

Effects of Productive vs. Receptive Vocabulary Tasks on Low-Proficiency EFL Learners' Long-Term Retention and Transfer

Lin Zhengyun^{1*}, Chandra Mohan Vasudeva Panicker²

^{1&2} School of Education, Lincoln University College, Malaysia

¹ School of Foreign Language JiangXi Institute of Technology, China

*Corresponding author: 1257261983@qq.com

Abstract

With the rapid development of artificial intelligence in foreign language education, large language models have provided new pathways for vocabulary teaching. Traditional vocabulary learning mainly relies on receptive tasks (e.g., flashcard-based apps like Baicizhan), which often results in low-proficiency learners recognizing words but being unable to use them, or forgetting quickly after short-term memory. This study adopted a quasi-experimental design with 70 low-proficiency non-English majors (CET-4: 300–424, Gaokao: 60–89 out of 150) from Jiangxi Institute of Technology (JXIT). The participants were randomly divided into two groups, with 35 students in the experimental group and 35 in the control group. The experimental group adopted an AI-supported productive vocabulary-learning model, while the control group engaged in conventional receptive vocabulary learning through the Baicizhan platform. The intervention lasted eight weeks, with 30 minutes daily, 15 theme-based CET-4 words per day, totaling 600 words. Data were collected via pretest, immediate posttest and delayed posttest. Independent samples t-test, paired samples t-test, and effect size analysis were employed. The results revealed that the experimental group significantly outperformed the control group in both immediate and delayed posttests. The findings verify that compared with conventional receptive vocabulary learning, AI-supported productive vocabulary learning effectively improves long-term vocabulary retention and lexical transfer competence among low-proficiency EFL learners. This study provides a replicable AI-based teaching paradigm for vocabulary instruction targeting low-proficiency learners in private universities and offers new empirical evidence for second language vocabulary acquisition theories.

Keywords: *AI-Assisted Learning, Baicizhan, Low-Proficiency EFL Learners, Long-Term Retention, Lexical Transfer, Productive Vocabulary Learning*

Introduction

Vocabulary plays a crucial role in second language proficiency, serving as an indispensable foundation for learners' listening, speaking, reading, and writing abilities (Nation, 2022). However, vocabulary acquisition remains a persistent challenge for low-proficiency EFL learners in Chinese private universities. Although they often devote substantial time to memorizing words, they frequently struggle with rapid forgetting and difficulty using them appropriately in authentic contexts. A noticeable imbalance especially frequently exists between receptive vocabulary knowledge and productive vocabulary ability, with learners able to recognize words in reading tasks but unable to

actively use them in speaking or writing (Laufer & Goldstein, 2004). Such problems are especially common among learners with relatively weak English foundations, who tends to rely heavily on rote memorization and repetitive recognition-based exercises while learning vocabulary.

At present, commercial vocabulary applications, such as Baicizhan and Shan bay, have become popular tools for vocabulary learning among Chinese college students. These applications mainly emphasize receptive learning activities, including form-meaning matching, picture recognition, and multiple-choice tasks. Such practices facilitate short-term memorization, but learners are provided with insufficient chances for contextualized output, active retrieval, and deeper semantic processing. According to Fergus I. M. Craik and Robert S. Lockhart's Depth of Processing Theory, memory retention is closely associated with the depth of cognitive processing involved during learning, and deeper semantic processing is more likely to produce durable long-term memory than shallow recognition-based learning (Craik & Lockhart, 1972). Similarly, studies on productive vocabulary learning suggest that active retrieval and contextualized language production are important for vocabulary internalization and transfer (Laufer & Hulstijn, 2001).

The emergence of large language models has brought fresh possibilities to optimize traditional vocabulary teaching and learning. Unlike rigid, exercise-oriented commercial vocabulary apps, modern AI learning tools can deliver contextualized interactive practice, real-time error feedback, and individualized learning support. Such adaptive features are particularly valuable for low-proficiency learners, who consistently need targeted guidance to improve their language skills. Swain's (1985) Comprehensible Output Hypothesis well supports the value of this AI-enabled learning mode. She argued that language production is far more than a simple display of acquired knowledge. Instead, it pushes learners to identify their linguistic deficiencies, test out language hypotheses, and engage in deeper cognitive processing of language rules. Within AI-supported learning environments, learners can practice and revise contextualized vocabulary output repeatedly, a flexible learning experience that traditional receptive training is unable to provide.

