ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

An Investigation of the Effectiveness of AI-Generated Contextual Texts for Extracurricular Vocabulary Acquisition among Tenth-Grade Students

Liu Yumei[1,a,*], Jin Zixi[2,b]

[1]*School of Foreign Languages, Lingnan Normal University, Zhanjiang, China*
[2]*Chow Yei Ching School of Graduate Studies, City University of Hong Kong, Hong Kong, SAR, China*
[a]*1045944776@qq.com,* [b]*zixijin2-c@my.cityu.edu.hk*
[*]*Corresponding author*

Abstract: *This quasi-experimental study investigates the effectiveness of AI-generated contextual texts extracurricular resources for vocabulary acquisition among 16 Chinese tenth-grade students with documented lexical difficulties. Grounded in the New Curriculum Standards' emphasis on vocabulary learning and the comprehensible input hypothesis, the research compared a traditional word-list memorization approach (control group) with an AI-mediated contextual learning approach using ChatGPT 3.5-generated passages (treatment group). Over three treatment cycles, participants engaged in 20-minute after-class study sessions, followed by productive vocabulary post-tests and semi-structured interviews. Despite non-significant statistical differences in post-test scores ($p > .05$) attributable to limited sample size and treatment duration, convergent analysis of quantitative and qualitative data yielded pedagogically significant findings. AI-generated texts demonstrated capacity to: (1) scaffold lexical learning through thematically coherent, difficulty-calibrated contexts aligned with the "i+1" principle; (2) promote deeper semantic encoding via output-driven assessment tasks; and (3) function as schematic bridges integrating new vocabulary with existing knowledge frameworks. Qualitative feedback revealed that while the control group experienced rapid engagement decline and lexical attrition, the treatment group reported enhanced contextual inferencing, accelerated acquisition, and sustained motivational appeal, though concerns about long-term sustainability and learner autonomy emerged. Four evidence-based pedagogical implications are proposed: hybrid resource integration combining AI texts with explicit phonological instruction; standards-aligned technology mediation; systematic strategy-based instruction; and transformation of AI output into varied productive exercises. Study limitations include small sample size, absence of delayed post-tests, and restricted student-AI interaction. Future research should employ longitudinal designs, delayed retention measures, and learner-controlled AI models to establish robust effect sizes and scalability.*

Keywords: *AI-generated Contextual Texts; Vocabulary Acquisition; Extracurricular Learning; Contextualized Instruction; Senior High School Students*

1. Introduction

Vocabulary constitutes the cornerstone of language proficiency. As Wilkins (1972, p.111) aptly asserted, “Without grammar, very little can be conveyed; without vocabulary, nothing can be conveyed”. Despite its paramount importance, vocabulary learning among Chinese senior high school students remains characterized by low efficiency, attributable to intense curricular pressure (Liang, 2018) and ineffective learning strategies (Wang, 2022). Compounding this issue is the tendency among some educators to neglect required lexical items that do not explicitly appear in reading passages (Wang, 2011). Given the inherent constraints of instructional time, extracurricular vocabulary acquisition emerges as a critical mechanism to supplement and extend in-class learning. *The Digital Literacy Framework for Teachers* (Ministry of Education of China, 2022, p. 3) mandates educators' capacity to “integrate digital technology resources into teaching and learning activities, encompassing digital instructional design, implementation, assessment, and collaborative education.” In response, English teachers in basic education have been actively exploring the pedagogical applications of Generative AI (GAI) to enhance student engagement and facilitate personalized instruction. Within the domain of vocabulary learning, preliminary research efforts have yielded promising insights. Zhou (2023) integrated ChatGPT as a conversational agent to generate word definitions, exemplar sentences, and related quizzes, reporting

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

positive effects on both vocabulary retention and learner motivation despite the absence of structured pedagogical scaffolding. Similarly, Leung et al. (2024) developed an AI-powered vocabulary learning application that contextualizes target words within interest-driven scenarios, observing significant improvements in student engagement alongside modest gains in lexical performance. Furthermore, (2022) demonstrated that AI-generated texts surpass traditional textbook materials in lexical exposure density, adaptability, and contextual customization per unit of time and length. Building upon these findings, this study investigates the efficacy of utilizing AI-generated, contextualized thematic texts as extracurricular learning materials for required vocabulary—particularly lexical items in non-reading sections—of the 2019 edition of the PEP *Senior High School English textbook*. The primary objectives are to alleviate students' after-school learning burden while enhancing both their interest and efficiency in vocabulary acquisition.

