frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 November 2025
DOI 10.3389/feduc.2025.1614680

OPEN ACCESS

EDITED BY
Aslina Baharum,
Sunway University, Malaysia

REVIEWED BY
Eduardo Encabo-Fernández,
University of Murcia, Spain
Hieronimus Canggung Darong,
Santu Paulus Indonesian Catholic University, Indonesia

*CORRESPONDENCE
Mutasim Al-Deaibes
maldeaibes@aus.edu

RECEIVED 22 April 2025
ACCEPTED 08 October 2025
PUBLISHED 24 November 2025

CITATION
Aldamen H, Almashour M, Al-Deaibes M and AlSharefeen R (2025) Testing Krashen's input hypothesis with AI: a mixed-methods study on reading input and oral proficiency in EFL. *Front. Educ.* 10:1614680.
doi: 10.3389/feduc.2025.1614680

Testing Krashen's input hypothesis with AI: a mixed-methods study on reading input and oral proficiency in EFL

Hesham Aldamen[1], Mohamad Almashour[1], Mutasim Al-Deaibes[2]* and Rami AlSharefeen[3]

[1]Department of Language and Literature, The University of Jordan, Al Jubeiha, Jordan, [2]American University of Sharjah, Sharjah, United Arab Emirates, [3]Rabdan Academy, Abu Dhabi, United Arab Emirates

This study investigated the impact of AI-generated graded reading materials on the oral proficiency of adult EFL learners in a six-month intervention. Ninety participants generated weekly texts using proficiency-aligned prompts and were assessed through pre- and post-intervention ACTFL Oral Proficiency Interviews, complemented by learner reflective journals. Quantitative results suggested significant proficiency gains across all initial levels, while thematic analysis of journals highlighted perceived benefits in vocabulary development, autonomy, and fluency. Together, these findings provide preliminary evidence consistent with Krashen's Input Hypothesis, while also linking AI-mediated reading to broader frameworks of scaffolding, vocabulary acquisition, and cognitive load management. At the same time, important limitations must be noted. The study relied on a single non-certified rater, lacked a control group, and did not systematically monitor the linguistic properties of AI-generated texts. Attrition was concentrated among Novice High learners, raising concerns about bias in proficiency outcomes. These constraints require cautious interpretation, and the results should be viewed as suggestive rather than definitive. Despite these limitations, the study contributes to current discussions on AI in language education by illustrating how generative tools can provide scalable, proficiency-aligned input. It offers preliminary insights into the potential of AI-mediated reading to support oral proficiency development, while underscoring the need for more rigorous designs in future research.

KEYWORDS

AI in language learning, comprehensible input, oral proficiency, EFL, ACTFL, generative AI, input-based acquisition, reading intervention

1 Introduction

The integration of generative artificial intelligence (GAI) into language education has opened new avenues for personalized input delivery, autonomous learning, and scalable instruction. Among these developments, AI-powered platforms such as large language models (LLMs) offer the novel capability of generating level-appropriate written texts tailored to individual learners' linguistic needs (Kasneci et al., 2023; Van Brummelen, 2019). While much of the current discourse around AI in education emphasizes feedback, assessment, or content generation, relatively little is known about how AI-generated reading input may influence language development over time, particularly in relation to oral proficiency (Wang and Dang, 2024; Guo et al., 2024; Lee and Moore, 2024; Alnemrat et al., 2025). This study addresses that gap by empirically testing whether sustained exposure to AI-generated, level-aligned written input can lead to measurable gains in learners' spoken language skills.

This investigation builds on Krashen's Input Hypothesis (Krashen, 1982), which posits that language acquisition occurs when learners are exposed to comprehensible input slightly beyond their current level of competence. Although often associated with listening input, Krashen argued that reading can serve as a powerful source of language acquisition, including the development of speaking proficiency (Krashen, 2004). However, empirical studies isolating the impact of reading alone, particularly AI-mediated reading, on oral proficiency are scarce. This study contributes to the field by implementing a six-month reading intervention in which learners read one AI-generated text per week, tailored to their proficiency level based on ACTFL guidelines (American Council on the Teaching of Foreign Languages, 2024), with no interaction, discussion, or direct instruction involved.

While this study builds on Krashen's Input Hypothesis, it is important to distinguish it from other major perspectives in second language acquisition. The Interaction Hypothesis (Long, 1996) emphasizes the role of conversational negotiation and feedback, and Swain's Output Hypothesis (Swain, 1985) highlights the importance of learner production for noticing and restructuring interlanguage. By contrast, the present study isolates the effects of written input delivered through AI-generated texts, without interaction or output requirements. This design does not directly test the Interaction or Output Hypotheses but instead provides preliminary evidence consistent with input-based accounts of language development.

In addition to Krashen's theory, this study is informed by several supporting frameworks. Vygotsky's Zone of Proximal Development (Vygotsky, 1978) highlights the importance of providing input that is just beyond what a learner can do independently, a principle operationalized here through AI-generated texts aligned to learners' proficiency levels. Nation's vocabulary framework (Nation, 2001) underpins the lexical control embedded in the reading prompts, while Cognitive Load Theory (Sweller, 1988) supports the intentional management of complexity to maintain learner engagement and comprehension.

As AI technologies become increasingly embedded in educational environments, understanding their role in input-driven language acquisition becomes both a theoretical and practical imperative. Can AI, when guided by principled instructional design, deliver reading materials that contribute meaningfully to the development of oral language? And how do learners perceive this process over time? These questions are especially relevant in large mixed-proficiency classrooms, where individualized input has traditionally posed logistical challenges.

This study aims to address these issues by investigating the following research questions:

1. Does regular exposure to AI-generated, proficiency-level-aligned written input over 6 months improve oral proficiency among adult EFL learners?
2. Does the effectiveness of AI-generated reading input vary across Novice High, Intermediate Mid, and Advanced Mid learners?
3. How do learners perceive the usefulness and impact of AI-generated, level-appropriate reading input on their oral language development?

By combining quantitative proficiency outcomes with qualitative reflections, this study contributes to the evidence base supporting input-based learning models while also offering insight into the pedagogical role of AI-generated content in second language education.

2 Literature review

2.1 Comprehensible input and language acquisition

The foundation of this study rests on Krashen's (1982) Input Hypothesis, which posits that language acquisition occurs most effectively when learners are exposed to comprehensible input that is slightly above their current level of competence. Unlike learning through explicit instruction or correction, acquisition, in Krashen's view, is largely subconscious, developing as learners make sense of input they understand. While Krashen acknowledged that both written and oral language can serve as sources of comprehensible input, he emphasized the role of reading as a particularly rich and low-anxiety source for language development (Krashen, 2004). Sustained engagement with level-appropriate reading materials has been associated with gains in vocabulary, grammatical accuracy, and fluency.

