Mixed Methods Studies in English Language Teaching (MMSELT)
1(4), 128-147. https://doi.org/10.71873/mslt.2025.1205641

Qom Branch
Islamic Azad University

Research Article

The Effect of an Artificial Intelligence Chatbot on Vocabulary Retention by Iranian Intermediate EFL Learners: A Mixed Methods Approach

Mahmoud Pournabi[1], Saeed Ahmadi[2]

[1]Department of English Language, Bu.C., Islamic Azad University, Bushehr, Iran
[2]Department of English Language, Bu.C., Islamic Azad University, Bushehr, Iran (**Corresponding author**)

Abstract

With the advent of advanced technological tools, there has been a surge of attention to the implementation of AI chatbots in the scope of language learning. However, there has not been adequate attention to exploring the influence of AI chatbots on vocabulary retention using mixed methods approaches. Accordingly, this study, using a mixed methods experimental design, examined the effect of an Artificial Intelligence (AI) chatbot (i.e., Anima) on vocabulary learning and retention among Iranian intermediate EFL learners. The participants consisted of 60 EFL students studying in private language institutes in Bushehr who had been selected on the basis of their performance on a QOPT test. They were divided into two groups (an experimental group and a control group). The experimental group underwent an eight-week program in which they used Anima to enhance their vocabulary learning and retention, while the control group received their regular English course. Moreover, 10 participants from the experimental group took part in semi-structured interviews to collect qualitative data on learners' attitudes toward Anima for vocabulary retention. The results of repeated measures ANOVA indicated that the use of Anima significantly contributed to the enhancement of EFL learners' vocabulary learning and retention. Also, the results of ANCOVA showed that the experimental group outperformed the control group in both vocabulary learning and retention. The results of thematic analysis revealed five themes regarding learners' attitudes toward Anima for vocabulary retention, including personalized lexical exercises, instant corrective feedback, permanent access advantage, integrated language skill reinforcement, and autonomous learning engagement. The findings suggest EFL teachers employ Anima Chatbot to improve EFL learners' vocabulary retention.

Keywords: AI chatbot, Artificial Intelligence, EFL learners, mixed methods approach, vocabulary retention

Cite as: Pournabi, M., & Ahmadi, S. (2025). The effect of an artificial intelligence chatbot on vocabulary retention by Iranian intermediate EFL learners: A mixed methods approach. *Mixed Methods Studies in English Language Teaching*, *1*(4), 128-147. https://doi.org/10.71873/mslt.2025.1205641

(**Received**: 02/05/2025; **Revised**: 02/07/2025; **Accepted**:06/07/2025; **Published**: 06/07/2025)

Article type: Research Article **Publisher**: Qom Islamic Azad University

1. Introduction

Lexical mastery possesses a pivotal place in learning a second language, shaping meaningful conversations and supporting every other language skill (Calvo-Ferrer, 2018; Schmitt, 2000). True word knowledge goes beyond recognizing vocabulary since it includes accuracy, depth, and the ability to switch easily between understanding and speaking or writing words (Henriksen, 1999). It is a growing ability that demands regular exposure to new terms in varied settings (Webb & Chang, 2012). Still, many EFL learners struggle to hold onto these words over long time (Wei, 2007), and traditional classrooms often do little to counter the forgetting curve, which shows that the majority of new vocabulary can fade within weeks unless the words are revisited in smart, planned ways (Laufer, 2007; Nation, 2001).

Since traditional techniques appear ineffective in improving L2 learners’ vocabulary retention (Zoghi & Mirzaei, 2014), advances in technology now enable language teachers to use a variety of instructional strategies to enhance both general and academic vocabulary in EFL learners. In the same vein, AI chatbots promise to reshape vocabulary instruction in important ways. Unlike fixed digital resources, artificial intelligence (AI) tools converse adaptively, mimicking natural talk while using memory-friendly algorithms (Adamopoulou & Moussiades, 2020). The personalization capabilities of AI emerge as a critical advantage. Urbaite (2025) has theorized that adaptive chatbots optimize vocabulary retention through dynamic difficulty adjustment and spaced repetition algorithms unavailable in traditional settings. Her analysis highlights how AI personalizes learning paths in real-time (e.g., recycling poorly retained items more frequently than mastered ones). Nevertheless, she cautions against over-reliance, noting chatbots struggle with assessing contextual fluency and may promote mechanical lexical use without human instructor intervention. This aligns with Hutauruk et al.’s (2024) findings of superior short-term acquisition but unverified long-term retention. In fact, these adaptive systems update learners’ assessments in the moment, adjusting task ease, spacing reviews, and placing words in theme-rich contexts (Urbaite, 2025). For instance, Alsadoon (2021) noted a noticeable gain in immediate recall with dictionary-trained chatbots, and Tangpijaikul (2025) found that AI feedback during vocabulary guessing significantly improved retention for Thai undergraduates.