The Senior High School English Curriculum Standards (2017 Edition, 2020 Revision) (hereinafter referred to as the New Curriculum Standards, 2020, p. 22) explicitly stipulates that "vocabulary is not mere memorization or isolated drills, but rather comprehensive language practice activities within specific thematic contexts." Empirical support for this principle is provided by Vu and Peters (2022), whose experimental findings revealed that reading with textual input enhancement significantly outperformed reading-only conditions in vocabulary gains, whereas no significant difference was observed between reading-while-listening and reading-only approaches. This underscores the primacy contextually enriched, thematically coherent discourse in fostering effective vocabulary learning. Consequently, this study employs PEP Senior High School English Compulsory Course 2 (2019) as the curricular framework, utilizing GAI to generate discourse that embeds target vocabulary within meaningful contexts. A quasi-experimental design was implemented with first-year senior high school students in Zhanjiang, Guangdong Province, allocating 20 minutes of extracurricular time for intervention. The effectiveness of this approach was empirically validated through pre-tests, post-tests, and semi-structured interviews.

To ensure optimal text difficulty aligned with participants' proficiency levels, both linguistic complexity and thematic content of the AI-generated materials were benchmarked against the Level 1 academic achievement requirements specified in the New Curriculum Standards. Grounded in Piaget's constructivist principle that new knowledge is built upon existing schemas, the contextual scenarios deliberately designed to mirror students' authentic life experiences and integrate their prior interdisciplinary knowledge, thereby strengthening lexical knowledge construction and enhancing learning outcomes. According to the Level 1 academic proficiency descriptors (New Curriculum Standards, 2020, pp. 47–48), learners upon completing their first year of senior high school should demonstrate the following competencies:

a) During decoding target words in contexts, students can understand the function, denotation, and connotation meaning, and the writer's attitude and intention of vocabulary in specific contexts;

b) During the practice of learning, students can identify different semantic and logical relationships shown in contexts;

c) During exposure to target words and their collocations in contexts, students can purposefully select vocabulary and grammatical structures to express the precise and logical relationship between meanings.

2. Research Design and Methodology

This study employed a quasi-experimental design involving 16 first-year senior high school participants. Sampling was conducted through purposive selection of students from a public high school in Zhanjiang, Guangdong Province, who exhibited persistent challenges in vocabulary acquisition as evidenced by their average English scores falling span from 80/150 to 107/150 across three consecutive school assessments. These participants demonstrated strong intrinsic motivation to improve their lexical competence in preparation for the heightened academic demands of grades 11 and 12, thereby ensuring high levels of engagement and cooperation throughout the intervention period.

2.1 Research Instruments

Four principal instruments were utilized to ensure comprehensive data triangulation: structured interviews, pre- and post-tests, and a selected Generative AI model.

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

2.1.1 Interviews

A semi-structured interview protocol was administered post-intervention to elicit participants' perceptions regarding the efficacy and acceptability of AI-assisted vocabulary learning. The interview comprised four core items:

a) To what extent do you perceive this AI-enhanced approach as effective for expanding your vocabulary repertoire?

b) How does this method compare with your previous vocabulary learning strategies in terms of preference, and what factors inform this preference?

c) Would you be inclined to continue this approach if implemented as optional extracurricular assignments?

d) What suggestions or additional feedback do you have regarding this pedagogical intervention?