Importantly, Krashen suggests that oral proficiency can develop from reading alone, provided learners are consistently exposed to meaningful and contextually appropriate input. The present study builds on this perspective by focusing exclusively on written input as the primary instructional intervention. At the same time, it acknowledges that other theories, such as Long's (1996) Interaction Hypothesis and Swain's (1985) Output Hypothesis, highlight the complementary roles of negotiation and production in SLA. By intentionally omitting these dimensions, the study maintains a clear theoretical focus on input, while recognizing that interaction and output remain critical components of broader language development.

Recent work also suggests that generative AI offers a new context for examining these classic frameworks, since large language models can deliver individualized, level-aligned reading materials at scale (Kasneci et al., 2023; Wang et al., 2025). This provides an opportunity to investigate whether Krashen's claims regarding input-driven acquisition extend to AI-mediated environments, where learners receive texts dynamically generated to match their proficiency.

2.2 Level-appropriate input and the zone of proximal development

Vygotsky's (1978) Zone of Proximal Development (ZPD) underscores the importance of providing learners with input that is appropriately challenging. The ZPD defines the space between what a learner can do independently and what they can accomplish with support. In the context of language learning, this zone can be operationalized by aligning input difficulty with the learner's current proficiency level while gradually nudging them toward more advanced performance. Instruction within this zone enables learners to scaffold their understanding and internalize new linguistic forms, particularly when the material is neither too simple to be redundant nor too complex to be discouraging.

To implement this framework in a scalable way, the present study used AI-generated texts tailored to learners' functional language

abilities based on the ACTFL Proficiency Guidelines (American Council on the Teaching of Foreign Languages, 2024). With carefully engineered prompts, the AI generated level-appropriate texts that increased slightly in complexity over time, thereby delivering sustained input within each learner's ZPD.

2.3 Lexical control and vocabulary acquisition

Nation's (2001) framework for vocabulary acquisition emphasizes the importance of repeated exposure to high-frequency vocabulary in comprehensible input. Learners must encounter words in diverse contexts to build depth of lexical knowledge. Reading is particularly effective for this purpose, especially when materials are written at an appropriate level to allow for repeated exposure without overwhelming cognitive processing.

In this study, the prompts guiding AI text generation are designed to control for lexical load by favoring common, high-utility vocabulary aligned with ACTFL descriptors for each proficiency band. This control supports incidental vocabulary acquisition and contributes to the development of fluent and flexible spoken language. Nation's focus on coverage and frequency aligns closely with the goals of the intervention and informs the prompt design process.

2.4 Managing input complexity through cognitive load theory

Cognitive Load Theory (Sweller, 1988) offers a complementary perspective by highlighting the mental effort required to process input. Learning is most effective when extraneous cognitive load is minimized, and intrinsic load is appropriately matched to the learner's capacity. If reading input is too complex, learners may become cognitively overwhelmed and unable to extract useful patterns or meaning. Conversely, overly simple texts may fail to challenge or advance linguistic development.

The use of AI to generate reading input presents both opportunities and risks in this regard. On one hand, prompt engineering allows for careful control of text difficulty, sentence structure, and topic familiarity. On the other hand, AI output may vary in complexity or relevance, necessitating prompt refinement and testing to ensure alignment with learner needs (Mollick and Mollick, 2023a, 2023b). In this study, prompt design is informed by principles of cognitive load management to ensure that input remains accessible while still promoting language development.

2.5 AI as a tool for scalable, adaptive language input

Recent developments in generative artificial intelligence have introduced new possibilities for adaptive language learning environments. Large Language Models (LLMs), such as GPT-4, can be prompted to simulate tutors, mentors, and other instructional roles, producing personalized textual content across a range of complexity and subject matter (Mollick and Mollick, 2023b; Yoon et al., 2023; Mahapatra, 2024). When combined with structured instructional design, these tools offer the potential to personalize language input at scale, reducing the need for instructor intervention and allowing for continuous learner engagement.

In contrast to instructor-led simulations, which often require extensive scripting and high development costs, prompt-based input generation allows learners to receive custom texts in real time, aligned with their proficiency level and learning goals. This approach democratizes access to tailored language input and aligns well with Krashen's emphasis on learner autonomy and low-anxiety acquisition environments. However, the success of such tools depends heavily on prompt quality, learner training, and ethical oversight to mitigate issues of hallucination, inconsistency, and cultural bias (Mollick and Mollick, 2023a, 2023b; Mahapatra, 2024; Mollick and Mollick, 2024; Mzwri and Turcsányi-Szabo, 2025).

Recent empirical research has begun to examine how these tools affect SLA outcomes. A scoping review of generative AI in language education found that most applications have targeted writing, grammar, and vocabulary development, with relatively limited attention to speaking proficiency (Wang et al., 2025). Han (2024) similarly called for systematic research into how AI-generated input can shape oral proficiency, noting that current studies often prioritize short-term engagement measures over longitudinal outcomes. Studies of AI-powered chatbots indicate that they can foster gains in speaking confidence, interactional skills, and learner engagement (Du and Daniel, 2024), while mixed-methods investigations of multiple AI tools, including ChatGPT, Grammarly, and Duolingo, report improvements in vocabulary, writing accuracy, and motivation (Seddik, 2025). Despite these promising developments, few studies have directly tested whether sustained exposure to AI-generated reading input alone can foster oral proficiency. The present study addresses this gap by examining whether extended engagement with AI-generated, proficiency-aligned texts contributes to measurable speaking gains.

2.6 The role of reflection in deepening language awareness

Although the primary intervention in this study is written input, learner reflection plays a supporting role by facilitating metacognitive awareness. Reflective practices have long been associated with deeper learning outcomes, as they prompt learners to consider how they engage with input and what strategies contribute to progress (Dewey, 1933; Schön, 1983). In language learning, reflective journaling has been shown to reinforce vocabulary retention, self-regulation, and goal-setting behaviors.

In the context of this study, learners maintain reflective journals throughout the intervention but submit them only at its conclusion. This design supports ongoing self-monitoring without introducing external evaluation during the learning process. Thematic analysis of these journals provides insights into learners' evolving perceptions of AI-generated texts, their engagement with the materials, and their perceived growth in oral proficiency.

Taken together, these frameworks underscore the relevance of investigating AI-generated reading input as a potential driver of oral proficiency development. Krashen highlights the sufficiency of comprehensible input, Vygotsky emphasizes developmental alignment through scaffolding, Nation underscores the role of lexical frequency,

and Sweller illustrates the importance of managing cognitive load. Reflection further supports learner awareness and metacognition. While each theory offers a lens for understanding input-driven learning, recent reviews point out that empirical evidence on sustained AI-generated input and its oral proficiency outcomes remains scarce (Han, 2024; Wang et al., 2025). By integrating these classic perspectives with emerging research on AI in SLA, the present study addresses this gap through a six-month intervention designed to test whether generative AI can provide effective, scalable input across different proficiency levels.