Despite growing academic attention on AI-assisted vocabulary learning, existing empirical research still has notable shortcomings. Existing intervention studies are mostly short-term in design and often involve limited vocabulary learning scopes or mixed-ability learner groups, which weakens the reliability and generalizability of their findings. Moreover, most prior research prioritizes learners' receptive vocabulary recognition while overlooking productive vocabulary transfer and long-term knowledge retention, which are the ultimate goals of vocabulary learning. Research focusing specifically on low-proficiency undergraduate learners in Chinese private universities also remains insufficient, leaving an underexplored research niche in local EFL contexts. Beyond that, most current studies treat AI tools merely as supplementary learning aids, rather than designing them as systematic pedagogical support to foster learners' productive vocabulary competence.

Against such a research backdrop, this study sets out to explore whether AI-based productive vocabulary tasks can better help low-proficiency EFL learners retain and transfer vocabulary knowledge in the long term, compared with conventional app-based receptive learning. Drawing on three classic theoretical frameworks—the Depth of Processing Theory (Craik & Lockhart, 1972), the Comprehensible Output Hypothesis

(Swain, 1985), and Vygotsky's Scaffolding Theory (1978)—this study compares the learning effects of AI-mediated productive vocabulary training and traditional receptive vocabulary learning. Its findings aim to provide empirical evidence for targeted vocabulary teaching design for low-proficiency learners and offer practical references for the effective integration of AI technology into EFL teaching in private university settings.

Review of Related Literature

In recent years, second language vocabulary acquisition research has made great progress in both theoretical development and technological applications. Scholars worldwide have extensively explored the limitations of traditional learning modes, the pedagogical value of output tasks, and the educational potential of artificial intelligence. Internationally, vocabulary knowledge is widely recognized as the foundation of language competence (Nation, 2022). A great number of empirical studies indicates that receptive learning via flashcard-based applications such as Quizlet and Mondly mainly enhances word recognition rather than deep lexical mastery. This approach often results in superficial learning characterized by “passive recognition without productive use” and rapid forgetting, with limited development of collocational competence and transfer ability (Jensen & Cadierno, 2022; Li & Zhang, 2022; Yang & Wang, 2024). Theoretically, Swain's Comprehensible Output Hypothesis (1985), Craik and Lockhart's Depth of Processing Theory (1972), and Task-Based Learning (TBLT) collectively demonstrate that productive tasks such as sentence creation, dialogue practice, and story retelling trigger deeper semantic processing, strengthen form-meaning-use connections, and promote long-term retention and lexical transfer (Swain, 1985; Yang et al., 2024; Hu, 2025). Recent research on working memory further suggests that learners with low working memory capacity benefit more from brief, low-cognitive-load productive tasks (Jin & Luo, 2026). With the rapid advancement of generative AI, studies have confirmed its ability to provide contextualized input, personalized scaffolding, real-time feedback, and low-anxiety interaction, effectively enhancing productive competence and lexical transfer, especially for low-proficiency learners (Alshahrani, 2024; Chen & Liu, 2025; Liu & Ma, 2025). A comprehensive meta-analysis by Şimşek et al. (2026), including 1,823 participants across 18 studies, revealed a large overall effect size ($g = 1.309$) for AI-assisted vocabulary learning, highlighting its effectiveness for low-proficiency learners. However, existing international research still suffers from short intervention periods, overreliance on receptive tasks, and insufficient empirical evidence focusing on Chinese low-proficiency learners.

In China, research reveals that Chinese university students, particularly those at private institutions with low English proficiency (CET-4 scores between 300 and 425), rely heavily on commercial apps such as Baicizhan for rote memorization, leading to persistent problems including weak retention, poor collocation, and a substantial gap between receptive and productive vocabulary (Zhou et al., 2016; Yu et al., 2022; Wu & Zhao, 2026). Yang et al. (2024) further noted that domestic studies on productive tasks often focus only on lexical depth rather than productive fluency. Recent domestic empirical research has shown that AI-assisted productive tasks significantly improve long-term retention and transfer compared to traditional app-based learning (Hao & Qi, 2025; Wang et al., 2024). A meta-analysis by Wang et al. (2025) also confirmed that 8-week interventions yield the most stable effects. Nevertheless, domestic scholarship still exhibits noticeable research gaps. Most local interventions adopt short-term designs, underutilise productive learning tasks, pay insufficient attention to private university low-

proficiency populations, and rarely conduct long-term tracking of vocabulary retention and transfer performance.