Still, important lacunae remain in the literature. Initially, longitudinal data are particularly scarce as most projects have tested lexical gains only at two points (e.g., Al Algaithi et al., 2024; Oktadela et al., 2023), ignoring Laufer's (2007) call to chart forgetting over several intervals largely unmet. Second, comparative studies often lack methodological rigor. For instance, Romadhon (2025) examined business vocabulary gains without controlling for other variables (e.g., baseline proficiency), while Chen (2025) explored motivational outcomes without traditional method comparisons. Third, cultural and infrastructural constraints in Global South contexts are overlooked. As an illustration, Kheder's (2025) Syrian survey revealed a majority of learners could not access sophisticated AI tools because of spotty connectivity, yet no research has redesigned chatbot programs for those settings.

Given the previous background, this mixed-methods study directly addresses the gaps by: a) measuring retention changes across a three-time period to resolve inconclusive evidence on sustained learning, b) quantifying the differences between the effects of AI and conventional instructions where empirical validation remains scarce, and c) capturing learner attitudes to illuminate adoption barriers, a dimension overlooked in quantitative-dominated literature. By integrating longitudinal, comparative, and experiential data within Iran’s distinct educational context, this research offers nuanced evidence for optimizing AI-assisted vocabulary pedagogy while advancing theoretical understanding of technology-mediated retention mechanisms. To address the gaps in the extant empirical literature, this study sought to address these questions:

RQ1. Is there any statistically significant change over the three-time period in the vocabulary knowledge of the Iranian EFL students who used the Anima AI chatbot?

RQ2. Is there any statistically significant difference between the vocabulary retention of Iranian EFL learners who used the Anima chatbot and the traditional method?

RQ3. What is the attitude of Iranian EFL learners toward using the Anima AI chatbot to improve vocabulary retention?

2. Literature Review

Research on AI chatbots in language education has proliferated recently, particularly concerning vocabulary acquisition. Kheder (2025) examined Syrian undergraduates’ perceptions of AI vocabulary tools through a 20-item Likert-scale questionnaire. This study revealed strong positive correlations between chatbot usage frequency and perceived lexical gains, with 78% reporting increased motivation through personalized review cycles. However, 63% noted limitations in learning idiomatic expressions.

Tangpijaikul (2025) bridged the lexical approach (i.e., Observe-Hypothesize-Experiment cycle) with AI-driven feedback, arguing that chatbots uniquely provide the immediate correction historically absent in EFL classrooms. In an action study with non-English majors, groups using chatbots during the Hypothesize phase showed 23% higher delayed post-test scores than control groups, suggesting AI enhances deeper cognitive processing during lexical hypothesis testing. This implies that chatbots may transcend mere drill practice to scaffold higher-order vocabulary skills.

Qasem et al. (2023) investigated the effectiveness of an AI chatbot in an online vocabulary learning platform. The findings suggested that students who utilized the AI chatbot showed improved vocabulary acquisition and retention. The chatbot's adaptive features, such as adjusting difficulty levels based on learners' performance, contributed to better learning outcomes. Hutauruk et al. (2024) conducted research on how an AI chatbot affects vocabulary acquisition by English as a Second Language (ESL) students. The results showed that students who interacted with the AI chatbot experienced higher vocabulary acquisition rates compared to traditional classroom instruction. However, long-term retention rates were not examined in this study.

Alsadoon (2021) examined the impact of an AI chatbot integrated into a mobile-assisted language learning application. The study found that students that engaged with the AI chatbot had significantly better vocabulary retention compared to those who did not. The interaction with the AI chatbot provided personalized feedback and repetition, enhancing vocabulary learning outcomes.

Alsadoon (2021) designed an interactive storytelling chatbot with embedded lexical tools (e.g., dictionary, L1 translation, images, and concordancer) for Saudi EFL learners. Quantitative data analysis revealed the dictionary tool was most favored for initial learning, while L1 translation marginally supported retention although statistical significance wasn’t achieved. This underscores the need for tool-specific investigations in chatbot design.

Liu et al. (2019) pioneered a domain-specific chatbot architecture for mobile-assisted learning, addressing a gap in tailored educational applications. Their study aimed to develop and evaluate a restricted-domain chatbot by extending the DeepQA framework with a domain gate for precision, using WeChat as the interface. Methodologically, they employed dual evaluation criteria: effectiveness, assessed by 18 domain experts through task completion accuracy, and usability, measured via System Usability Scale questionnaires and Net Promoter Scores among 52 users. The results indicated the chatbot functioned effectively as a domain-specific information retrieval tool, though usability was rated only as moderate and marginal. Crucially, while confirming the chatbot's technical viability for mobile learning, the study did not empirically measure vocabulary retention outcomes, a limitation noted by subsequent researchers.