3 Methodology

3.1 Research design

This study employs a longitudinal, quasi-experimental, mixed-methods design to investigate the effect of AI-generated, proficiency-aligned written input on the oral proficiency development of adult EFL learners. The design integrates both quantitative and qualitative components to capture measurable changes in oral proficiency and explore learners' perceptions of the AI-mediated input. The six-month intervention involves participants independently generating weekly reading materials using pre-designed, well-tested, and well-engineered prompts provided by the instructor. Participants will read these AI-generated materials without peer or instructor interaction, in line with Krashen's Input Hypothesis, which isolates the role of comprehensible written input in oral language acquisition. Participants were provided with detailed instructions prior to the intervention; these are included in Appendix A.

Pre- and post-intervention assessments are conducted using the ACTFL Oral Proficiency Interview (OPI), while qualitative insights are gathered from reflective journals. The mixed-methods approach enables both comparative analysis of proficiency outcomes and exploration of learner experiences with AI-generated texts.

3.2 Participants

The study includes 90 adult undergraduate students enrolled in English as a Foreign Language (EFL) courses at a public university in Jordan. Participants are stratified into three groups based on initial proficiency level: Novice High, Intermediate Mid, and Advanced Mid, with 30 students in each group. These levels are determined through individual ACTFL Oral Proficiency Interviews conducted by the researcher, who received extensive training in ACTFL protocols during his doctoral studies in the United States.

All participants voluntarily agree to participate and provide written informed consent. Demographic data such as age, gender, and academic major are collected for descriptive analysis. Participants have no history of extended residence in an English-speaking country, ensuring relative homogeneity in their language exposure context.

3.3 Materials and instruments

Three primary instruments are used in this study: the ACTFL Oral Proficiency Interview (OPI), AI-generated reading texts created through ChatGPT, and participant reflective journals.

The ACTFL OPI is administered at the beginning and end of the intervention period to assess changes in speaking proficiency. Ratings are assigned according to ACTFL proficiency levels and are treated as ordinal data for the purpose of analysis. The same researcher conducted both interviews to ensure consistency across assessments. While the researcher received extensive ACTFL training during doctoral studies in the United States, they are not formally certified by ACTFL. As such, external rater reliability cannot be assured, and all proficiency level assignments should be interpreted within this methodological constraint.

Participants generate their own reading texts using prompts provided by the researchers. These prompts are engineered to produce level-appropriate input based on the ACTFL Proficiency Guidelines (American Council on the Teaching of Foreign Languages, 2024). Each prompt controls for text type, vocabulary frequency, sentence complexity, and communicative function. However, the specific texts generated by each participant were not reviewed or analyzed by the researcher prior to reading. This design choice reflects a learner-driven, ecologically valid implementation of AI tools, but it also introduces potential variability in lexical range, topical relevance, or appropriateness across participants. Participants generate and read one new text each week for 6 months. The proficiency-aligned prompts are included in Appendix B.

Reflective journals serve as the qualitative instrument. Although students maintain the journals throughout the intervention, they submit them only once, at the end. The journals document their perceptions of the usefulness, clarity, difficulty, and impact of the AI-generated texts on their speaking abilities.

To strengthen rating reliability, all OPIs were audio-recorded and archived. The researcher had received extensive training in ACTFL procedures during doctoral studies and conducted pilot calibrations using benchmark recordings prior to data collection. While an ACTFL-certified external rater was not available for this study, a subset of 20% of the recordings was independently reviewed by a second trained researcher, and agreement rates were compared. This process provided a partial reliability check, although the absence of formal ACTFL certification remains a limitation that should be acknowledged in interpreting results.

3.4 Content analysis of AI-generated texts

To assess the quality and alignment of the AI-generated input, a content analysis was conducted on a stratified sample of 15 texts, five from each proficiency band: Novice High, Intermediate Mid, and Advanced Mid. Texts were analyzed for lexical coverage, lexical diversity, and syntactic complexity. Lexical coverage was measured against the first 1,000 and 2,000 most frequent word families using Nation's framework. Lexical diversity was estimated using type–token ratio. Syntactic complexity was evaluated through mean sentence length, clauses per sentence, and frequency of subordinate clauses. Analyses were conducted using AntWordProfiler and Coh-Metrix. These measures provided an indication of whether the AI-generated input reflected the lexical and syntactic expectations of the targeted ACTFL proficiency levels.

3.5 Procedures

The study includes three phases: pre-intervention, intervention, and post-intervention. In the pre-intervention phase, all participants complete an OPI to determine their initial proficiency level. To ensure that participants were familiar with the AI platform and the structured use of prompts, a short orientation session was conducted at the beginning of the study. During this session, the researcher explained how to input the assigned prompts, interpret the AI-generated texts, and follow the weekly reading and reflection procedures outlined in Appendix A. Participants were also advised on how to handle irrelevant or unclear AI outputs using consistent redirection strategies.

3.5.1 OPI

Triangulation of OPI outcomes with reflective journal insights increases the credibility of the findings. The absence of peer or instructor interaction during the intervention controls for external variables, helping to isolate the effect of written input.

Attrition was monitored across the intervention period. Of the 90 learners who began the study, 82 completed both the pre- and post-intervention assessments. The eight learners who withdrew were all from the Novice High group, with reasons including scheduling conflicts and limited sustained engagement. Baseline demographic and proficiency characteristics of completers and non-completers were compared, and no significant differences were observed apart from the initial proficiency distribution. Nevertheless, the uneven attrition concentrated among lower-proficiency learners represents a potential source of bias and is considered in the interpretation of results.

3.6 Inter-rater reliability of oral proficiency ratings

All Oral Proficiency Interviews (OPIs) were audio-recorded to enable subsequent reliability checks and calibration. The primary rater, while extensively trained in ACTFL protocols during doctoral study in the United States, was not formally ACTFL-certified. To strengthen reliability, a subset of 18 pre- and post-intervention recordings (20 percent of the sample, evenly distributed across proficiency bands) was independently re-rated by a second rater with advanced training in ACTFL procedures. Prior to re-rating, both raters engaged in calibration using benchmark recordings aligned with ACTFL proficiency descriptors to ensure consistency in scoring criteria.

Inter-rater agreement was calculated using Cohen's κ for categorical proficiency levels and the Intraclass Correlation Coefficient (ICC) for ordinal consistency. Cohen's $\kappa = 0.82$ indicated substantial agreement, while the ICC (two-way random, absolute agreement) was 0.87 with a 95 percent CI of [0.78, 0.93], reflecting high consistency between raters. Discrepancies between raters were discussed and resolved, but reliability indices were calculated on initial ratings to provide an unbiased estimate of agreement.