To sum up, despite widespread scholarly recognition of the value of productive tasks, deep cognitive processing and AI-enabled instruction in vocabulary learning, current research still lacks systematic eight-week intervention studies. There is also a lack of direct empirical comparisons between AI-assisted productive vocabulary training and conventional app-based receptive learning, as well as targeted evidence for low-proficiency non-English majors in private university settings. Hence, this study adopts an eight-week quasi-experimental design with low-proficiency students from a private university in Jiangxi. It compares how two learning approaches—AI-supported productive vocabulary practice and Baicizhan-based receptive learning—shape learners' immediate vocabulary acquisition, delayed retention and lexical transfer abilities. The research outcomes intend to provide practical empirical support for AI-driven vocabulary pedagogy in private university EFL classrooms and enrich context-specific data for second language vocabulary acquisition theories.

Research Questions

RQ1: Does AI-supported productive vocabulary learning yield better long-term vocabulary retention than Baicizhan-based receptive learning among low-proficiency EFL learners?

RQ2: Does AI-supported productive vocabulary learning facilitate better productive vocabulary transfer than Baicizhan-based receptive learning among low-proficiency EFL learners?

RQ3: What are low-proficiency EFL learners' learning experiences and perceptions regarding AI-supported productive vocabulary learning?

Methodology

This study used a quasi-experimental design with pre-test, immediate post-test, and delayed post-test to examine how two vocabulary-learning approaches affect low-proficiency EFL learners' vocabulary development. Quantitative data were collected to examine the differences in learners' short-term vocabulary mastery, long-term vocabulary retention, and productive vocabulary transfer ability between AI-supported productive vocabulary learning and traditional app-based receptive vocabulary learning. Overall, the quantitative statistical analysis ensured objective and credible research findings for verifying the effectiveness of different vocabulary learning interventions.

Research Design

A quasi-experimental design with pre-test, immediate post-test, and delayed post-test measurements was employed in this study. Participants were divided into an experimental group and a control group with all teaching and learning conditions strictly controlled to be identical, except for the core independent variable of vocabulary learning mode. The experimental group received an eight-week AI-supported productive vocabulary learning intervention, while the control group adopted conventional receptive vocabulary learning via the Baicizhan application.

Before the formal intervention, a vocabulary pre-test was conducted to verify the baseline homogeneity of vocabulary proficiency between the two groups. Upon the completion of the eight-week intervention, an immediate post-test was implemented to assess learners' short-term vocabulary mastery and productive transfer ability. Four weeks after the intervention, a delayed post-test was administered to measure learners' long-term vocabulary retention effects. Strict variable control measures were implemented throughout the experiment to eliminate the interference of irrelevant confounding factors and guarantee the validity and reliability of the experimental results.

Participants

The participants of this study were 70 second-year non-English major undergraduates from Jiangxi Institution of Technology, a private university in southeast China. All participants had received unified college English teaching for one academic year before the experiment, with consistent daily English learning arrangements and no participation in additional off-campus English training courses, which ensured the basic homogeneity of their learning background and English foundation. To screen typical low-proficiency EFL learners suitable for this intervention, authoritative and objective English proficiency indicators were adopted as selection criteria. Specifically, participants' CET-4 scores ranged from 300 to 425, and their National College Entrance Examination English scores (full score = 150) were between 60 and 89. This score range effectively screened learners with weak but improvable English foundations, avoiding extreme cases of zero-basis learners or proficient learners that might interfere with experimental validity. Furthermore, a pre-experiment vocabulary test confirmed that participants' initial mastery rate of target vocabulary ranged from 20% to 40%, indicating that most target words were new and learnable for participants, which successfully prevented floor and ceiling effects in experimental data.

All participants were randomly assigned to the experimental group and the control group, with 35 students in each group. Independent samples t-test results verified no significant difference in baseline vocabulary level and overall English proficiency between the two groups ($p > 0.05$). None of the participants had long-term experience in using AI learning tools or professional vocabulary learning applications, which effectively controlled potential confounding variables. All participants volunteered to participate in this study and signed informed consent forms.

Target Vocabulary Materials

The target vocabulary materials were screened from 739 official CET-4 high-frequency words embedded in the Baicizhan platform developed by Chengdu Superlove Technology Co., Ltd. This word bank was compiled in accordance with the latest national CET-4 syllabus and a corpus of 18 authentic CET-4 examination papers from the latest three years. Words in this list frequently appear in the listening, reading, translation, and writing sections of CET-4 examinations, serving as a authoritative and practical vocabulary resource for Chinese undergraduates preparing for CET-4 tests.

