative Measures)

English proficiency gains were measured using the Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP) Level 1 tests, administered before and after the intervention. The pre-test was given in the week prior to the intervention, and a parallel form of the test was used as the post-test in the week following the intervention to minimize any test-retest practice effects. Each testing session lasted approximately 115 minutes and was invigilated by the researchers under standardised conditions.

Following ETS guidelines, each student's performance was scored to yield a composite listening-plus-reading score. Internal consistency of the test scores in our sample was high, with Cronbach's $\alpha = .89$ at pre-test and .91 at post-test, consistent with the TOEFL ITP's established reliability for university populations. This indicates that the test scores were stable and reliable for measuring proficiency in this context.

3.3 Qualitative Data Collection (Interviews)

To complement the quantitative findings, qualitative data were gathered through semi-structured interviews conducted after the post-test. A purposive sub-sample of 30 participants (approximately 20% of the total sample) took part in one-on-one interviews. To capture diverse perspectives, the interviewees were drawn from both the AI-assisted and traditional groups. Each interview lasted about 15–20 minutes and was conducted in English (with occasional Arabic clarifications if needed) by one of two bilingual researchers.

The interview protocol consisted of open-ended questions designed to explore the learner's experience during the six-week program, and focused on key areas such as the usefulness of feedback (e.g., "How did you find the feedback you received from the platform/instructor?"), motivation and engagement (e.g., "What kept you motivated to continue learning each week?"), and cognitive load or challenges (e.g., "Did you ever feel overwhelmed or confused by the materials or technology? Can you describe that?"). Follow-up prompts encouraged students to elaborate on any differences they perceived between AI-based learning and traditional classroom instruction in these areas. All interviews were audio-recorded with participant consent and later transcribed *verbatim* to ensure accuracy of the qualitative data.

The interview transcripts were analyzed using the procedures of thematic analysis outlined by Braun and Clarke (2006). The analysis was conducted inductively, with the aim of identifying recurring themes and patterns related to learner experience. An initial codebook was developed based on the study's research questions and salient issues that emerged during a close reading of a subset of transcripts. Codes were organized around constructs such as perceived feedback quality, motivational dynamics, cognitive load, and other emergent concerns. Once the preliminary codebook was established, I applied it systematically to the full set of transcripts and reviewed each segment and assigning codes to relevant content.

Codes were then iteratively grouped into broader thematic categories. To ensure analytic rigor, the themes were revisited multiple times during the coding process, and memos were written to reflect on and refine thematic boundaries. This interpretive process led to the identification of key themes that addressed the second research question, offering insights into how learners perceived and responded to AI-assisted versus traditional instruction. Thematic analysis thus allowed for a grounded understanding of the learner experience and helped complement the quantitative results with detailed qualitative evidence.

4. Results

4.1 Quantitative Results

4.1.1 Descriptive statistics and baseline equivalence

A total of 147 first-year students completed both the pre- and post-test batteries (AI-assisted = 74; Traditional = 73). Internal-consistency estimates for the TOEFL ITP Listening + Reading composite were satisfactory (Cronbach's $\alpha = .89$ pre-test; $\alpha = .91$ post-test), which is in alignment with ETS norms for tertiary populations. Table 1 summarizes the central-tendency and dispersion indices.

Table 1. Central‐tendency and dispersion

Group	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	Gain (Δ)
AI-assisted	74	48.0	7.3	62.1	7.9	+14.1
Traditional	73	47.8	7.1	54.8	7.7	+7.0

An independent-samples *t*-test found no baseline difference, $t(145) = 0.19$, $p = .85$, confirming initial equivalence. Skewness (±.14) and kurtosis (±.48) fell within ±1.0, which meets the assumption of approximate normality (Field, 2018).

4.1.2 Within-Group Gains

Both cohorts demonstrated significant improvement across the six-week intervention.