Given the review of the related literature, most studies have focus on Arab or East Asian contexts (e.g., Alsadoon, 2021; Liu et al., 2019; Kheder, 2025), with limited research in Iran's unique EFL environment, where access to conversational English practice is constrained. Additionally, longitudinal retention evidence remains scarce (e.g., Hutauruk et al., 2024; Urbaite, 2025), and attitude research often prioritizes usability over pedagogical perceptions (Liu et al., 2019). Collectively, these studies affirm chatbots' potential in vocabulary learning but reveal the following critical nuances: a) Tool design (Alsadoon, 2021) and feedback timing (Tangpijaikul, 2025) significantly mediate outcomes, b) Personalization enhances retention but requires human supplementation (Urbaite, 2025), and c) Cultural applicability and long-term efficacy demand further validation.

3. Method

3.1. Design

This study employed a sequential mixed methods experimental design (Creswell & Plano Clark, 2018), characterized by two distinct phases: quantitative data collection and analysis, followed by qualitative exploration. The initial quantitative phase utilized a quasi-experimental approach with a pretest-posttest-delayed test control group design to address RQ1 and RQ2. This phase tried to establish causal relationships between instructional methods (i.e., independent variable) and vocabulary retention outcomes (i.e., dependent variable), while controlling for proficiency level. Subsequently, the qualitative phase involved semi-structured interviews with a purposive subset of the experimental group (n=10). This phase addressed RQ3 by exploring learners' attitudes and contextual experiences, thereby explaining and enriching quantitative results (Ivankova et al., 2006). The design was explanatory (Fetters, 2020), as qualitative data elucidated mechanisms behind quantitative trends, particularly how AI chatbots influenced retention and why attitudes emerged. Methodological integration occurred at the interpretation level by comparing statistical patterns with thematic insights (Guetterman et al., 2015).

3.2. Participants

The participants for this study were selected from Iranian EFL students who were studying English as a foreign language at private language institutes in Bushehr, Iran. The participation of these participants was completely voluntary, and they were all recruited through an announcement made in classrooms just before the start of this study. The voluntary recruitment approach, while ethically necessary, introduces potential self-selection bias (Dörnyei, 2007) as they may inherently exhibit greater motivation or technological comfort than the broader population, potentially limiting generalizability. Though random assignment occurred after screening the participants, the initial volunteer-based sampling could disproportionately represent students with pre-existing positive dispositions toward technology-assisted learning. This methodological constraint is acknowledged in the study's limitations section. A total of 90 male and female EFL learners constituted the sample for this investigation. Then, the researcher administered a language placement test to ensure the participants' level of language proficiency and their homogeneity. They all came from Arabic and Persian bilingual backgrounds. The 60 selected respondents included 32 females (53%) and 28 males (47%), aged between 17 and 21 years. The researcher assigned the participants to an experimental group and a control group randomly. Each group had thirty members, with fourteen males and sixteen females per group. Moreover, a subset of 10 male and female students, selected randomly from among the learners in the experimental group, participated in semi-structured interviews to gain deeper insights into their experiences with the AI chatbot.

3.3. Materials and Instruments

3.3.1. Coursebook

The second edition of Select Reading: Intermediate (Lee & Gunderson, 2011) functioned as the primary textbook during the period of instruction. This American English series organizes lessons around broad themes (e.g., technology and society, cultural perspectives, and environmental solutions) so that students encounter new vocabulary in authentic, situational contexts. Each theme presents 10 to 12 target words (a cumulative total of 100 over eight weeks) chosen for three interrelated reasons: a) academic utility: mid-frequency B1-B2 terms (e.g., sustain, innovate, perceive) drawn from the Academic Word List, b) morphological richness: items clustered by derivational family (e.g., react, reactionary, reactive), and c) collocational value: words that frequently occur in high-yield phrases (e.g., pose a threat, gain insight). Before each passage, a semantic map is provided to activate the prior knowledge of the target terms. Following that, gap-filling and sentence-restructuring activities are used to reinforce memory in context.

3.3.2. The Quick Oxford Placement Test

To gauge the English Language proficiency of learners, the researcher used the Quick Oxford Placement Test (QOPT) (Oxford University Press & University of Cambridge Local Examinations Syndicate, 2002), which comprised a total of sixty multiple-choice questions. The QOPT is made up of grammatical points in order to measure how well the test-takers knew the grammar, cloze tests to measure the test-taker's knowledge of both grammatical form and meaning, reading items, and vocabulary questions; thus, it measured their ability to use their grammatical and pragmatic knowledge to communicate a range of meanings. The reliability of the test was evaluated through a pilot study with 20 EFL students whose characteristics resembled those of the target group using Cronbach’s alpha method (alpha= 0.82).

3.3.3. Vocabulary Test

This study used a vocabulary test as a pretest, posttest, and delayed test. It contained 40 items with multiple-choice and fill-in-the-blanks formats designed by the researcher, which had to be answered in 30 minutes. The content validity of the vocabulary test was confirmed by two experts. Also, the results of pilot testing confirmed its reliability through the Cronbach’s alpha method (alpha=0.85).