To align with participants' CEFR A2-B1 proficiency level and the eight-week intervention schedule, further screening was conducted on the original 739 words. Basic vocabulary that most learners had already mastered and obscure sophisticated words with low examination frequency were eliminated. Ultimately, 600 qualified core CET-4 high-frequency words were confirmed as the formal experimental materials. These words were categorized into eight thematic units, with one exclusive learning theme arranged per

week throughout the intervention. Learners learned 15 target words daily and 75 words weekly. The thematic grouping strengthened contextual connections between vocabulary items and facilitated learners' practical lexical application. The selected vocabulary maintained a reasonable part-of-speech proportion, including 40% nouns, 35% verbs, and 25% adjectives. This distribution conforms to the practical lexical application demands of daily English communication and CET-4 writing tasks. The pre-experiment vocabulary test results confirmed participants' low initial mastery of target vocabulary, ensuring the effectiveness and credibility of the subsequent intervention data.

Research Instruments

This study employed three parallel vocabulary test papers as the core research instrument, implemented as the pre-test, immediate post-test, and delayed post-test respectively. The three parallel tests maintained equivalent structures, full scores, difficulty levels, and item distributions. To minimize test practice effects and memorization bias, each test version adopted different sentence contexts, contextual prompts, and item arrangements while retaining unified assessment dimensions and evaluation standards.

Each test paper consisted of three sections to comprehensively measure three core dimensions of vocabulary learning outcomes: receptive vocabulary knowledge, productive vocabulary retrieval ability, and vocabulary transfer competence. The receptive vocabulary section adopted multiple-choice items to assess learners' ability to recognize and comprehend target words in contextualized sentences. The productive vocabulary section included sentence completion and guided sentence production tasks, which aimed to measure learners' productive word retrieval and contextualized lexical application ability. In addition, each test paper contained a contextualized vocabulary transfer writing task. Participants were required to complete an 80–100-word short paragraph in a new communicative context by applying designated target vocabulary items. This task was designed to examine learners' ability to transfer vocabulary knowledge from passive recognition and controlled production to spontaneous, meaningful contextualized use, rather than evaluating their overall English writing proficiency. The writing topics across the three tests were equivalent in difficulty and structural complexity but differed in situational prompts to avoid repeated memorization and ensure authentic test performance.

Instead of testing all 600 target words directly, the tests adopted a stratified proportional sampling approach. Test items were proportionally and evenly selected from different thematic units and instructional stages of the eight-week intervention, ensuring the representativeness and comprehensiveness of the assessment. The total score of each test paper was 100 points. To guarantee content validity, all test items were developed based on the CET-4 high-frequency vocabulary corpus consistent with the experimental learning materials and were reviewed by two experienced college English teachers specializing in vocabulary instruction and language assessment. Prior to the formal experiment, a pilot test was conducted among non-participant students with equivalent English proficiency. The results showed that the Cronbach's α coefficient of the test exceeded 0.80, indicating satisfactory internal consistency and reliable test stability.

For the vocabulary transfer writing task, the standardized scoring rubric covered four core dimensions: vocabulary accuracy, contextual appropriateness, grammatical correctness, and lexical diversity. Two experienced college English teachers independently rated all writing responses. Inter-rater reliability was calculated to ensure scoring consistency, and

the average score of the two raters was adopted as the final score for each participant to guarantee scoring objectivity and credibility.

Intervention Procedures

The entire intervention lasted eight weeks, with 30 minutes of fixed vocabulary learning arranged every weekday. All learning activities were completed uniformly under teacher supervision, ensuring consistent learning duration and learning intensity for all participants. Except for the differentiated vocabulary learning modes for the two groups, all other learning conditions, including the learning environment, schedule arrangement, and review requirements, remained completely identical to strictly control irrelevant variables.

Experimental Group: AI-Supported Productive Vocabulary Learning

Participants in the experimental group completed daily standardized productive vocabulary learning tasks guided by the Coze AI intelligent agent. The AI system was preset with a fixed and unified learning procedure to avoid random and invalid interactive content, ensuring standardized intervention quality throughout the experiment. The daily learning process comprised four core links: vocabulary presentation, productive sentence making, thematic contextual dialogue, and daily learning review.

Firstly, the AI systematically presented the phonetic symbol, part of speech, core meaning, common collocations, and typical example sentences of each target word. Secondly, learners were required to create independent original sentences with each target word, and the AI provided real-time error correction and targeted optimization guidance for grammatical errors, collocation mistakes, and contextual inappropriateness. Thirdly, the AI built exclusive thematic scenarios based on the weekly learning theme, guiding learners to complete contextual interactive dialogues using target words. Finally, the AI summarized learners' daily error types and key vocabulary knowledge to support targeted consolidation. The entire learning process focused on active lexical output and deep semantic processing to promote effective vocabulary internalization.