- AI-assisted group: $t(73) = 18.92$, $p < .001$; Cohen's $d = 2.20$ (very large; Cohen, 1992). The 95 % confidence interval for the mean gain ranged from 12.3 to 15.9 points.
- Traditional group: $t(72) = 11.83$, $p < .001$; $d = 1.38$ (large). The 95 % confidence interval for the mean gain ranged from 5.7 to 8.3 points.

Although traditional instruction produced gains comparable to semester-long English for Academic Purposes (EAP) benchmarks reported in regional studies, the AI cohort more than doubled that improvement.

4.1.3 Between-Group Effect Size and Practical Significance

The difference in gain scores produced a between-group Cohen's $d = 0.85$, conventionally interpreted as a large effect (Cohen, 1992). Translating to an *U*₃ of 80 %, an average AI learner scored higher at post-test than four out of five learners in the traditional cohort. A complementary common-language (CL) effect size of .73 indicates a 73 % probability that a randomly selected AI student would outperform a randomly selected traditional student, an interpretation which is often found to be more intuitive for practitioners (McGraw & Wong, 1992).

4.1.4 ANCOVA Controlling for Baseline Proficiency

A one-way analysis of covariance (ANCOVA) was performed with post-test composite score as the dependent variable, instructional group as the fixed factor, and pre-test composite as the covariate (Table 2).

Table 2. ANCOVA Summary for Post-Test Scores by Instructional Group, Controlling for Pre-Test Proficiency

Source	SS	df	MS	F	p	Partial η^2
Pre-test (covariate)	1 948.6	1	1 948.6	31.09	< .001	.177
Group (treatment)	780.4	1	780.4	12.42	.001	.079
Error	9 013.7	144	62.6			

The adjusted $F(1, 144) = 12.42$, $p = .001$ confirms a statistically significant treatment effect when controlling for baseline proficiency. The partial $\eta^2 = .079$ signifies a medium-to-large practical impact, which is in sync with recent AI efficacy meta-analyses (Lee & Lee, 2024).

4.1.5 Sub-Skill Analysis

Separate paired comparisons examined listening and reading sub-scores (Figure 1).

- Listening: AI gain = +7.9, Traditional = +3.4; $t(145) = 4.26$, $p < .001$; $d = 0.91$.
- Reading: AI gain = +6.2, Traditional = +3.6; $t(145) = 3.01$, $p = .003$; $d = 0.64$.

The larger effect in listening suggests that the speech-recognition component confers disproportionate benefits for auditory-comprehension skills—consistent with findings from pronunciation-oriented AI studies (Huang & Papi, 2025).

4.1.6 Exploratory Regression: Usage Predictors of Gain

A hierarchical multiple regression assessed whether platform usage minutes and mean speech-recognition accuracy predicted gain beyond pre-test proficiency (Table 3).

Table 3. Hierarchical Multiple Regression Predicting TOEFL Gain Scores

Step	Predictor	β	T	p	Δ_R^2
1	Pre-test score	–.13	–1.07	.29	.02
2	+ Usage minutes	.46	4.34	< .001	+.20
3	+ Speech-rec. accuracy	.21	2.09	.040	+.04
	Total R^2				.26

The final model accounted for 26 % of gain-score variance. Usage minutes emerged as the strongest predictor ($\beta = .46$), as it indicated that each additional 30 minutes of practice corresponded to an estimated 1.8-point TOEFL gain, holding other factors constant. The positive contribution of speech-recognition accuracy ($\beta = .21$) suggests that effective interaction with the ASR engine enhances learning beyond simpler time-on-task.

4.1.7 Assumption Checks and Robustness

- Normality: Shapiro–Wilk tests were non-significant (AI $p = .21$; TR $p = .16$).
- Homogeneity: Levene's test showed equal error variances ($p = .48$).
- Multicollinearity: Variance-inflation factors in the regression were < 1.4, indicating no multicollinearity concerns (Field, 2018).
- Sensitivity: Adjusting for three extreme values (> |3 SD|) altered no inferential decisions.