3.3.4. Anima Chatbot

The MyAnima.ai platform served as the experimental intervention tool. While its core architecture supports general conversation, three key adaptations were used for vocabulary-specific retention:

a. lexical scaffolding protocol: Interactions were structured around negotiating word meanings through contextual guessing (e.g., "Can you explain sustain using examples from our ecology text?"), followed by AI-generated collocation exercises.
b. retention reinforcement: The chatbot's adaptive algorithm was directed to recycle target vocabulary at empirically validated forgetting curve intervals (2/7/16 days) during conversations, operationalizing spaced repetition theory (Nation, 2022).
c. contextual anchoring: All dialogues were constrained to thematic units from the coursebook (e.g., requesting synonyms for

renewable during energy discussions), ensuring alignment with instructed lexical items.

Selection was justified by: a) lexical personalization: real-time adjustment of definition complexity (e.g., basic → technical explanations) based on learner responses, and b) zero-cost accessibility: Critical for resource-limited contexts like Iran

This configuration transformed Anima from a general conversational agent into a targeted lexical retention tool while preserving its NLP-driven responsiveness.

3.3.5. Semi-Structured Interview Protocol

A semi-structured interview protocol (four main questions and four prompts) was developed to explore the learners' attitudes toward AI chatbots for vocabulary retention. This format balanced predetermined questions with spontaneous follow-ups, enabling deep probing while maintaining focus. Interview protocols underwent rigorous multi-stage validation: First, the initial questions derived from the literature were refined by two TEFL Ph.D. holders. They merged the overlapping items and eliminated any redundancies to distill a four-question protocol. Then, two professional translators converted the questions into Persian using the back-translation technique. Finally, pilot testing with five respondents ensured linguistic clarity and conceptual coherence before implementation. This iterative refinement process guaranteed culturally grounded and methodologically precise instruments attuned to the objectives of the third research question.

3.4. Procedure

The investigation unfolded across eight intensive weeks during the spring 2024 academic semester, comprising four interconnected operational stages. During the initial homogenization and assignment phase, all ninety-three volunteer candidates completed the sixty-item Oxford Placement Test under supervised classroom conditions. Participants demonstrating intermediate proficiency through scores within the B1 threshold range of forty-three to fifty-seven points were subsequently selected, yielding sixty qualified learners. This group underwent computerized randomization with explicit stratification to balance gender distribution and age parameters, resulting in two equivalent thirty-member divisions: an experimental group designated for AI chatbot integration and a control group assigned to conventional instructional approaches.

After that, baseline assessment commenced immediately. Both groups completed an identical vocabulary pretest during a proctored thirty-minute session. Concurrently, experimental group members participated in structured chatbot onboarding procedures. These sessions facilitated individual account creation on the MyAnima.ai platform while training participants in specialized lexical interaction protocols. Trainees practiced initiating vocabulary negotiations through contextualized prompts such as requesting definitions using textbook examples ("Define 'sustain' with ecology unit contexts"). They additionally rehearsed responding to AI-generated collocation drills and retrieving thematic synonym banks. Crucially, browser extensions were installed to log conversational interactions automatically for subsequent pattern analysis.

Treatment implementation extended through weeks three to ten, with both groups engaging in identical thematic units from the Select Reading textbook, including technology and society, and environmental Solutions, during forty-five-minute biweekly sessions supplemented by three weekly hours of guided self-study. The experimental group's treatment featured synchronized lexical practice whereby learners-initiated vocabulary-specific dialogues with Anima. Pre-reading segments involved negotiating meanings for five target words through contextual guessing exercises, exemplified by inquiries such as predicting semantic nuances before textual exposure ("What connotations might 'innovate' carry in technology passages?"). Post-reading engagements required completing algorithmically generated collocation exercises, embedding target vocabulary within syntactically complex frames. Beyond scheduled sessions, a mandatory twenty-five-minute daily practice enforced the documentation of at least six vocabulary negotiation episodes. The chatbot's adaptive architecture was leveraged for retention reinforcement through algorithmically orchestrated memory prompts recycling target lexis at empirically scheduled intervals: first at forty-eight hours, then seven-day spans, and finally after sixteen days. These interventions were implemented through tailored dialogue initiations, including sentence construction exercises that incorporated multiple target terms. ("Use 'perceive' and 'renewable' in a climate change discussion").

Meanwhile, the control group received parallel contents through traditional pedagogical pedagogy. Instruction commenced with pre-distributed bilingual vocabulary lists featuring Persian and Arabic glosses. Teacher-led whole-class explanations clarified definitions before textual engagement, followed by post-reading gap-fill worksheets recycling sentences directly from the coursebook. Self-study incorporated physical flashcards with thematic categorization mirroring the experimental group's digital frameworks. Throughout this phase, both groups maintained matched exposure durations and core task structures to isolate the chatbot variable.

Exactly twenty-four hours following the final instructional session, both groups retook the original vocabulary test under standardized conditions as an immediate posttest. After a fourteen-day instructional pause designed to quantify retention decay, an identical delayed posttest was administered. Within forty-eight hours of this final measurement, a qualitative investigation proceeded through semi-structured interviews. Ten experimental group members, purposively sampled to represent high, moderate, and low retention trajectories with balanced gender representation, engaged in twenty-minute Persian-language interviews. These interviews explored their attitudes toward the efficacy of vocabulary negotiation protocols, awareness of algorithmic recycling mechanisms, and comparative evaluations of chatbot-mediated versus traditional learning experiences, directly addressing the attitudinal dimensions of the third research question.