Control Group: Baicizhan-Based Receptive Vocabulary Learning

Participants in the control group learned the same 600 target vocabulary words via the Baicizhan application. The daily learning quantity, learning frequency, and eight-week intervention cycle were completely consistent with those of the experimental group. The learning mode was limited to traditional receptive vocabulary training, including picture flashcard recognition, word meaning matching, and multiple-choice tests. All output-related functions, personalized dialogue interactions, and manual editing functions of the application were manually disabled throughout the intervention.

Learners in the control group only completed passive word recognition and mechanical repetitive review of target vocabulary. Their learning process mainly relied on shallow form-meaning mapping, without any active vocabulary output, contextual application, or targeted error correction feedback, which was distinctly different from the deep productive learning mode of the experimental group.

Data Analysis Methods

This study adopted purely quantitative statistical analysis to process experimental data. All data were analyzed via SPSS 26.0 statistical software. A 2 (Group: AI vs. Baicizhan) $\times$ 3 (Time: Pretest, Immediate Posttest, Delayed Posttest) mixed-design ANOVA was

conducted to examine overall effects. Independent samples t-tests were used for between-group comparisons, and paired samples t-tests for within-group changes. Cohen's d was calculated to assess effect size.

Analysis

Table 1 presents the descriptive statistics of the vocabulary test scores for Group A (AI-assisted) and Group B (Baicizhan-based) across the pre-test, immediate post-test, and delayed post-test.

Table 1: Descriptive Statistics of Vocabulary Test Scores

	Pretest		Immediate posttest		Delayed posttest	
	Mean	SD	Mean	SD	Mean	SD
Group A	55.91	6.79	79.00	10.81	72.37	12.04
Group B	55.20	11.09	67.91	13.94	65.37	16.38

As shown in Table 1, the two groups demonstrated highly comparable vocabulary proficiency before the intervention. The mean pretest scores of the group A (55.91) and group B (55.20) were very close, indicating no substantial initial difference between the two groups. This result suggests that the participants shared a relatively homogeneous vocabulary foundation at the beginning of the experiment, thereby ensuring the validity of subsequent comparisons.

After the 8-week intervention, both groups improved in the immediate posttest; however, the group A achieved substantially greater gains than group B. The mean score of the AI group increased from 55.91 to 79.00, while the Baicizhan group improved from 55.20 to 67.91. The independent samples t-test further confirmed a statistically significant difference between the two groups in the immediate posttest ($t = 3.718$, $p = .001$), with a large effect size ($d = 0.889$). These findings indicate that AI-supported productive vocabulary learning was significantly more effective than receptive flashcard learning in facilitating vocabulary acquisition among low-proficiency EFL learners.

Figure 1 further illustrates the vocabulary score trajectories of the two groups across the three testing stages. As shown in Figure 1, the AI group demonstrated a steeper improvement trend after the intervention and maintained a relatively more stable retention pattern in the delayed post-test compared with the Baicizhan group.

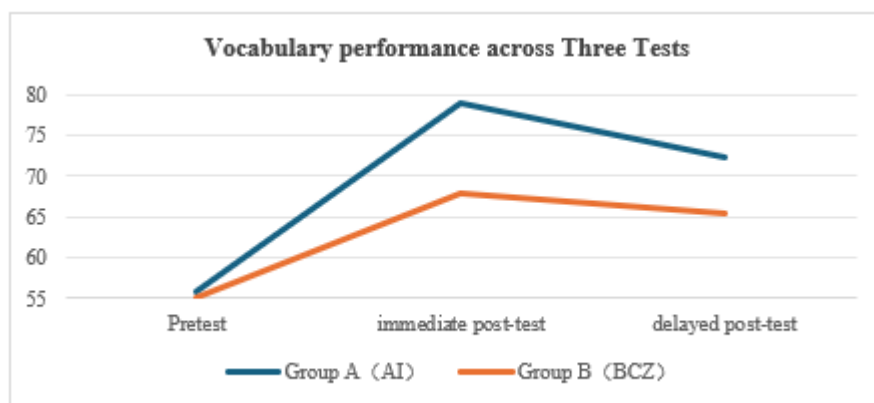


Figure 1. Vocabulary performance across Three Tests

The superior performance of the AI group may be attributed to the nature of the productive learning tasks. Unlike the receptive learning activities provided by Baicizhan, the AI-supported learning model required learners to actively retrieve and use target vocabulary through contextualized sentence production and guided dialogue interaction. Such tasks encouraged learners to integrate lexical meaning, contextual adaptation, and communicative intent simultaneously rather than relying solely on passive recognition and form-meaning matching. This finding strongly supports Depth of Processing Theory (Craik & Lockhart, 1972), which argues that deeper semantic processing contributes to stronger and more durable memory encoding.