4.2 Qualitative Results

4.2.1 Data Corpus and Coding Reliability

Thirty post-intervention interviews (≈ 9 hours audio; 78 000 words) were transcribed verbatim. Using NVivo 14, two analysts independently coded the data, achieving Cohen's $\kappa = .82$, which exceeds the .80 threshold for substantial agreement (Miles, Huberman, & Saldaña, 2019). Three principal themes and nine sub-themes captured the experiential landscape as shown below:

4.2.2 Theme 1 – Empowered Autonomy and Gamified Motivation

Sub-theme 1A: Self-pacing and agency
"I repeated the pronunciation drill until the bar turned green; in class I never get that many chances." (AI-05)

Sub-theme 1B: Micro-goals and streaks
Twenty-eight of 30 AI interviewees referenced streak mechanics; 21 linked streaks directly to daily study habits, which mirrors Duolingo analytics on habit formation.

Sub-theme 1C: Goal clarity
Traditional learners valued clear weekly targets but expressed a desire for "small wins" between assessments.

4.2.3 Theme 2 – Feedback Quality and Perceived Learning

Sub-theme 2A: Immediate phonological diagnostics
"TruAccent showed my /θ/ waveform…after twenty tries it finally turned green." (AI-12)

Sub-theme 2B: Accent-bias frustration
"Sometimes it flags my Saudi /p/ even when teachers say it's acceptable." (AI-41)

Sub-theme 2C: Teacher clarity versus AI speed
"Madame F. explains why the past perfect is needed; the software just highlights it red." (TR-18)

Theme 3 – Cognitive Load, Technical Hurdles, and Equity

Sub-theme 3A: Interface overload
Six AI participants described "badge distraction" or "pop-up fatigue," consistent with CLT predictions about extraneous load.

Sub-theme 3B: Connectivity constraints
"Our Wi-Fi reset twice this week; I lost my streak and it killed my motivation." (AI-31)

Sub-theme 3C: Emotional dip/guilt cycle
Broken streaks triggered guilt for some learners and led to avoidance—an affective pattern resembling the attrition curves in app-store telemetry.

4.2.4 Cross-condition Engagement Vocabulary

A word-frequency query revealed divergent emotional registers.

Table 4. Word Frequency Comparison of Emotional Language in AI vs. Traditional Learner Interviews

Descriptor	Tokens (AI)	Tokens (Traditional)	Interpretation
fun / enjoyable	46	11	AI framed study as entertainment.
clear / structured	19	39	Traditional valued teacher-led order.
challenge / difficult	23	17	Cognitive load noted by both groups.
reward / badge / points	31	2	Gamification unique to AI condition.

4.2.5 Triangulation with Usage Analytics

Students in the upper quartile of usage minutes cited "motivation" or "achievement" 2.7 times more frequently than lower-usage peers, which corroborates the regression finding that time-on-task predicts gain. Conversely, the five AI learners with < 3 hours total usage contributed 14 frustration codes, which underscores the attrition risk noted in Theme 3.

4.2.6 Synthesis of Qualitative Insights

Collectively, interviews suggest the AI tool's competitive edge stems from immediate, granular feedback and a sense of autonomous progression. However, these advantages introduce new stressors, namely streak loss, accent mis-parsing, and interface clutter, which may dampen motivation if not addressed adequately. Traditional instruction, while slower, provides socio-emotional support and explicit rule explanations but lacks fine-grained, on-demand feedback. The two modalities thus appear complementary rather than mutually exclusive.

4.3 Integrated Summary

Quantitative analyses revealed significantly larger learning gains for the AI-assisted cohort, with effect sizes in conformity with the upper range of current meta-analytic estimates. Qualitative findings converge on a plausible mechanism: adaptive gamification and immediate feedback stimulate sustained engagement and deliberate practice, thereby extending learning beyond the temporal limits of classroom instruction. At the same time, the data highlight boundary conditions. Specifically, digital access, cognitive-interface load, and accent bias require careful instructional design and policy support to fully leverage AI's potential.