2.1.2 Generative AI Model Selection

In this study, the AI model functioned as a pedagogical tool for instructors to generate contextualized texts embedding target vocabulary for after-class assignments. The design deliberately minimized direct student-AI interaction to ensure accessibility and scalability across diverse educational contexts, particularly in settings with limited technological infrastructure. This teacher-mediated approach enables broad student benefits through a single AI-accessible device, representing a cost-effective solution suitable for resource-constrained regions.

To identify the most pedagogically appropriate AI model, four mainstream GAI platforms—Copilot, ChatGPT-3.5, Kimi, and Wenxin-Yiyan—were systematically evaluated. Model selection was determined through robustness testing, which assesses output consistency across varied prompts. Drawing upon the Directional Expectation Test (DIR) methodology proposed by Ribeiro et al. and adapted by Gui, Xi et al. (2024), this study evaluated each model's stability in generating pedagogically suitable texts. The evaluation criteria for model suitability were operationalized as follows:

a) Capacity to generate coherent narratives incorporating all target lexical items;

b) Production of semantically plausible and logically structured discourse;

c) Adherence to specified length parameters (about 150 words per text);

d) Generation of content aligned with ethical and moral educational standards.

The selected model demonstrated superior robustness across these dimensions and was subsequently employed as the primary text-generation instrument. The standardized prompt template utilized for text generation was as follows:

"*I am an intermediate English learner. I need you to write a short story for me to memorise some new words. *The story should be around 150 words, and there must be a clear plot line in the story.*The story should contain ALL the following words or their transformed forms (if needed);*Except for the given words, please make the other vocabulary in the story as simple as possible.*All the language in the story must be grammatically and semantically acceptable. *Mark the new words in the story: acute dozen sour voyage environment character insect dismiss independent content possession figure govern corporation flame*"

2.1.3 Robustness Test Procedure

Three target word lists, each comprising 15 lexical items systematically sampled from the glossary appendix of the New Curriculum Standards (pp. 121-178), were employed to evaluate model consistency. The comparative performance metrics of the four AI models across these test sets are presented.

The robustness test results revealed that both Wenxin-Yiyan and Kimi exhibited critical deficiencies in lexical fidelity, with target word omission rates ranging from 60% (9/15 words) to 93% (14/15 words). Furthermore, Kimi generated a text of 221 words, substantially exceeding the designated length parameters. While Copilot and ChatGPT 3.5 demonstrated satisfactory performance across the metrics of word count, target word inclusion, and thematic appropriateness, analysis of linguistic complexity yielded divergent outcomes. Copilot's output registered a prohibitively high mean readability level of 44, whereas ChatGPT 3.5 achieved the lowest mean complexity score of 17.33, aligning most closely with the target proficiency level. Consequently, ChatGPT 3.5 was selected as the

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

research instrument for subsequent text generation tasks.

Table 1: Results of the Robustness Test.

List No.	AI Model	Metrics			
		Words in Texts	Target Words Contained	Level	Theme
1	**Copilot**	161	15/15	40	Bass the insect got a new job as a fruit tester.
2		134	15/15	39	Anna's advice to her brother
3		183	15/15	44	Buddy the blind dog saved a cat.
Mean		**159.33**	**15/15**	**41**	
1	**Chat GPT 3.5**	128	15/15	24	Emma's adventure during her voyage
2		142	15/15	14	Emma's writing workshop experience
3		162	15/15	14	Tom the blind settled the crisis of river blocking.
Mean		**144**	**15/15**	**17.33**	
1	**Kimi**	117	15/15	31	**A strange sour smelled insect inspired a group of adventurers**
2		184	14/15	24	Lucy and Tom helped each other
3		221	15/15	10	Tom the blind saved a dog
Mean		**174**	**14.67/15**	**21.67**	
1	**Wenxin-Yiyan**	175	9/15	38	The adventure of Alice the entomologist
2		168	15/15	64	Emily turned to Dr. Smith for academic help.
3		179	14/15	44	Max the dog saved its blind host.
Mean		**174**	**12.67/15**	**48.67**	

2.1.4 Test

The study employed a pre-test/post-test design to isolate the effects of the intervention and measure gains in lexical competence. The pre-test was designed to establish baseline data by confirming that target lexical items were unfamiliar to participants, thereby minimizing the confounding effect of prior knowledge. The post-test assessed participants' ability to apply the acquired vocabulary in context-specific tasks.