3.7 Sensitivity analyses of effect sizes

To assess the robustness of the unusually large effect sizes observed, sensitivity analyses were conducted. Bootstrapped confidence intervals with 5,000 resamples were calculated for the Wilcoxon signed-rank effect size. In addition, tied ratings were conservatively recoded as non-improvements to evaluate whether the results remained significant under stricter assumptions. These procedures provided a test of whether the large effect sizes could be attributed to statistical artifacts.

3.8 Qualitative analysis

Learner journals were analyzed using thematic analysis (Braun and Clarke, 2006). The process included familiarization with the data, initial coding, theme generation, review, and definition. Two researchers independently coded the journals and compared results to ensure consistency. Discrepancies were discussed until consensus was reached, and themes were refined iteratively. Data saturation was assumed when no new themes emerged. Representative excerpts are presented in Table 1 to illustrate each theme.

3.9 Limitations

This study has several limitations. First, while the Oral Proficiency Interviews were conducted by a researcher with extensive ACTFL training during his doctoral studies in the United States, the researcher was not formally certified, which may affect the external reliability of the proficiency assessments. Second, although participants received detailed instructions and an orientation session, individual variation in how the AI prompts were used may have influenced the consistency of the generated input. Third, the study's focus on adult EFL learners in a university setting limits the generalizability of the findings to other age groups or learning contexts. Fourth, the exclusive use of written, non-interactive input excludes listening, speaking, and multimodal resources, which may reduce ecological validity when compared to real-world language environments. Finally, while the reflective journals provided valuable qualitative insight, they remain subjective in nature and may be shaped by participants' introspective abilities, engagement, and motivation.

3.10 Ethical considerations

Ethical approval for the study is obtained from the university's research ethics board. Informed consent is collected from all participants, and confidentiality is ensured through the use of

TABLE 1 Pre- and post-intervention ACTFL proficiency distributions (n = 90 Pre-OPI; n = 82 Post-OPI).

Proficiency level	Pre-OPI (n = 90)	Post-OPI (n = 82)
Novice high	30	10
Intermediate low	0	12
Intermediate mid	30	8
Intermediate high	0	22
Advanced mid	30	2
Advanced high	0	28

Post-OPI data includes only those participants (n = 82) who completed both assessments.

participant codes. No personally identifying information is included in the analysis or publication. Participants are informed of their right to withdraw at any point without consequence. Journals and assessment data are securely stored and accessible only to the researcher.

4 Results

4.1 Quantitative results

Eighty-two participants completed both the pre- and post-intervention Oral Proficiency Interviews (OPIs). Prior to the intervention, the 90 participants were evenly distributed across three ACTFL proficiency levels: Novice High ($n = 30$), Intermediate Mid ($n = 30$), and Advanced Mid ($n = 30$). Eight participants did not complete the post-intervention OPI.

Changes in post-intervention proficiency levels are presented in Table 1. Several participants advanced beyond their initial proficiency level, with no instances of regression observed.

4.1.1 Proficiency shifts by group

A breakdown of individual progress within each group further illustrates the extent of proficiency gains:

- Novice High Group ($n = 22$ completed post-OPI):
  - 10 participants remained at Novice High.
  - 12 participants advanced to Intermediate Low.
- Intermediate Mid Group ($n = 30$):
  - 8 participants remained at Intermediate Mid.
  - 22 participants advanced to Intermediate High.
- Advanced Mid Group ($n = 30$):
  - 2 participants remained at Advanced Mid.
  - 28 participants advanced to Advanced High.

The direction and magnitude of these shifts are summarized in Table 2.

4.1.2 Statistical significance of proficiency gains

To assess whether these gains were statistically significant, a Wilcoxon signed-rank test was conducted. The ACTFL levels were coded numerically as follows: Novice High = 1, Intermediate Low = 2, Intermediate Mid = 3, Intermediate High = 4, Advanced Mid = 5, and Advanced High = 6. Each participant's pre- and post-intervention level was paired and tested. The analysis revealed a statistically significant upward shift in proficiency levels, W = 0.00, $p < 0.001$ ($p = 3.43 \times 10^{-15}$). This result indicates that the observed improvements in oral proficiency were highly unlikely to be due to chance. Descriptive statistics revealed clear upward trends in median proficiency levels across all groups. The overall median ACTFL level increased from Intermediate Mid (Median = 3, IQR = 2–5) to Intermediate High (Median = 4, IQR = 3–6). For the Novice High group, the median increased from 1 to 2; for the Intermediate Mid group, from 3 to 4; and for the Advanced Mid group, from 5 to 6. To ensure full transparency, the distribution of paired differences is reported as follows: 62 positive differences, 20 ties, and 0 negative differences. The sum of negative ranks was therefore W = 0 W = 0 W = 0. Using the large-sample normal approximation, the Wilcoxon signed-rank test yielded $Z = 6.85$Z = 6.85Z = 6.85, $p < 1 \times 10 - 11$p < 1 \times 10^{−11} p < 1 × 10–11, with an effect size of $r = 0.87$r = 0.87r = 0.87, calculated as r = Z/Nr = Z/\sqrt{N}r = Z/N. A complementary sign test produced $p = 2.17 \times 10 - 19$p = 2.17 \times 10^{−19}p = 2.17 × 10–19. These extreme values are mathematically consistent with a dataset in which all non-tied participants improved, though they exceed the magnitude typically observed in educational interventions.

TABLE 2 Proficiency level changes by pre-intervention group ($n = 82$).

Pre-OPI group	Stayed at same level	Moved up one level	Moved up two levels
Novice high ($n = 22$)	10	12	0
Intermediate mid ($n = 30$)	8	22	0
Advanced mid ($n = 30$)	2	28	0

All participants who improved advanced by one proficiency level; no downward movement was observed.

4.2 Sensitivity analyses of effect sizes

Sensitivity analyses suggested that the observed proficiency gains were robust. Bootstrapped estimates of the Wilcoxon effect size produced a mean $r = 0.85$ with a 95 percent CI of [0.78, 0.90], consistent with the originally reported $r = 0.87$. When tied ratings were conservatively recoded as non-improvements, the Wilcoxon signed-rank test remained significant ($W = 8.0$, $p < 0.001$), with an adjusted effect size of $r = 0.79$. These findings indicate that while the effect sizes are unusually large for an educational intervention, they remain stable across multiple analytic approaches and are unlikely to be an artifact of the statistical method.

Attrition analysis further indicated that of the 90 participants who began the study, 82 completed the post-intervention OPI. Dropouts were not evenly distributed: 8 occurred in the Novice High group, whereas all Intermediate Mid and Advanced Mid learners completed the intervention. A Fisher's exact test indicated that attrition was significantly concentrated among Novice High participants ($p = 7.55 \times 10 - 5$p = 7.55 \times 10^{−5}p = 7.55 × 10–5). Tables 3, 4 summarize these results.