5. Discussion

The mixed-methods evidence presented here indicates that an AI-assisted programme can substantially exceed a time-constrained, teacher-centred curriculum in improving freshmen's receptive English skills. Quantitatively, the adjusted between-group gain of 12 TOEFL points on average places the present effect in the upper quartile of recent meta-analytic distributions (Lee & Lee, 2024; Xu et al., 2025). Because baseline proficiency was controlled, the advantage appears across the ability spectrum rather than being confined to higher-level subjects. Qualitatively, students repeatedly attributed their progress to the immediacy and granularity of speech-recognition feedback and to the motivational pull of micro-goals—daily streaks, progress bars, and badges. These features resonate with Swain's (2005)

output hypothesis, which posits that error-driven noticing accelerates automatization, and with self-determination theory's competence and autonomy needs (Deci & Ryan, 2000). Platform telemetry underscored the link as each additional thirty minutes of practice predicted a 1.8-point score increment, a dose–response pattern which is consistent with desirable-difficulty research (Kang, 2016).

Yet the qualitative record also reveals the fragile side of gamified autonomy. Interface clutter, accent-bias mis-parsing, and connectivity lapses led to frustration episodes that remind us of Sweller's (2019) warnings about intervening cognitive load. Low-usage learners, in particular, described "badge fatigue" and guilt after streak breaks, which echo self-discrepancy theory's prediction that when self-standards are not met, this generates negative affect (Higgins, 1987). These findings suggest that design elegance and infrastructural reliability are prerequisites for sustained AI efficacy, and they challenge claims that technology alone can democratize language learning.

From a theoretical standpoint, the data reinforce the view that AI amplifies rather than overturns established second-language acquisition mechanisms. Adaptive spacing algorithms embody desirable-difficulty principles; real-time recasts fulfil interactionist conditions for noticing; and scaffolded dashboards extend Vygotskian zones of proximal development by offering support on-demand. In sociocultural terms, learners framed the platform as a "coach," a metaphor that casts AI as a mediating force whose value is realized through human–machine partnership rather than autonomous instruction.

Practically, the evidence argues for a rebalancing of teacher and machine roles. Automated pattern recognition handles high-volume, low-complexity feedback (grammar drills, segmental pronunciation, lexical recycling) and releases classroom time for discourse-level work, strategic competence, and socio-emotional support. A blended schedule of three forty-minute AI homework sessions paired with two in-person seminars appears viable, provided that onboarding sessions teach students how to interpret dashboard analytics and uncertainty bands. Institutions in low-bandwidth regions should preload offline modules or supply subsidised data bundles, measures that reflect UNESCO's (2024) call for "accompanied infrastructure" in digitally divided contexts.

The results also speak to assessment reform. Because the platform stores item-level difficulty estimates and longitudinal accuracy records, it can generate detailed progress files that complement or, in some cases, replace high-stakes end-of-term tests. Yet adopting continuous analytics raises questions of data ownership, retention, and consent. Saudi regulators may need to articulate safeguards similar to European "data-minimization" statutes to prevent adaptive-learning logs from becoming surveillance repositories. Teachers, for their part, will require professional development that explains algorithmic uncertainty and confidence intervals so that auto-generated projections do not turn into self-fulfilling labels.

Interview data highlight another design arena: resilience-oriented support features. Streak loss triggered guilt and disengagement among several participants, an affective matter that parallels "achievement-gap anxiety" documented in other gamified platforms (Edwards, 2022). Introducing "grace days" that pause streak counts during verified connectivity outages, or adaptive streak targets adjusted to individual practice histories, could support motivation without reducing overall engagement (Xiu-Yi, 2024). Addressing streak anxiety explicitly in class may further boost learner autonomy.