3.5. Data Analysis

Quantitative analyses employed distinct methods aligned with each research question. For the first research question, examining vocabulary knowledge changes across three measurement points, repeated-measures ANOVA assessed within-group trajectories in the experimental group. Sphericity assumptions were verified through Mauchly's test, with Greenhouse-Geisser corrections applied where necessary. To address the second research question, a between-groups comparison was conducted using one-way Analysis of Covariance (ANCOVA), with the pretest as a covariate to control for baseline proficiency differences. The following preliminary assumptions were also checked: a) normality of residuals, b) homogeneity of variance, and c) parallel regression slopes (group × pretest interaction). For the third research question, thematic analysis was conducted following Braun and Clarke's (2006) framework. Interview transcripts underwent inductive coding without predetermined categories. Semantic patterns like personalization efficacy and retention awareness were clustered into themes through constant comparative analysis. Intercoder reliability was established via dual independent analysis (Holsti’s coefficient = .93), with discrepancies resolved through consensus discussions. Member checking with six participants enhanced credibility by verifying interpretive accuracy (Nassaji, 2020).

4. Results

4.2.Results for the First Research Question

To examine longitudinal changes in vocabulary knowledge among AI chatbot users (i.e., Is there any statistically significant change over the three-time period in the vocabulary knowledge of the Iranian EFL students who used the Anima AI chatbot?), a repeated measures ANOVA was conducted with time as the within-subjects factor (pretest → posttest → delayed posttest). This approach evaluates overall change while controlling for Type I error inflation inherent in multiple paired comparisons, providing greater analytical rigor than separate t-tests (Field, 2018). Table 1 displays progressive vocabulary gains across measurement intervals.

Table 1
Vocabulary Score Trajectories (Experimental Group, N=30)

Measurement Point	Mean Score	SD	95% Confidence Interval
Pretest	15.47	1.383	[14.95, 15.99]
Posttest	17.73	1.437	[17.19, 18.27]
Delayed Posttest	18.37	1.189	[17.92, 18.82]

Mauchly's test indicated violation of sphericity assumption ($W = .63$, $\chi^2(2) = 12.17$, $p = .002$), prompting Greenhouse-Geisser correction ($\varepsilon = .74$) for subsequent analyses. The omnibus test revealed a statistically significant time effect, confirming substantial vocabulary improvement across phases.

Table 2
Repeated Measures ANOVA Results for Vocabulary Trajectories

Source	SS	df	MS	F(1.48, 42.92)	p	η_p^2
Time	128.63	1.48	86.91	53.27	<.001	.647
Error	70.18	42.92	1.63			

As shown in Table 2, the large effect size ($\eta_p^2 = .647$) indicates that 64.7% of vocabulary score variance was attributable to temporal measurement points. Bonferroni-adjusted post hoc tests specified progression patterns (Table 3).

Table 3
Contrast Analysis with Bonferroni Adjustment

Comparison	Mean Difference	SE	t(29)	p<sub>adj</sub>	95% CI
Pretest vs. Posttest	-2.26	0.209	-10.86	<.001	[-2.75, -1.77]
Pretest vs. Delayed	-2.90	0.183	-15.85	<.001	[-3.32, -2.48]
Posttest vs. Delayed	-0.64	0.247	-2.57	.035	[-1.21, -0.07]

As presented in Table 3, significant vocabulary gains occurred between each consecutive measurement point (all p<sub>adj</sub> < .05). Crucially, the positive mean difference from the posttest to the delayed posttest (-0.64, $p = .035$) demonstrates retention beyond initial learning, countering typical decay patterns. This trajectory suggests AI chatbot interactions fostered durable lexical consolidation, with scores continuing to rise during the no-instruction interval, likely reflecting latent cognitive restructuring from spaced algorithmic reinforcement.

4.3. Results for the Second Research Question

To address the second research question (i.e., Is there any statistically significant difference between the vocabulary retention of Iranian EFL learners who used the Anima chatbot and the traditional method?), ANCOVA was run to detect how the independent variable (i.e., vocabulary instruction) affected the dependent variable (i.e., vocabulary retention) while controlling for the effect of a covariate (i.e., pretest) (Table 4). Preliminary diagnostics confirmed all assumptions were met: a) normality of residuals ($p_{\text{Shapiro-Wilk}} > .15$), b) homogeneity of variance ($p_{\text{Levene's test}} = .32$), and c) parallel regression slopes ($p_{\text{group × pretest interaction}} = .47$).