The pre-test instrument was adapted from Nation's (1990, pp. 31-33) Vocabulary Size Test (VST), grounded in his theoretical framework distinguishing receptive from productive vocabulary knowledge. Receptive knowledge encompasses the ability to identify and retrieve the meaning of lexical items in written or auditory input, including awareness of their typical grammatical patterns and collocational behavior. Productive knowledge, conversely, entails command of a word's phonological, orthographic, and semantic properties, alongside the capacity to deploy it appropriately with correct grammatical structures and collocations in context. Crucially, mastery of receptive knowledge is a prerequisite for the development of productive knowledge.

The receptive format of the VST was deemed suitable for this study's pre-test as it effectively gauges whether target words are novel to learners. In principle, an ideal pre-test would yield a mean score of zero, indicating complete unfamiliarity with the lexical items. However, to account for the probability of correct guessing, this study established an acceptable scoring range of 0% to 50%. For a ten-item test, for instance, random guessing can be statistically modeled as a Bernoulli trial exhibiting the following characteristics:

$$x \sim B(10, 0.25) \tag{1}$$

$$x = 0,1,2,3,4,5,6,7,8,9,10 \tag{2}$$

The probability for test-takers to get a scoring rate more than 50 %will be:

$$f(x) = P(x = 5) + P(x = 6) + P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) \tag{3}$$

Fill the relevant value into the formular for Probabilities for a binomial random variable X:

$$P(x) = \binom{n}{x} p^x (1 - P)^{n-x} \tag{4}$$

$$\text{n=10, P=0.25} \tag{5}$$

After the calculation, 7.6% is the value of $f(x)$.

Statistical analysis reveals that for a ten-item test, the probability of achieving a scoring rate exceeding 50% through random guessing alone is merely 7.6%. This implies that when accuracy surpasses this threshold, the likelihood that correct answers stem from genuine receptive knowledge rather than chance rises to 92.4%. Such performance would indicate that test-takers possess prior familiarity with the target lexical items, necessitating their replacement to prevent pre-existing

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

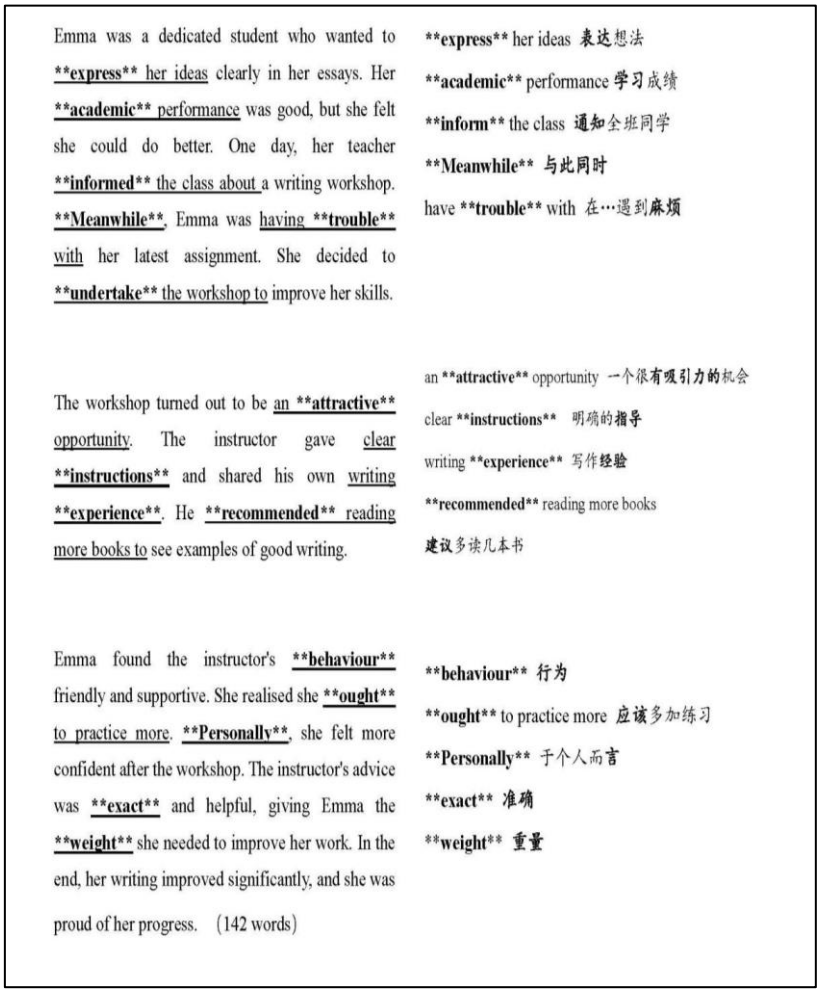
vocabulary knowledge from confounding the experimental results.