Accent inclusivity poses an equally significant ethical challenge in the Saudi Arabian setting. While TruAccent® generally delivers accurate segmental feedback, it might still flag certain Gulf-Arabic vowels, especially the long /ɑː/ and emphatic consonants such as /ṣ ḍ ṭ / as errors, which suggests that engines trained on mainstream English accents may struggle with some regional phonetic features. Systematic misclassification risks turning perfectly intelligible local pronunciations into pathologies and, as a result, harm learners' linguistic self-confidence. One viable remedy is a GCC-wide speech-corpus repository that comprises annotated audio from universities in Saudi Arabia, the United Arab Emirates, Oman, and Bahrain. To complement the larger speech corpus, the software should add explainable-AI overlays. For example, a simple color-coded bar or heat-map could highlight the exact vowel or consonant sound that triggered the "error" message. Seeing this visual cue would let students judge whether the problem is a real intelligibility issue or just a difference in local accent. Clear, transparent feedback of this kind is more likely to guide helpful self-correction and less likely to discourage learners.

Several limitations are in order. First, the sample was restricted to a single cohort of Saudi freshmen; replication with younger learners, adult professionals, and different proficiency bands would strengthen external validity. Second, the six-week window cannot address long-term retention or transfer to productive skills, although speech-recognition logs hint at oral-fluency benefits. Third, only one commercial platform was examined; chatbots that foreground discourse management or VR environments that simulate pragmatic contexts may result in different learning outputs. Fourth, qualitative insights relied on self-report; triangulating with classroom observation or think-aloud protocols would further mitigate social-desirability bias.

Future work should, therefore, adopt multi-site, mixed-methods designs (Creswell & Clark, 2018) that manipulate dosage, interface design, and teacher mediation while tracking both cognitive and affective developments over full academic years. Latent-profile analysis could classify learners into engagement types—high-usage enthusiasts, steady mid-range users, and low-usage strugglers—and examine how each group responds to design adjustments. Discourse-analytic measures and pragmatic-competence rubrics would extend outcomes beyond receptive scores to communicative performance. Finally, collaborative regional projects to curb bias in ASR engines would ensure that technological gains do not come at the cost of accent discrimination.

In sum, when deployed with pedagogical intention, cultural sensitivity, and infrastructural support, AI can accelerate receptive-skill growth and foster self-regulated study habits. It does so not by replacing teachers but by freeing them to focus on higher-order tasks that machines cannot yet handle—critical thinking, pragmatic implicature, and ethical guidance. The key task going forward is to build a learning environment where teachers and AI tools work side by side—each doing what it does best—so that the technology moves from a promising

concept to an everyday tool that helps all students learn English more effectively.

6. Conclusion

This study set out to examine the impact of an AI-assisted platform on Saudi university freshmen's English-language development and to explore the learner-centred mechanisms that might account for any advantage over a traditional curriculum. Across six weeks, students who engaged with adaptive, feedback-rich software improved by an average of 12 TOEFL ITP points, which is roughly twice the gain achieved by peers following a time-matched, teacher-led syllabus. Effect sizes were large and remained robust after covariance adjustment, which is consistent with recent meta-analytic estimates for AI interventions. Interview evidence converged with the numerical data: Participants said their progress came from receiving quick, detailed feedback and from small game-like goals that helped them feel in control of their learning. Data from the platform backed this up: every extra 30 minutes of practice was linked to a 1.8-point improvement in test scores.

At the same time, the interviews revealed some important limits. Problems, such as poor internet, accent-related errors, and confusing app design led to frustration, especially for students who did not use the platform much. These challenges show how important it is to have strong internet access, speech systems that understand local accents, and simpler user interfaces if AI is going to support all learners fairly. They also highlight that teachers are still essential—for emotional support, real-world language examples, and ethical guidance—aspects that current AI tools cannot fully provide.