The delayed post-test results further revealed important differences in long-term retention between the two learning approaches. Although both groups experienced some degree of forgetting after the intervention, the AI group maintained a considerably higher retention level than the Baicizhan group. More specifically, the AI group retained a delayed post-test mean score of 72.37; whereas the Baicizhan group obtained a lower mean score of 65.37. The statistical analysis also revealed a significant difference between the two groups in the delayed post-test ($p = .046$, $d = 0.487$), suggesting that AI-supported productive vocabulary learning contributed to stronger long-term vocabulary retention.

To further examine long-term retention performance, Figure 2 compares the vocabulary development trajectories of the two groups in terms of immediate learning gain, forgetting loss, and net retention. As illustrated in Figure 2, the AI group achieved substantially larger immediate learning gains (+23.09) than the Baicizhan group (+12.71). Although the AI group also showed a moderate decline in the delayed posttest, its final net retention gain (+16.46) remained considerably higher than that of the Baicizhan group (+10.17). These findings suggest that AI-supported productive vocabulary learning not only promotes stronger vocabulary acquisition but also leads to more durable long-term retention.

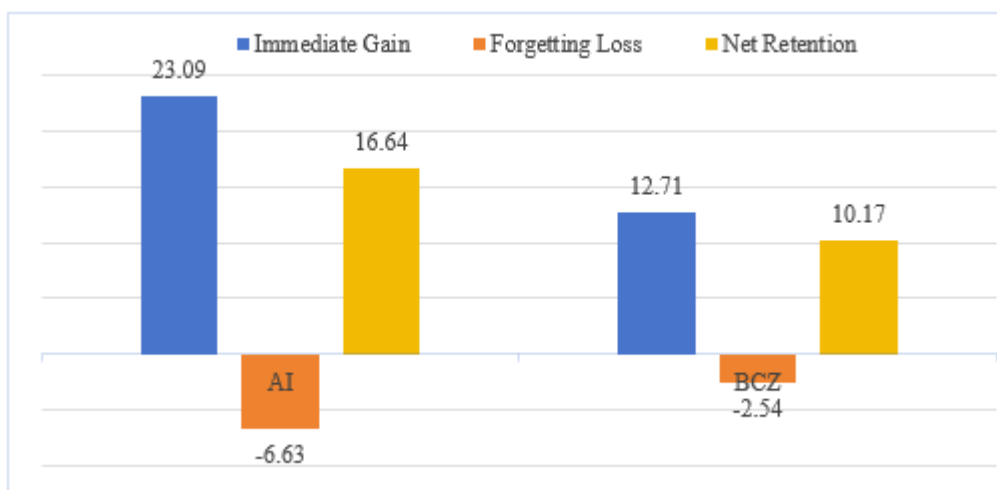


Figure 2. Vocabulary Retention Performance Between Groups

A key explanation lies in the repeated retrieval and contextualized application embedded in productive tasks. During AI dialogue interactions, learners actively retrieved target vocabulary, constructed original sentences, and responded to communicative prompts. This repeated productive retrieval strengthened lexical accessibility and reinforced memory consolidation, consistent with the retrieval practice effect in cognitive psychology: active retrieval enhances long-term retention more effectively than passive review (Karpicke & Roediger, 2008).

Furthermore, the AI system delivered immediate corrective feedback, adaptive prompts, and contextual scaffolding throughout the process. This dynamic support helped learners identify lexical gaps, revise inaccurate expressions, and gradually internalize vocabulary knowledge—aligning with Vygotsky’s (1978) Scaffolding Theory, which highlights guided assistance for effective learning. For low-proficiency learners, who often lack confidence and autonomous strategies, AI-mediated interaction offered a low-anxiety environment for repeated productive practice.

In addition to vocabulary retention, the present study also investigated learners’ vocabulary transfer ability through a writing task. Table 2 presents the descriptive statistics of writing scores across the three tests. At the pretest, the two groups were comparable in writing performance, with Group A scoring 8.54 (SD = 1.40) and Group B scoring 8.34 (SD = 1.39), indicating no significant baseline difference. After the 8-week intervention, Group A improved to 10.17 (SD = 2.26) in the immediate posttest, while Group B reached only 9.14 (SD = 2.05). In the delayed posttest, Group A maintained a stable score of 10.11 (SD = 2.18), whereas Group B declined slightly to 8.89 (SD = 2.32). Learners in the AI group demonstrated greater ability to use target vocabulary accurately and appropriately in new communicative situations. Their writing showed richer lexical variety, more natural collocations, and better contextual adaptation compared with the Baicizhan group.