4.3 Inter-rater reliability of oral proficiency ratings

Analysis of the double-rated subset showed strong inter-rater agreement. Cohen's $\kappa = 0.82$ indicated substantial agreement, and the ICC (two-way random, absolute agreement) was 0.87 with a 95 percent CI of [0.78, 0.93], reflecting high consistency between raters. These results suggest that the oral proficiency ratings were reliable and that rater bias was unlikely to account for the observed proficiency gains.

4.4 Content analysis of AI-generated texts

To evaluate the quality and alignment of the AI-generated input, a sample of 15 texts (five per proficiency level) was analyzed for lexical

TABLE 3 Wilcoxon signed-rank and sign test results for oral proficiency gains.

Metric	Value
Positive differences	62
Ties	20
Negative differences	0
Wilcoxon WWW	0
ZZZ	6.85
ppp (two-tailed)	<0.00000000001
Effect size *r*	0.87
Sign test ppp	2.17×10^{-19}

Effect size r was calculated as Z/NZ/\sqrt{N}Z/N. Extreme values are consistent with a dataset in which all non-tied participants improved.

TABLE 4 Attrition by proficiency group.

Group	Started	Completed	Dropped	Dropout %
Novice high	30	22	8	26.7
Intermediate mid	30	30	0	0.0
Advanced mid	30	30	0	0.0
Total	90	82	8	8.9

Fisher's exact test indicated that attrition was significantly concentrated among Novice High learners ($p = 7.55 \times 10 - 5p = 7.55 \times 10^{-5}p = 7.55 \times 10–5$).

coverage, lexical diversity, and syntactic complexity. Results are presented in Tables 5, 6.

The analysis showed that Novice High texts were dominated by high-frequency vocabulary, with 88 percent of tokens drawn from the first 1,000 most frequent word families and a mean sentence length of 7.4 words. These texts contained minimal subordination and reflected short, formulaic sentence patterns consistent with ACTFL descriptors at this level.

Intermediate Mid texts demonstrated broader lexical range, with 72 percent of tokens within the first 2,000 word families. Mean sentence length increased to 11.8 words, and subordination appeared more regularly, averaging 8–10 instances per 100 sentences.

Advanced Mid texts exhibited greater lexical diversity (type–token ratio = 0.48) and more complex syntax, with mean sentence length of 18.2 words and frequent use of subordination (20 + instances per 100 sentences). Lexical coverage dropped to 59 percent within the first 2,000 word families, indicating exposure to lower-frequency vocabulary, which aligns with the expectations of advanced proficiency.

Taken together, these results indicate that the AI-generated texts reflected lexical and syntactic features appropriate to each proficiency band, providing learners with input aligned to their developmental stage.

4.5 Qualitative results

To explore participants' perceptions of AI-generated, level-aligned written input and its effect on their oral language development, a thematic analysis was conducted on 55 reflective journals submitted at the end of the six-month intervention.

TABLE 5 Inter-rater reliability of oral proficiency ratings (Subset of $n = 18$).

Measure	Value	95% CI	Interpretation*
Cohen's κ (categorical agreement)	0.82	—	Substantial agreement
ICC (two-way random, absolute)	0.87	[0.78, 0.93]	High consistency

Values based on 18 pre- and post-intervention OPI recordings (20 percent of sample). Reliability indices calculated on initial ratings prior to adjudication.

Journals were written independently by students throughout the study but submitted only once at the end of the intervention. Fourteen journals were excluded due to insufficient length, off-topic content, or lack of reflection. The analysis followed Braun and Clarke's (2006) six-phase process: familiarization with the data, generation of initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report.

Five major themes emerged from the analysis. These are summarized in Table 7, which provides descriptions and exemplar quotations illustrating learners' perceptions of autonomy, vocabulary development, fluency, content relevance, and challenges.

Themes were further supported by participant reflections and organized to represent a broad range of perspectives across all proficiency levels. All quotes are attributed using participant ID and ACTFL level.

4.5.1 Theme 1: clarity and comprehension support

Many participants, particularly at the Novice and Intermediate levels, reported that the AI-generated texts were readable and manageable. The accessibility of the input enabled them to engage with the content more independently and with less reliance on dictionaries or peers.

- P03 (Novice-High): "Good reading. Not many hard word. I can read all and I understand. I no need ask my friend help this time."
- P67 (Novice-High): "Before, I stop when reading English. Now I can read all. Still not perfect but better. Make me happy."
- P44 (Intermediate-Mid): "Sometimes the text was a bit long, but I still could get the meaning. I did not have to check dictionary too much."

4.5.2 Theme 2: confidence and motivation

Participants frequently reported that the reading texts contributed to improved self-confidence and greater willingness to speak in class. Several students linked this increased confidence to their ability to recall vocabulary and sentence structures encountered in the readings.

- P09 (Novice-High): "I feel brave now. I speak more when teacher ask me. Because I read before. I remember words from the text."
- P19 (Novice-High): "I think I want to read more. It not boring. I understand and that make me want to learn English."
- P58 (Intermediate-Mid): "I got used to reading in English. I still make mistake when speaking, but I try more. The reading help with that."

TABLE 6 Lexical and syntactic features of sampled AI-generated texts by proficiency level (n = 15).

Proficiency level	Lexical coverage (% within 1 k word families)	Lexical coverage (% within 2 k word families)	Type–token ratio	Mean sentence length (words)	Clauses per sentence	Subordination frequency (per 100 sentences)
Novice high	88%	95%	0.36	7.4	1.0	Rare (≤2)
Intermediate mid	65%	72%	0.41	11.8	1.4	Moderate (8–10)
Advanced mid	49%	59%	0.48	18.2	1.9	Frequent (20+)

Lexical coverage calculated with reference to Nation's word-family lists. Syntactic measures derived using Coh-Metrix. Values are means across five sampled texts per proficiency band.

TABLE 7 Thematic analysis of learner journals.

Theme	Description	Exemplar quotes
Increased autonomy	Learners reported greater independence in managing their own reading pace and choosing when to engage with texts.	"With AI texts, I could read on my own time and check meanings without asking the teacher."
Vocabulary expansion	Students emphasized repeated exposure to new words and expressions, which supported speaking confidence.	"The AI readings gave me many new words, and I started to use them when I spoke with my classmates."
Fluency development	Journals highlighted gradual improvement in oral fluency, linked to frequent practice with level-appropriate input.	"After some weeks, I felt I could speak faster without stopping to think so much."
Perceived relevance of content	Learners reflected positively when passages were engaging and aligned with personal interests, though some noted variability.	"Some topics were really interesting, like technology and travel, but others felt less useful to me."
Challenges and frustration	A minority reported difficulty with occasional complex passages or motivation lapses.	"One or two texts were too hard for me, and I lost interest quickly."

4.5.3 Theme 3: efficiency and accessibility

Participants across all proficiency levels valued the manageable pace of the intervention. Many described the single weekly reading task as appropriately balanced in terms of time, length, and difficulty. The ability to complete readings at their own pace was also noted as a benefit.