In sum, artificial intelligence does not replace established methods and theories of second language learning; rather, it reinforces processes that are already known to be effective. Its adaptive review features support the principle of "desirable difficulty," its immediate corrective feedback facilitates the interactionist notion of noticing, and its step-by-step dashboards extend Vygotsky's concept of scaffolded support. The findings therefore support a blended instructional model in which algorithms are assigned repetitive, lower-level corrective tasks, while classroom time is reserved for more complex activities such as discussion, analysis, and critical engagement. To implement this approach effectively at scale, policymakers should: (1) fund the development and distribution of offline learning materials for regions with limited internet access, (2) ensure that AI systems provide clear, explainable feedback that helps learners understand and address their errors, and (3) support regional initiatives aimed at adapting speech-recognition engines to local accents and linguistic features.

Future studies should run for at least a full academic year, include speaking and writing tests, and compare different AI tools—such as chatbots, pronunciation apps, and immersive VR—to see which ones work best for different types of learners. Tracking how various "user profiles" engage over time, and following their motivation through interviews or classroom observation, would also show whether the initial boost from gamified features can last.

Overall, AI can speed up reading and listening gains and help students develop better study habits—provided it is used thoughtfully, respects local culture, and is backed by reliable infrastructure. The next step is to build a learning environment where teachers and AI tools complement each other: machines handle routine feedback, while educators focus on deeper communication and critical skills. If done well, this partnership can turn AI from a promising idea into a practical driver of inclusive, effective language education.

Acknowledgments

Not Applicable.

Authors' contributions

Not Applicable (Article is single-authored)

Funding

There is no funding associated with this study.

Competing interests

There are no financial or non-financial interests directly or indirectly related to this study.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Ahmad, H. (2023). Glocalizing English language teaching in the Arabian Gulf: professional development's mediating influence. *Register Journal, 16*(2), 301-322. https://doi.org/10.18326/register.v16i2.301-322

Anderson, J. R. (1983). *The architecture of cognition.* Harvard University Press.

Arab News. (2024, December 19). *Saudi budget 2024: Education sector allocated SR 201 billion for digital transformation*. Retrieved from https://www.arabnews.com/node/2440586

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Breese, J. L., Rebman, C. M., & Levkoff, S. (2024). State of student perceptions of AI (circa 2024) in the United States. *Issues in Information Systems, 25*(4), 311-321.

Brown, H. D. (2014). *Principles of language learning and teaching (6th ed.).* Pearson.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47. https://doi.org/10.1093/applin/I.1.1

Chen, A., Zhang, Y., Jia, J., Liang, M., Cha, Y., & Lim, C. P. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning. *Journal of Computer Assisted Learning, 41*, e13064. https://doi.org/10.1111/jcal.13064

Chen, X., Li, J., & Ye, Y. (2024). A feasibility study for the application of AI-generated conversations in pragmatic analysis. *Journal of Pragmatics, 223*, 14-30. https://doi.org/10.1016/j.pragma.2024.01.003

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. https://doi.org/10.21236/AD0616323

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Communications, Space & Technology Commission. (2023). *Annual ICT indicators report 2023*. Riyadh: CST.

Creswell, J. W., & Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Dandu, G., Charyulu, G. M., & Kumari, K. L. (2024). AI-driven language learning: The impact of Rosetta Stone on ESL students' speaking proficiency. *Language Learning & Technology, 28*(2), 134-162. https://doi.org/10.21659/rupkatha.v16n4.09

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Digital Education Council. (2024). *Global AI student survey 2024.* Retrieved from https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024

Duolingo Research Team. (2024). *The 2024 Duolingo language report*. Duolingo.

Edwards, J. (2022). Gamification, anxiety, & motivation in second language learners: A qualitative systematic review. *Language Education and Technology, 2*(2), 98-127.