Table 2: Descriptive Statistics of Writing Scores

	Pretest		Immediate posttest		Delayed posttest	
	Mean	SD	Mean	SD	Mean	SD

Group A	8.54	1.40	10.17	2.26	10.11	2.18
Group B	8.34	1.39	9.14	2.05	8.89	2.32

This finding suggests that productive vocabulary learning may facilitate the transformation of receptive vocabulary knowledge into active productive competence. While receptive flashcard learning may improve vocabulary recognition, it does not necessarily guarantee learners' ability to retrieve and apply vocabulary during authentic communication. In contrast, AI-supported productive tasks continuously pushed learners to produce language, test lexical hypotheses, and refine vocabulary usage through interaction. This result is consistent with Swain's (1985) Comprehensible Output Hypothesis, which highlights the crucial role of language output in second language development.

Overall, the findings consistently demonstrate that AI-supported productive vocabulary learning shows clear advantages over receptive flashcard learning in terms of immediate vocabulary acquisition, long-term retention, and vocabulary transfer ability. The study suggests that vocabulary learning effectiveness depends not merely on repetition frequency, but more importantly on the depth of cognitive processing, productive retrieval, and contextualized lexical engagement during learning. Moreover, the findings highlight the pedagogical potential of AI-assisted vocabulary learning systems as interactive scaffolding environments rather than simple digital memorization tools, particularly for low-proficiency EFL learners in private undergraduate institutions.

Findings and Discussion

Based on the statistical analyses of pretest, immediate posttest, and delayed posttest data, the present study yielded five key findings.

First, AI-supported productive vocabulary learning led to significantly better vocabulary performance than receptive flashcard learning. The AI group (Group A) outperformed the Baicizhan group (Group B) on both the immediate posttest ($p = .001$) and the two-week delayed posttest ($p = .046$), confirming the immediate and sustained advantages of productive practice.

Second, productive learning fostered superior long-term vocabulary retention. While both groups experienced some forgetting over time, the AI group maintained substantially higher delayed posttest scores (72.37 vs. 65.37) and a larger net retention gain (+16.46 vs. +10.17), indicating more durable lexical memory.

Third, AI-mediated practice significantly enhanced vocabulary transfer to writing. In the writing task, the AI group produced more accurate, contextually appropriate target words with richer collocations and better coherence. Their writing scores were consistently higher across all tests, especially in delayed retention ($p = .026$), reflecting stronger productive competence.

Fourth, productive tasks promoted deeper lexical processing and more stable encoding. Unlike passive recognition, AI-driven dialogue, sentence construction, and immediate

feedback required active retrieval, semantic analysis, and contextual adjustment. This aligns with Depth of Processing Theory (Craik & Lockhart, 1972) and retrieval practice effects (Karpicke & Roediger, 2008), strengthening long-term memory traces.

Finally, AI tools show clear pedagogical value for low-proficiency EFL learners in private undergraduate settings. When designed as interactive, output-oriented scaffolding systems rather than simple memorization tools, AI-supported tasks effectively improve vocabulary acquisition, retention, and transfer.

Conclusion

This quasi-experimental study compared the effects of AI-supported productive vocabulary learning and conventional receptive flashcard learning over eight weeks with low-proficiency EFL learners at JXIT as subjects. The experimental data revealed that the AI group achieved far better results than the control group. Their advantages lay in long-term vocabulary retention and lexical transfer, and the differences between the two groups were statistically significant. The findings indicate that AI-enabled contextualized practice encourages deeper cognitive engagement, strengthens lexical retrieval, and enhances long-term retention. By prompting learners to use words in meaningful interactions and receive timely feedback, AI systems help transform passive receptive knowledge into flexible productive competence, as predicted by Swain's (1985) Comprehensible Output Hypothesis.

Pedagogical implications for low-proficiency EFL instruction are clear. First, vocabulary teaching should shift from rote memorization to contextualized productive use. Second, AI tools should be integrated as adaptive scaffolding environments—not just digital flashcards—to support output practice, error correction, and gradual skill development. Third, instruction should prioritize sustainable retention and transfer over short-term recognition gains. For AI application development, the results highlight that interaction-driven, output-focused design better supports lexical growth than receptive-only models. Future tools could embed more authentic communicative scenarios and personalized feedback loops.