- P06 (Intermediate-Mid): "Having just one text each week was good. Not too much. I read it on my own time, no stress."
- P29 (Intermediate-Mid): "The length of the texts was okay. Not too short, not too long. I could understand most of it, so I finish reading."
- P45 (Advanced-Mid): "It was manageable with my schedule. The reading level was not too high. Sometimes a little easy, but still okay for review."

4.5.4 Theme 4: skepticism and trust issues

Several participants expressed lingering doubts about the accuracy and reliability of texts generated by ChatGPT, despite knowing the source. Concerns centered not on the origin of the content, but on whether AI output could be fully trusted without human verification. These reflections reflect a cautious stance toward AI as a learning tool and highlight students' tendency to cross-check or avoid relying on AI-generated texts for high-stakes tasks. For example, Participant P80 (Intermediate Mid) stated that they checked Google after reading, even though the source was clear, illustrating that the student is aware of the source but still feels the need to verify externally. Similarly, Participant P48 (Advanced Mid) reported not using the AI-generated content in writing due to doubts about its reliability, which reflects a strategic choice to limit dependence on AI-generated input for formal or academic purposes.

- P33 (Novice High): "I do not know is it right. Who write it? AI not same like teacher. Maybe is wrong."
- P80 (Intermediate Mid): "Sometimes I do not trust it 100%. It's AI and no name on the text. I check Google after reading."
- P48 (Advanced-Mid): "The facts seemed okay, but I still had doubts. I did not use it in writing because not sure it's reliable."

4.5.5 Theme 5: concerns about depth and over-reliance

Some participants, particularly at higher proficiency levels, indicated a desire for more intellectually demanding content. Others reflected on a tendency to rely exclusively on the AI-generated texts, potentially limiting their broader language exposure and critical engagement with other sources.

- P24 (Intermediate-Mid): "After some weeks, I noticed I stopped reading other English things. I think I got lazy. Just reading AI text and no more."
- P59 (Advanced-Mid): "It helped me with fluency, but I needed more analysis. Sometimes the ideas too basic. No much deep thinking."
- P76 (Advanced-Mid): "By the third month, I was kind of bored. Texts felt the same. I wanted more complexity, more academic language maybe."

These themes provide insight into participants' experiences with AI-generated input and offer qualitative support for the observed gains in oral proficiency. The next section will interpret these findings in light of the study's theoretical framework and pedagogical implications.

5 Discussion

This study investigated the impact of AI-generated, level-aligned written input on the oral proficiency development of adult EFL learners over a six-month period. The findings reveal that consistent exposure to tailored written texts led to statistically significant improvements in oral proficiency across all three initial proficiency groups. Additionally, learner reflections highlighted perceived benefits in comprehension, motivation, and confidence, while also raising concerns about depth, over-reliance, and trust in AI-generated content. This discussion interprets these results in light of the study's theoretical foundations and prior research.

The supplementary content analysis provided empirical evidence that the AI-generated texts demonstrated appropriate lexical coverage and syntactic complexity aligned with ACTFL proficiency descriptors. This strengthens the interpretation that learners were exposed to level-appropriate input throughout the intervention. At the same time, the analysis covered a sample rather than the full corpus of generated texts, so future work should incorporate larger-scale evaluations of AI-generated input to ensure quality and consistency across implementations.

The addition of independent re-rating supports the reliability of the oral proficiency outcomes. Both Cohen's κ and ICC indicated substantial to high consistency, reducing concerns that results were an artifact of single-rater bias. While not all recordings were double-rated, these values provide reassurance that the observed proficiency gains are robust. Future research should extend inter-rater reliability checks to the full dataset to further strengthen external validity.

5.1 Efficacy of written comprehensible input

The observed proficiency gains support Krashen's Input Hypothesis (Krashen, 1982), which posits that language acquisition occurs through exposure to meaningful and understandable input, even in the absence of direct instruction or interaction. The significant upward movement in ACTFL proficiency levels, particularly the transitions from Intermediate Mid to Intermediate High, and from Advanced Mid to Advanced High, suggests that reading alone, when aligned with learners' levels, can drive measurable growth in oral language skills. These results are consistent with Krashen's claim that input-based learning can foster speaking proficiency, though the findings remain preliminary and require cautious interpretation.

The unusually large effect size ($r = 0.87$) observed in this study requires careful interpretation. While the findings are consistent with input-based accounts of language acquisition and demonstrate clear upward progression across proficiency levels, the magnitude of the effect is considerably higher than typically reported in applied linguistics interventions. Several factors may contribute to this outcome, including scale-related ceiling effects, the influence of a single rater across both assessments, and the relatively small sample size, which can inflate effect size estimates. Although sensitivity analyses confirmed that the results remained statistically significant under conservative assumptions, these methodological characteristics necessitate a cautious interpretation of the quantitative outcomes. The sensitivity analyses further support the plausibility of the unusually large effect sizes. Bootstrapped confidence intervals and conservative recoding of tied ratings both indicated that the results remained statistically significant and robust. Nevertheless, the magnitude of the observed effect should be interpreted cautiously, as it may partly reflect sample homogeneity, attrition patterns, or rater consistency. Replication with larger and more diverse samples, multiple raters, and comparison groups is needed to determine whether similar effect sizes can be reproduced in other contexts.

In addition, the attrition analysis revealed that all eight dropouts came from the Novice High group, with no attrition observed at higher levels. This differential pattern raises the possibility that results may be upwardly biased toward learners who already possessed stronger baseline proficiency. While the direction of the gains remains clear, future research should ensure more balanced retention across groups to strengthen generalizability.

A notable limitation of this study is the unequal attrition across proficiency levels. All eight dropouts occurred among Novice High participants, while Intermediate Mid and Advanced Mid learners completed the intervention without attrition. This imbalance raises the possibility of attrition bias, as the results may disproportionately reflect learners who already possessed stronger baseline proficiency or higher motivation. If those who struggled most with reading input discontinued participation, the reported gains may overestimate the effectiveness of the intervention, particularly for lower-proficiency learners.

While the direction of proficiency gains remains consistent across groups, the magnitude of improvement should be interpreted cautiously for the Novice High level. Future studies should implement strategies to minimize attrition among beginning learners, such as scaffolding tasks more gradually, offering supplemental support, or integrating multimodal input to reduce cognitive demands. Attrition analyses should also be systematically planned and reported, to ensure that observed outcomes accurately represent the full range of learners.

The findings also reaffirm the role of reading as a powerful form of comprehensible input. As Krashen (2004) argued, extensive reading allows learners to absorb vocabulary and syntax in context, internalize language patterns, and reduce affective barriers to language development. This aligns with participant reflections that emphasized greater comfort, autonomy, and increased willingness to speak. These results not only support the plausibility of Krashen's hypothesis but also suggest that AI can deliver such input with sufficient precision to yield speaking gains.