The post-test was adapted from the productive vocabulary assessment framework developed by Nation and Laufer (1999, pp. 36-37), which operationalizes the principles of the Output Hypothesis by compelling learners to generate target words in authentic contexts. In this cued-recall format, each target word is removed from a meaningful sentence and replaced with a blank, with the first few letters provided as orthographic prompts. Test-takers must retrieve and produce the appropriate lexical item based on contextual cues and the initial letter hints, as demonstrated in the following exemplar:

The book covers a series of isolated epis_____from history.

However, such minimalist contexts compromise ecological authenticity. To address this limitation, the present study replaced isolated sentences with AI-generated short passages that provide richer contextual scaffolding while preserving the core principle of eliciting active lexical production—a pedagogically sound approach grounded in the Output Hypothesis of second language acquisition. Prior to post-test administration, participants received explicit instructions that alternative responses were permissible, thus encouraging the retrieval of varied lexical and collocational forms congruent with contextual constraints. Scoring criteria stipulated that credit be awarded to any response demonstrating both grammatical accuracy and semantic coherence within the passage.

2.2 Experimental Procedures


Emma was a dedicated student who wanted to **express** her ideas clearly in her essays. Her **academic** performance was good, but she felt she could do better. One day, her teacher **informed** the class about a writing workshop. **Meanwhile**, Emma was having **trouble** with her latest assignment. She decided to **undertake** the workshop to improve her skills.

express her ideas 表达想法
academic performance 学习成绩
inform the class 通知全班同学
Meanwhile 与此同时
have **trouble** with 在…遇到麻烦

The workshop turned out to be an **attractive** opportunity. The instructor gave clear **instructions** and shared his own writing **experience**. He **recommended** reading more books to see examples of good writing.

an **attractive** opportunity 一个很有吸引力的机会
clear **instructions** 明确的指导
writing **experience** 写作经验
recommended reading more books
建议多读几本书

Emma found the instructor's **behaviour** friendly and supportive. She realised she **ought** to practice more. **Personally**, she felt more confident after the workshop. The instructor's advice was **exact** and helpful, giving Emma the **weight** she needed to improve her work. In the end, her writing improved significantly, and she was proud of her progress. (142 words)

behaviour 行为
ought to practice more 应该多加练习
Personally 于个人而言
exact 准确
weight 重量

Figure 1: Generated Text Sample. The Chinese characters are the translation of target words and collocations.

The sixteen participants were randomly assigned to two groups (n=8 each) using stratified sampling to ensure equivalence in gender distribution and English proficiency levels. Group A (A1–A8) served as the control group, while Group B (B1–B8) functioned as the experimental (treatment) group. Both cohorts convened in separate unoccupied classrooms during a 20-minute post-instructional period.

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

Following pre-test administration, the control group received a word list identical to the textbook glossary entries, including Chinese translations and phonetic transcriptions. Participants were permitted to employ any self-regulated memorization strategy (e.g., silent reading, oral rehearsal, orthographic copying) but were instructed to maintain individual focus without peer interaction. Concurrently, the experimental group received the AI-generated passage in which all target lexical items and their collocations were typographically enhanced through bold formatting in both English and Chinese versions. Participants were directed to study the text holistically within the 20-minute timeframe, with emphasis on comprehending contextual usage patterns.