Several limitations must be acknowledged. The sample was limited to one institution, restricting generalizability. The eight-week intervention period only captured short-to-medium effects; longer-term outcomes remain unexamined. Future research could include larger, more diverse samples, extended interventions, and mixed-methods data (e.g., interaction logs, learner interviews) to explore underlying learning mechanisms. In summary, AI-supported productive vocabulary learning is a pedagogically effective, learner-centred approach for low-proficiency EFL learners. It offers a practical, sustainable way to improve vocabulary acquisition, retention, and transfer in private undergraduate contexts.

References

- Alshahrani, S. (2024). ChatGPT-assisted vocabulary acquisition for EFL learners. *Forum for Linguistic Studies*, 7(9), 10257. <https://doi.org/10.30564/fls.v7i9.10257>
- Chen, L., & Liu, H. (2025). *Supporting low-proficiency L2 learners with custom GPT-scaffolded pedagogy*. Computer Assisted Language Learning. <https://doi.org/10.1080/0958822.2539152>

- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Hao, C., & Qi, X. F. (2025). Effects of LLM chatbots on college students' self-regulated vocabulary learning. *Overseas English*, (14), 42–44. (in Chinese)
- Hu, J. (2025). Task-based instruction and vocabulary learning. *Advances in Education*, 15(3), 391–399. <https://doi.org/10.12677/ae.2025.153415>
- Jensen, C., & Cadierno, F. (2022). The effects of app-based vocabulary learning on young EFL learners. *Language Learning & Technology*, 26(2), 1–18.
- Jin, X., & Luo, N. (2026). Working memory capacity and lexical retention in AI-assisted vocabulary learning. *System*, 103, 103890. <https://doi.org/10.1016/j.system.2025.103890>
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. <https://doi.org/10.1126/science.1152408>
- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. <https://doi.org/10.1111/j.0023-8333.2004.00260.x>
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>
- Li, Y., & Zhang, L. (2022). Mobile-assisted vocabulary learning: Limitations of receptive memorization. *Frontiers in Psychology*, 13, 993224. <https://doi.org/10.3389/fpsyg.2022.993224>
- Liu, Y., & Ma, X. (2025). AI-generated context and lexical productive competence. *Language Learning*, 75(2), 456–489.
- Nation, I. S. P. (2022). *Learning vocabulary in another language (3rd ed.)*. Cambridge University Press.
- Şimşek, B., Direkci, B., & Koparan, B. (2026). Artificial intelligence-assisted vocabulary teaching: A meta-analysis. *Frontiers in Artificial Intelligence*, 9, 1806700. <https://doi.org/10.3389/frai.2026.1806700>
- Swain, M. (1985). *Communicative competence: Some roles of comprehensible input and comprehensible output in its development*. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

- Wang, S. L., et al. (2025). A meta-analysis of AI chatbots' effects on EFL learning. *Heilongjiang Higher Education Research*, (2), 112–118. (in Chinese)
- Wang, Y. Y., Wu, J., Chen, F., Wang, Z., Li, J., & Wang, L. (2024). Empirical assessment of AI-powered tools for vocabulary acquisition. *IEEE Access*, 12, 131892–131905. <https://doi.org/10.1109/ACCESS.2024.3446657>
- Wu, F. Q., & Zhao, X. (2026). AI-assisted apps and college English vocabulary learning. *Overseas English*, (4), 93–97. (in Chinese)
- Yang, L. R., Song, J. W., & Yang, X. M. (2024). A review of L2 vocabulary acquisition based on continuation tasks. *Journal of Beijing International Studies University*, (3), 30–44. (in Chinese)
- Yang, W., & Wang, F. (2024). Design features and limitations of Chinese vocabulary apps. *Modern Educational Technology*, 34(5), 78–83. (in Chinese)
- Yang, Y., Li, M., & Zhang, H. (2024). *Productive tasks and lexical fluency development*. *System*, 98, 102587.
- Yu, S. L., Dong, Z., & Huang, S. Y. (2022). A review of English learning strategies and vocabulary size. *Modern Linguistics*, 10(7), 1411–1417. (in Chinese) <https://doi.org/10.12677/ml.2022.107188>
- Zhou, J. R., Yang, X. J., & Mao, M. H. (2016). A multimodal analysis of mobile vocabulary apps. *Journal of Harbin Vocational College*, (1), 113–115. (in Chinese)