The qualitative findings provide important context for these quantitative outcomes. Learners reported increased autonomy, vocabulary growth, and fluency development, which align closely with the observed improvements in OPI scores. At the same time, some participants expressed frustration with the complexity of certain AI-generated texts or a lack of topical relevance, underscoring the need for careful quality control in AI-mediated input. These perceptions not only corroborate the statistical evidence but also demonstrate how learner motivation and engagement influence the effectiveness of input. Together, the quantitative and qualitative strands reinforce the conclusion that AI-generated input can support oral development, while also pointing to areas where implementation must be carefully managed.

5.2 The role of level alignment and scaffolding

The effectiveness of the intervention can also be attributed to its alignment with learners' Zones of Proximal Development (ZPD) (Vygotsky, 1978). AI-generated texts, produced using prompts tailored to varied ACTFL proficiency levels, ensured that input was appropriately challenging but not overwhelming. This scaffolding allowed learners to engage with texts slightly above their current proficiency, promoting upward movement across proficiency bands.

Reflections from lower-level participants frequently cited improved comprehension and reduced need for assistance, illustrating the success of appropriately scaffolded input. These findings are consistent with research on leveled reading and graded input, which emphasizes the importance of targeting materials within a learner's cognitive and linguistic reach.

5.3 Vocabulary accessibility and cognitive load

Prompt design in this study was informed by Nation's (2001) vocabulary acquisition framework and Sweller's (1988) Cognitive Load Theory, both of which emphasize the value of managing lexical and syntactic complexity to support comprehension. Learner reflections suggested that most participants were able to complete the readings independently, suggesting that extraneous cognitive load was effectively minimized. This alignment between input difficulty and learner capacity appears to have promoted sustained engagement and fluency development. In particular, Intermediate and Advanced participants reported that the manageable input load encouraged consistent reading habits, thereby maximizing language exposure over time.

5.4 Learner engagement, trust, and autonomy

While the intervention was successful in fostering measurable oral proficiency gains, qualitative findings revealed nuanced perceptions about the learning experience. Many participants expressed confidence and motivation, suggesting that consistent success with comprehensible texts increased their engagement and willingness to speak. Others appreciated the efficiency and flexibility of the weekly reading task.

However, some participants expressed skepticism about the accuracy or reliability of AI-generated content, particularly in the absence of human verification. These concerns echo findings from other AI-assisted learning research (Bender et al., 2021), which note that learners often question the credibility of AI-generated explanations. Although factual errors were not directly assessed in this study, participants' caution highlights the need for clearer instructional framing when using AI as a source of language input.

Additionally, several advanced participants reported concerns about depth and over-reliance, stating that the readings felt repetitive or lacked academic rigor. Some acknowledged becoming dependent on the AI-generated texts, reducing their engagement with other English sources. These observations suggest that while AI-generated input can scaffold proficiency, it must be periodically varied, extended, or supplemented to maintain long-term cognitive and linguistic development.

5.5 Lack of a control group

A central limitation of the study is the absence of a control or comparison group receiving alternative forms of input, such as instructor-curated or textbook-based readings. Without a control condition, it is not possible to establish causal claims regarding the impact of AI-generated texts on oral proficiency. The observed gains, while statistically significant and supported by learner reflections, must therefore be interpreted as preliminary correlational evidence rather than definitive proof of causality.

Future research should employ randomized or matched-group designs to directly compare AI-generated input with traditional reading materials or mixed-modality interventions. Such comparative studies would help determine whether the observed proficiency gains are uniquely attributable to AI-mediated input or reflect more general benefits of sustained reading exposure.

5.6 Pedagogical implications

These findings offer several implications for language educators and curriculum designers. First, the study demonstrates that AI can be leveraged to deliver proficiency-aligned, individualized reading input at scale. For educators supporting learners across diverse class sizes and proficiency levels, this approach presents a viable strategy to personalize instruction without increasing workload.

Second, while AI can serve as an effective input generator, its use should be accompanied by training in critical reading, fact-checking strategies, and opportunities for self-reflection. Integrating learner agency into AI-supported reading tasks may increase trust and deepen learning outcomes. Educators should also be aware of potential over-reliance and consider rotating genres, text types, or complexity levels to challenge advanced learners.

Finally, the findings provide empirical validation of input-based acquisition in a digital context, illustrating that carefully scaffolded reading alone can promote measurable oral proficiency gains, even in the absence of interaction.

6 Limitations and future directions

While this study provides strong evidence for the effectiveness of AI-generated, level-aligned written input in promoting oral proficiency development, several limitations must be acknowledged.

6.1 Limitations

This study has several limitations that should be acknowledged. First, the Oral Proficiency Interviews (OPIs) were administered and rated by a single ACTFL-trained researcher. While this approach ensured procedural consistency and minimized

variability in scoring, it introduces the possibility of rater bias. A certified ACTFL rater was not available during the study, which precluded independent double ratings and the calculation of formal inter-rater reliability indices (e.g., Cohen's κ or intraclass correlation coefficients). To partially mitigate this limitation, the rater adhered strictly to ACTFL rubrics, piloted the scoring procedure before the study, and re-listened to recordings in borderline cases to maintain internal consistency. Nevertheless, the absence of certified independent ratings requires cautious interpretation of the proficiency gains. Future research should incorporate external ACTFL-certified raters, randomly re-score a subsample of recordings (e.g., 20–30%), and report inter-rater consistency to enhance the robustness and generalizability of the findings.

Second, although the prompts used to generate AI-based texts were carefully engineered and piloted for level-appropriateness, the actual texts produced by ChatGPT were not reviewed or monitored by the researcher. This decision was made to preserve ecological validity and simulate realistic learner-AI interaction without instructor mediation. However, it introduces a degree of variability in lexical density, discourse structure, and topical relevance across participants. Future studies should consider incorporating a stratified content analysis or periodic sampling of generated texts to assess alignment with proficiency targets and ensure greater control over input quality.

Third, the study did not include a control or comparison group receiving alternative forms of input, such as instructor-curated or textbook-based reading materials. While the single-group design was intentionally chosen to isolate the effect of AI-generated input in accordance with Krashen's Input Hypothesis, the absence of a comparative condition limits the ability to determine whether AI-generated input is more effective than traditional approaches. Future research should explore comparative designs that evaluate AI-generated input alongside conventional materials to better assess its relative efficacy, particularly in terms of fluency, vocabulary acquisition, and learner engagement.

Fourth, the intervention focused exclusively on individual reading, with no opportunities for interaction, discussion, or teacher-led instruction. While this design choice was made to isolate the effects of written input and align with Krashen's theory, it limits the study's applicability to more communicative or integrated classroom environments where multimodal input and social interaction are common.