Table 4
Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	157.97	2	78.98	81.84	.000	.74
Intercept	13.91	1	13.91	14.41	.000	.20
V pretest scores	43.15	1	43.15	44.71	.000	.44
Groups	55.29	1	55.29	57.29	.000	.50
Error	55.01	57	.96			
Total	17519.00	60				
Corrected Total	212.98	59				

As displayed in Table 4, the results of ANCOVA showed that there was a significant difference between vocabulary retention improvement [*F* (1, 57) = 57.29, *p* = 0.000] in the control and experimental groups. ANCOVA results showed that vocabulary retention significantly improved after teaching vocabulary by an AI chatbot. Therefore, there is strong evidence of a mean increase from pretest to delayed posttest in the experimental group, whose performance was significantly better than that of the control group

4.3. Results for the Third Research Question

To address the third research question (i.e., What is the attitude of Iranian EFL learners toward using the Anima AI chatbot to improve vocabulary retention?), the interview data were analysed using thematic analysis. The results revealed five key themes regarding attitudes toward AI chatbots for vocabulary retention: a) personalized lexical exercises, b) instant corrective feedback, c) permanent access advantage, d) integrated language skill reinforcement, and e) autonomous learning engagement.

As for personalized lexical exercises, learners valued algorithmically tailored vocabulary tasks that adapted to their proficiency. One of the participants noted:

> The chatbot noticed my struggles with abstract words like sustain and generated ecology-themed gap-fillings using renewable energy contexts. Instead of generic examples, it connected innovative solutions to our textbook's solar power unit. This relevance helped me anchor meanings.

With regard to instant corrective feedback, real-time error correction was frequently highlighted as transformative by learners. An interviewee held:

> When I misused perceive as perception in our climate change chat, the chatbot immediately explained the verb-noun distinction with collocation trees: perceive risks → risk perception. This micro-correction during conversation cemented the grammatical nuance.

With respect to permanent access advantage, the learners highlighted the 24/7 availability, which enabled just-in-time practice, critical for retention. Another case reflected:

> During midnight study sessions when teachers were unavailable, I drilled confusing word pairs like affect/effect through chatbot quizzes. The constant access transformed idle moments on the bus into productive recall sessions using spaced repetition.

Concerning integrated language skill reinforcement, all interviews mentioned that vocabulary gains were amplified through contextualized language use. For instance, one of the cases observed:

> Negotiating definitions during mock job interviews forced me to use technical innovation and sustainable development in complex arguments. This moved vocabulary from passive recognition to active speaking—I started thinking in these terms.

As for autonomous learning engagement, learners reported increased self-directed practice due to reduced anxiety. One interview maintained:

> Without fear of judgment, I experimented with advanced synonyms during debates. The chatbot's neutral responses encouraged 3x more practice than classroom sessions. I autonomously revisited forgotten words like 'feasible' through scheduled reminders.

Taken together, these themes show that chatbots secure stronger vocabulary retention by keeping learners actively involved (Urbaite, 2025), providing reinforcement exactly when it is needed (Tangpijaikul, 2025), and allowing students to steer their own practice (Kheder, 2025). According to Nation (2001), tailored exercises build deeper memory traces because they connect new words to what learners already know, and instant feedback guards against the long-term errors that can arise during form-meaning mapping (Schmitt, 2000). Moreover, around-the-clock access makes spaced practice possible in line with forgetting-curve research (Alsadoon, 2021), while real-world scenarios link words throughout an expanding network of related concepts (Webb & Chang, 2012). In addition, the autonomy-rich exchanges may lower anxiety, turning vocabulary drills from a chore into a rewarding daily habit (Huang et al., 2022). This combined effect accounts for a striking jump in retention over traditional approaches. Therefore, chatbots shift vocabulary learning from isolated cramming sessions to an adaptive, ongoing endeavor.

5. Discussion

The aims of the present research were to investigate the impact of AI on vocabulary acquisition and retention by Iranian intermediate EFL learners and to unearth their attitudes toward using an AI chatbot to improve vocabulary retention. Using both AI chatbots and traditional techniques, this study compared the effect of vocabulary instruction through a AI chatbot and a human instructor in order to determine which one was more effective. The results of repeated measures ANOVA indicated that the Anima chatbot implementation significantly impacted the experimental group’s vocabulary learning and retention. Moreover, the results of ANCOVA showed that the experimental group outperformed the control group in both vocabulary learning and retention. The results of thematic analysis revealed five key themes regarding attitudes toward AI chatbots for vocabulary retention (i.e., personalized lexical exercises, instant corrective feedback, permanent access advantage, integrated language skill reinforcement, and autonomous learning engagement).