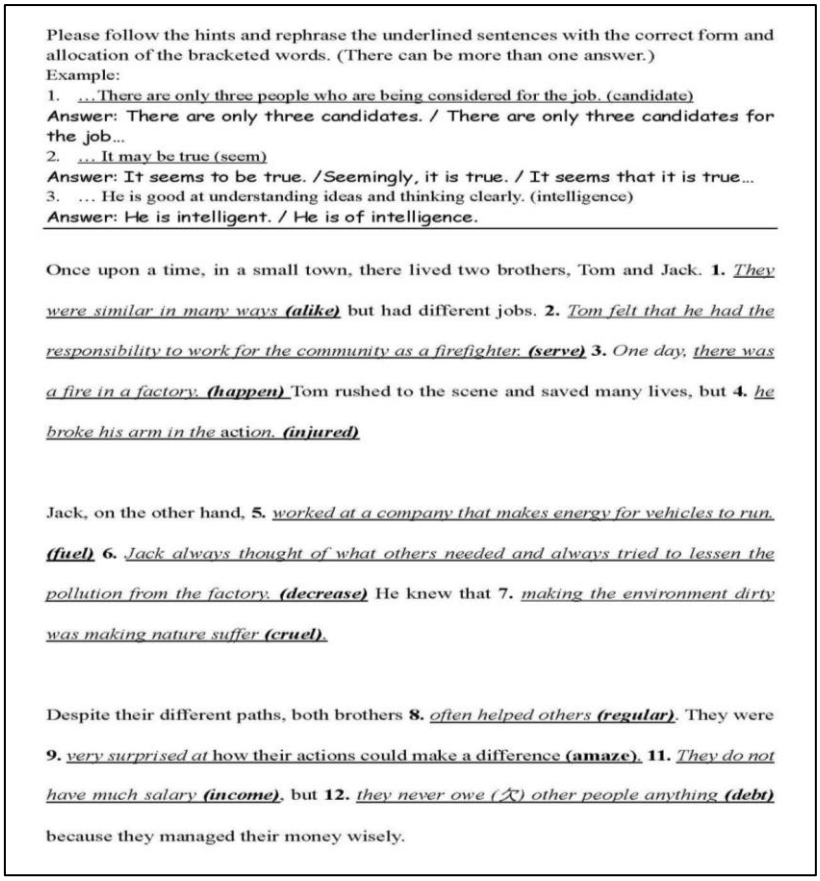
Upon completion of the study period, all instructional materials were collected from both groups to prevent further review. Subsequently, all participants completed the post-test, which required productive application of the target vocabulary in novel contexts. This three-cycle experimental protocol was administered on separate occasions with distinct word lists.

Following the final experimental cycle, semi-structured interviews were conducted with all participants to elicit their perceptions and evaluative feedback regarding the pedagogical intervention. To facilitate lexical focus, the AI-generated texts featured underlined target word collocations, with corresponding Chinese translations provided in a parallel column. Sample materials are displayed in Figure 1, Figure 2, and Figure 3.