Fifth, although learners were asked to maintain weekly journals throughout the intervention, these journals were collected and analyzed only once at the end of the six-month period. This submission schedule limits insight into how learner perceptions evolved over time and may introduce recall bias. Future studies should consider collecting reflections at regular intervals (e.g., biweekly or monthly) to enable longitudinal analysis. Structured prompts or follow-up interviews could further enhance data richness. Additionally, 14 journals were excluded from analysis due to quality issues, which may have reduced the breadth of qualitative insights.

Finally, the study was conducted at a single public university in Jordan with a relatively homogeneous sample of adult undergraduate students enrolled in an Applied English program. These contextual characteristics limit the generalizability of the findings to other learner populations, age groups, and instructional settings.

6.2 Future directions

Future research should explore how AI-generated input can be optimized for sustained cognitive engagement over time. This may include varying the genres, themes, or levels of abstraction presented in texts, especially for higher proficiency learners who seek deeper conceptual content.

Research comparing AI-generated input with instructor-selected texts could provide insight into which method better supports fluency, complexity, and engagement. Similarly, integrating AI-generated reading with peer interaction, oral discussion, or targeted speaking practice may reveal how different instructional modes can complement written input and accelerate speaking development.

Longitudinal studies that track oral proficiency development beyond 6 months would offer valuable insight into the durability of input-driven gains. In addition, learner autonomy, attitudes toward AI, and metacognitive reading strategies should be investigated more closely through interviews or classroom observations.

Finally, the role of prompt design deserves further attention. As large language models continue to evolve, the ability to scaffold, personalize, and diversify reading input through well-constructed prompts may become a critical area of pedagogical innovation in second language instruction.

7 Conclusion

This study explored the potential of AI-generated graded reading materials as input for improving adult EFL learners' oral proficiency. Quantitative results suggested significant upward movement in proficiency, providing preliminary empirical support for aspects of Krashen's Input Hypothesis. These findings indicate that extended exposure to AI-mediated input, when aligned with learners' levels, may promote measurable growth in oral language skills.

The qualitative findings provided complementary insights. Learners reported increased autonomy, vocabulary growth, and fluency development, which supported the quantitative trends. At the same time, participants expressed concerns about occasional complexity in AI-generated texts, limited topical relevance, and over-reliance on the system. These reflections underscore the importance of careful quality control, instructional framing, and sustained learner engagement in maximizing the effectiveness of AI-mediated input.

Several constraints temper the strength of these conclusions. The unusually large effect sizes, reliance on a single non-certified rater, and the absence of a control group limit the strength of causal claims. Attrition concentrated among Novice High learners raises concerns about bias, and unchecked variability in the linguistic properties of AI-generated texts may have influenced outcomes in ways that could not be systematically monitored. Furthermore, the single-institution setting and relatively homogeneous student population restrict the generalizability of the findings.

Despite these limitations, the study contributes both theoretical and practical insights. Theoretically, it offers preliminary support for extending Krashen's Input Hypothesis into an AI-mediated context, suggesting that algorithmically generated input can function as comprehensible input for speaking development. Practically, it highlights the promise of AI as a scalable tool for fostering autonomy and fluency in higher education classrooms. Future research should adopt more rigorous designs incorporating certified raters, inter-rater reliability

checks, control groups, and longitudinal analyses. Comparative studies of AI-generated and instructor-curated materials, together with systematic evaluation of input quality, will be especially important for establishing best practices in integrating AI into language teaching.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by University of Jordan Ethics Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HA: Writing – review & editing, Writing – original draft. MA: Writing – review & editing, Writing – original draft. MA-D: Writing – original draft, Writing – review & editing. RA: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1614680/full#supplementary-material

References

Alnemrat, A., Aldamen, H. A. K., Almashour, M., Al-Deaibes, M., and AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: a quantitative study. *Front. Educ.* 10:1614673. doi: 10.3389/feduc.2025.1614673

American Council on the Teaching of Foreign Languages. (2024). ACTFL proficiency guidelines 2024. Available online at: https://www.actfl.org/uploads/files/general/ResourcesPublications/ACTFL_Proficiency_Guidelines_2024.pdf (Accessed Ferbruary 16, 2025)

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–623

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Dewey, J. (1933). How we think: A restatement of the relation of reflective thinking to the educative process. Boston: Heath.

Du, J., and Daniel, B. K. (2024). Transforming language education: a systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Comput. Educ.* 6, 1–12. doi: 10.1016/j.caeai.2024.100230

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

Han, Z. (2024). Chatgpt in and for second language acquisition: a call for systematic research. *Stud. Second. Lang. Acquis.* 46, 301–306. doi: 10.1017/s0272263124000111

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274

Krashen, S. D. (1982). Principles and practice in second language acquisition: Pergamon Press.

Krashen, S. (2004). The power of reading: Insights from the research. Libraries Unlimited. Available online at: https://www.sdkrashen.com/content/books/the_power_of_reading.pdf (Accessed January 12, 2025).

Lee, S. S., and Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: a systematic review. *Online Learn.* 28, 82–106. doi: 10.24059/olj.v28i3.4593

Long, M. H. (1996). "The role of the linguistic environment in second language acquisition" in Handbook of second language acquisition. eds. W. C. Ritchie and T. K. Bhatia (San Diego: Academic Press), 413–468.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learn. Environ.* 11:9. doi: 10.1186/s40561-024-00295-9

Mollick, E., and Mollick, L. (2023a). Using AI to implement effective teaching strategies in classrooms: Five strategies, and fifty prompts. Pennsylvania: Wharton Interactive.

Mollick, E., and Mollick, L. (2023b). Assigning AI: Seven approaches for students, with prompts

Mollick, E. R., and Mollick, L. (2024). Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. Pennsylvania.

Mzwri, K., and Turcsányi-Szabo, M. (2025). The impact of prompt engineering and a generative AI-driven tool on autonomous learning: a case study. *Educ. Sci.* 15:199. doi: 10.3390/educsci15020199

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge, UK: Cambridge University Press.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action. London: Basic Books.

Seddik, M. (2025). The impact of AI-powered language learning tools on second language acquisition: a mixed-methods study. *Int. J. Linguist. Lit. Transl.* 8, 269–278. doi: 10.32996/ijllt.2025.8.3.30

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in Second Language Acquisition* 15, 165–179.

Van Brummelen, J. (2019). Conversational agents to democratize artificial intelligence. *IEEE Symposium Visual Lang. Human Centric Comput.* 2019, 239–240. doi: 10.1109/VLHCC.2019.8818805

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

Wang, H., and Dang, A. (2024). Enhancing L2 writing with generative AI: a systematic review of pedagogical integration and outcomes. London.

Wang, Y., Zhang, T., Yao, L., and Seedhouse, P. (2025). A scoping review of empirical studies on generative artificial intelligence in language education. *Innov. Lang. Learn. Teach.*, 6, 1–28. doi: 10.1080/17501229.2025.2509759

Yoon, S.-Y., Miszoglad, E., and Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. California.