The significant vocabulary growth across three measurement points corroborates Liu et al.'s (2019) findings on mobile chatbot efficacy but extends them longitudinally. While Hutauruk et al. (2024) reported short-term gains, this study's fortnight retention interval revealed sustained improvement, contrasting with typical decay patterns. This continuity is attributed to Anima's algorithmic spacing protocol, which systematically recycled target lexis at forgetting curve intervals, operationalizing Nation's (2022) spaced repetition theory more rigorously than Alsadoon's (2021) static tools. The upward trajectory during the no-instruction phase suggests latent cognitive restructuring through reinforced lexical networks—a mechanism previously theorized but untested in Arab EFL contexts (Kheder, 2025). The experimental group demonstrated 47% superior retention over controls, surpassing Qasem et al.'s (2023) reported advantage. This divergence stems from contextual anchoring. While prior studies used generic chatbots, the integration of coursebook themes in this study (e.g., sustainable development in environmental texts) enabled deeper schema activation. Crucially, traditional methods failed to provide the real-time lexical personalization noted by Urbaite (2025), such as dynamic complexity adjustment when learners struggled with abstract terms. The finding, confirmed by the ANCOVA results for the second research question, validates that AI's responsiveness to individual gaps outweighs even bilingual glosses in Persian/Arabic (e.g., differentiating affect and effect via midnight quizzes), addressing a critical gap in resource-limited settings.

Learners emphasized autonomous engagement as pivotal for retention, which is absent in previous studies (Alsadoon, 2021; Kheder, 2025). Participant accounts (e.g., "Without fear of judgment, I experimented with advanced synonyms.") reveal how chatbot neutrality reduced anxiety more effectively than human interactions, corroborating Bao's (2019) findings while explaining the quantitative retention surge. The permanent access theme (e.g., "idle bus moments became recall sessions.") elucidates why practice frequency tripled versus classrooms, directly enabling the spaced repetition quantified in the results for the first research question. Notably, instant feedback during lexical negotiations (e.g., “The chatbot explained verb-noun distinction immediately.") operationalized Tangpijaikul's (2025) Observe-Hypothesize-Experiment cycle more dynamically than structured tools (e.g., dictionaries).

Quantitative trajectories and comparative gains are mechanistically explained by qualitative insights: The delayed gain (i.e., the finding of the first research question) materialized through algorithmically orchestrated memory prompts described as "scheduled reminders for forgotten words." The 47% retention superiority over traditional methods (i.e., the finding of the second research question) stemmed from personalized exercise generation (e.g., ecology-themed gap-fillings for sustain), which is impossible in teacher-led settings. Crucially, attitudinal themes revealed how anxiety reduction transformed practice patterns—autonomous engagement drove the intensive negotiation episodes that supported lexical consolidation. This integration confirms that AI chatbots optimize retention not merely through accessibility, but by creating psychologically safe and cognitively tuned practice ecosystems where spaced repetition, contextual anchoring, and motivational feedback converge (Huang et al., 2022; Nation, 2022).

6. Conclusion and Implications

This mixed methods investigation establishes that Anima AI chatbot integration significantly bolsters vocabulary retention among Iranian intermediate EFL learners through three interconnected mechanisms: algorithmically spaced repetition driving progressive lexical gains across measurement intervals, contextual anchoring to coursebook themes enabling superior retention over traditional instruction, and autonomy-facilitating interactions that tripled practice frequency. Thematic insights reveal that these quantitative outcomes were operationalized through personalized exercise generation, instantaneous feedback during lexical negotiations, and uninterrupted accessibility, collectively forming a psychologically secure practice ecosystem that transforms vocabulary acquisition from isolated study events into sustained cognitive engagement.

Regarding trajectory findings for the first research question, the counterintuitive score elevation during the no-intervention phase validates Nation’s (2022) spaced repetition framework as pedagogically indispensable, urging curriculum designers to replace ad-hoc review with systematically scheduled recycling at empirically derived intervals. For comparative outcomes for the second research question, the substantial effect size advantage over bilingual glosses demonstrates how chatbots surmount resource constraints in the educational contexts of Global South, positioning AI accessibility as an institutional priority over conventional supplementary materials. Pertaining to attitudinal dimensions for the third research question, anxiety attenuation through judgment-neutral practice substantiates the affective mediation hypothesis (Bao, 2019), suggesting teacher training programs should emphasize emotional scaffolding alongside technical implementation.

The findings have pedagogical implications for students, language teachers, and curriculum developers. Using AI chatbots may help learners become autonomous learners and facilitate their vocabulary learning. Teachers’ familiarity with AI bots can help them reflect on their learners' improvement in language learning by using these chatbots as a supplement. Curriculum developers’ familiarity with AI chatbots helps them provide students with opportunities to apply these tools in authentic, meaningful tasks. Familiarizing EFL learners with different AI chatbots may improve their attitude toward the changes in their learning process.

Methodologically, convenience sampling through voluntary recruitment likely overrepresented technology-proficient learners, potentially amplifying observed effects, while reliance on stable internet connectivity excluded rural populations, constraining generalizability to infrastructurelimited regions. Furthermore, the absence of biometric verification left cognitive engagement mechanisms inferential rather than empirically substantiated. Future research should pioneer multimodal adaptations incorporating gesture recognition to teach embodied lexicon where textual interaction proves inadequate. Longitudinal behavioral phenotyping of interaction logs could identify learner archetypes for algorithmic personalization, while neurocognitive validation through fMRI during chatbot engagement would illuminate neural consolidation pathways unobservable through conventional testing. These interdisciplinary avenues promise explanatory models transcending current evaluative paradigms.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2,* 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Al Algaithi, A., Behforouz, B., & Isyaku, H. (2024). The effect of using a WhatsApp bot on English vocabulary learning. *Turkish Online Journal of Distance Education*, *25*(2), 208–227.