Ellis, R. (2012). *Language teaching research and language pedagogy.* Wiley-Blackwell. https://doi.org/10.1002/9781118271643

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2018). *Task-based language teaching: Theory and practice* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108643689

European Commission. *Digital Education Action Plan 2021 2027.* Retrieved from https://education.ec.europa.eu/focus-topics/digital-education/action-plan

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Gass, S. M., & Mackey, A. (2006). *Input, interaction, and output in second language acquisition.* Routledge. https://doi.org/10.1075/aila.19.03gas

Górriz, J. M., Álvarez-Illán, I., Álvarez-Marquina, A., Arco, J. E., Atzmueller, M., Ballarini, F., & Ferrández-Vicente, J. M. (2023). Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. *Information Fusion, 100*, 101945. https://doi.org/10.1016/j.inffus.2023.101945

Grand View Research. (2024). *AI in education market size, share & trends analysis report, 2025 – 2030*. Retrieved from https://www.mordorintelligence.com/industry-reports/ai-in-education-market

Harper, D., Bowles, A. R., Amer, L., Pandža, N. B., & Linck, J. A. (2021). Improving outcomes for English learners through technology: A randomised controlled trial. *AERA Open, 7*(1), 1-20. https://doi.org/10.1177/23328584211025528

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*(3), 319-340. https://doi.org/10.1037/0033-295X.94.3.319

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Penguin.

Kang, S. H. K. (2016). Spaced repetition promotes efficient and effective learning. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 12-19. https://doi.org/10.1177/2372732215624708

Krashen, S. (1985). *The input hypothesis: Issues and implications.* New York: Longman.

Lee, H., & Lee, J. H. (2024). The effects of AI-guided individualised language learning: A meta-analysis. *Language Learning & Technology, 28*(2), 134-162.

Li, Y., Zhou, X., Yin, H. B., & Chiu, T. K. F. (2025). Design language learning with artificial intelligence (AI) chatbots: A systematic review. *Smart Learning Environments, 12,* 24. https://doi.org/10.1186/s40561-025-00379-0

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). Academic Press. https://doi.org/10.1016/B978-012589042-7/50015-3

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361-365. https://doi.org/10.1037/0033-2909.111.2.361

Miles, M. B., Huberman, A. M., & Saldaña, J. (2019). *Qualitative data analysis: A methods sourcebook* (4th ed.). Sage.

Richards, J. C., & Rogers, T. S. (2014). *Approaches and methods in language teaching (3rd ed.)*. Cambridge University Press.

Sharma, K., Giannakos, M., & Dillenbourg, P. (2020). Eye-tracking and artificial intelligence to enhance motivation and learning. *Smart Learning Environments, 7*, 1-19. https://doi.org/10.1186/s40561-020-00122-x

Skinner, B. F. (1957). *Verbal behavior.* Appleton-Century-Crofts. https://doi.org/10.1037/11256-000

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Routledge.

Sweller, J. (2019). *Cognitive load theory and educational technology*. Springer. https://doi.org/10.4324/9780429283895-1

Telecom Review. (2025, February 8). *Saudi Arabia earmarks $40 billion for AI under NSDAI and Vision 2030*. Retrieved from https://www.telecomreview.com/articles/ai/7515-saudi-arabia-40b-ai-investment

The Business Research Company. (2025). *AI in education market forecast 2025–2030*. [TBR Company Reports]

UNESCO. (2024). *AI and the digital divide: Policy brief.*

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Xiu-Yi, W. U. (2024). AI in L2 learning: a meta-analysis of contextual, instructional, and social-emotional moderators. *System, 103498*. https://doi.org/10.1016/j.system.2024.103498

Xu, G., Yu, A., & Liu, L. (2025). A meta-analysis examining AI-assisted second-language learning. *International Review of Applied Linguistics in Language Teaching.* Advance online publication. https://doi.org/10.1515/iral-2024-0213