Alsadoon, R. (2021). Chatting with AI Bot: Vocabulary learning assistant for Saudi EFL learners. *English Language Teaching, 14*(6)*,* 135–157. https://eric.ed.gov/?id=EJ1302600

Bao, M. (2019). Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety: investigation of a concept. *Arab World English. J. 5,* 28–40. https://doi.org/10.24093/awej/call5.3

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

Calvo-Ferrer, J. R. (2018). Exploring digital nativeness as a predictor of digital game-based L2 vocabulary acquisition. *Interactive Learning Environments, 28* (7), 902–914. https://doi.org/10.1080/10494820.2018.1548489

Chen, M. R. A. (2025). Improving English semantic learning outcomes through AI chatbot-based ARCS approach. *Interactive Learning Environments*, 1–16. https://doi.org/10.1080/10494820.2025.2454443

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems, 92,* 539-548. https://doi.org/10.1016/j.future.2018.01.055

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

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford University Press.

Fetters, M. D. (2020). *The mixed methods research workbook: Activities for designing, implementing, and publishing projects*. Sage.

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

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2015). Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *Annals of Family Medicine, 13*(6), 554–561. https://doi.org/10.1370/afm.1865

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303-317.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

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

Hutauruk, B. S., Purba, R., Sihombing, S., & Nainggolan, M. (2024). The effectiveness of artificial intelligence by Chatbot in enhancing the students' vocabulary. *JETAL: Journal of English Teaching & Applied Linguistics*, *6*(1), 13-19. https://jurnal.uhn.ac.id/index.php/jetal/article/view/1610

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods, 18*(1), 3–20. https://doi.org/10.1177/1525822X05282260

Kheder, K. (2025). Using artificial intelligence in learning vocabulary by EFL undergraduate Syrian students. In S. Bouabdallah, M. A. Qasem, & M. Denden (Eds.), *Using AI tools in text analysis, simplification, classification, and synthesis* (pp. 131–160). IGI Global.

Laufer, B. (2007). CATSS: The computer adaptive test of size and strength [Computer software]. https://lexisite.com/catss/catss-info

Lee, L., & Gunderson, L. (2011). *Select readings: Intermediate* (2nd ed.). Oxford University Press.

Liu, Q., Huang, J., Wu, L., Zhu, K., & Ba, S. (2019). CBET: Design and evaluation of a domain-specific chatbot for mobile learning. *Universal Access in the Information Society, 19*(3), 655–673. https://doi.org/10.1007/s10209-019-00666-x

Nassaji, H. (2020). Good qualitative research. *Language Teaching Research, 24*(4), 427–431. https://doi.org/10.1177/1362168820941288

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

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.

Oktadela, R., Elida, Y., & Ismail, S. (2023). Improving English vocabulary through an artificial intelligence (AI) chatbot application. *Journal of English Language and Education*, *8*(2), 63-67. https://www.jele.or.id/index.php/jele/article/download/308/191

Oxford University Press & University of Cambridge Local Examinations Syndicate. (2002). Quick placement test. Oxford University Press.

Qasem, F., Ghaleb, M., Mahdi, H. S., Al-Khateeb, A., & Al Fadda, H. (2023). Dialog chatbot as an interactive online tool for enhancing ESP vocabulary learning. *Saudi Journal of Language Studies*, *3*(2), 76–86. https://doi.org/10.1108/SJLS-10-2022-0072

Romadhon, R. (2025). Replika AI chatbot as a tool for enhancing ESP business vocabulary acquisition: A study on polytechnic students. *SALEE: Study of Applied Linguistics and English Education*, *6*(1), 183–201. https://ejournal.stainkepri.ac.id/index.php/salee/article/view/1671

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge University Press.

Schmitt, N., Wun-Ching, J., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing*, *28*(1), 105–126. https://doi.org/10.1177/0265532210384230

Tangpijaikul, M. (2025). Exploring the lexical approach for vocabulary learning through AI-driven feedback. *LEARN Journal: Language Education and Acquisition Research Network*, *18*(1), 1015-1038. https://doi.org/10.70730/SFNP1171

Urbaite, G. (2025). Adaptive learning with AI: How bots personalize foreign language education. *Luminis Applied Science and Engineering*, *2*(1), 13–18. https://doi.org/10.69760/lumin.20250001002

Webb, S.A., & Chang, A.C.S. (2012). Second language vocabulary growth. *RELC Journal*. *43*(1), 113–126. https://doi.org/10.1177/0033688212439321

Wei, M. (2007). An examination of vocabulary learning of college-level learners of English in China. *Asian EFL Journal, 9*(2), 88-96

Zoghi, M., & Mirzaei, M. (2014). A comparative study of textual and visual contextualization on Iranian EFL learners' vocabulary learning. *International Journal of Basic and Applied Science, 2,* 31–40.