1. You can **express** it.
   A. to tell your feeling or thought
   B. to tell a lie
   C. to tell a joke
   D. to retell a story that you heard elsewhere
2. He has no **academic** background.
   A. relating to jobs
   B. relating to marriage
   C. relating to education
   D. relating to a personal relationship
3. They will **inform** you.
   A. to give someone money
   B. to tell someone about something
   C. to punish someone's fault
   D. to praise someone
4. **Meanwhile**, he did it.
   A. unfortunately
   B. something looks like another
   C. something else is happening at the same time
   D. something seems to happen
5. He has **trouble**.
   A. power and strength
   B. problems or difficulties
   C. money and wealth
   D. a good mood
6. He is **attractive**.
   A. awful to look at
   B. surprised to look at
   C. regretful to look at
   D. pleasant to look at
7. These are the **instructions**.
   A. a statement telling people how to do or use something
   B. an advertisement to advocate people to buy something
   C. a notice to remind people to do something
   D. an exercise for students to train their skills
8. I have some **experience**.
   A. extra money that you gain from hard work
   B. lesson that you gain from failure
   C. knowledge or skill that you gain from doing something
   D. an idea that you gain from reading books
9. I **recommend** it.
   A. to warn someone not to do something
   B. to advise someone to do something
   C. to persuade someone to give up something
   D. to tell someone that you own something
10. I know his **behaviour**.
    A. the things that a person does
    B. the things that a person is good at
    C. the things that a person like
    D. the things that a person hate
11. You **ought** to do it.
    A. someone wants to do something
    B. someone should do something
    C. someone feels happy to do something
    D. someone feels sorry to do something
12. **Personally**, he can do it.
    A. giving a positive opinion
    B. giving an objective opinion
    C. giving others' opinion
    D. giving your own opinion
13. It is **exact**.
    A. completely correct in every detail
    B. unclear, confusing, or not certain
    C. completely wrong in every detail
    D. almost correct
14. It has **weight**.
    A. how long something is
    B. how large something is
    C. how wide something is
    D. how heavy something is
15. It is **significant**.
    A. necessary
    B. important
    C. delicious
    D. ugly

Figure 2: Pre-test Sample.

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031



Please follow the hints and rephrase the underlined sentences with the correct form and allocation of the bracketed words. (There can be more than one answer.)
Example:
1. …There are only three people who are being considered for the job. (candidate)
Answer: There are only three candidates. / There are only three candidates for the job…
2. … It may be true (seem)
Answer: It seems to be true. /Seemingly, it is true. / It seems that it is true…
3. … He is good at understanding ideas and thinking clearly. (intelligence)
Answer: He is intelligent. / He is of intelligence.

Once upon a time, in a small town, there lived two brothers, Tom and Jack. **1.** *They were similar in many ways* ***(alike)*** but had different jobs. **2.** *Tom felt that he had the responsibility to work for the community as a firefighter.* ***(serve)*** **3.** *One day, there was a fire in a factory.* ***(happen)*** Tom rushed to the scene and saved many lives, but **4.** *he broke his arm in the* action. ***(injured)***

Jack, on the other hand, **5.** *worked at a company that makes energy for vehicles to run.* ***(fuel)*** **6.** *Jack always thought of what others needed and always tried to lessen the pollution from the factory.* ***(decrease)*** He knew that **7.** *making the environment dirty was making nature suffer* ***(cruel).***

Despite their different paths, both brothers **8.** *often helped others* ***(regular)***. They were **9.** *very surprised at* how their actions could make a difference **(amaze)**. **11.** *They do not have much salary* ***(income)***, but **12.** *they never owe (欠) other people anything* ***(debt)*** because they managed their money wisely.

Figure 3: Post-test Sample.

3. Results and Analysis

3.1 Qualitative Feedback from Participants

3.1.1 Control Group (GA) Perspectives

Participants in the control group expressed ambivalent evaluations of the traditional vocabulary memorization protocol. On the positive dimension, several students acknowledged that the pre- and post-test structure reinforced retention and strengthened reading comprehension skills. However, they identified substantive limitations inherent in the word-list approach: (a) insufficient attention to phonological accuracy, (b) precipitous decline in engagement following the initial session, and (c) rapid lexical attrition without systematic review. Additionally, participants reported that the 20-minute duration felt excessive, and the pedagogical materials lacked motivational appeal. Environmental distractions, including excessive noise, and the potential imposition of mandatory participation further diminished their willingness to engage. Several students indicated that their commitment would be contingent upon affective factors (e.g., daily mood) and assessment pressure (e.g., threat of next-day dictation). In essence, while GA participants valued the iterative testing component, they critiqued the method for its phonetic inadequacies, motivational deficits, and susceptibility to contextual interference.

3.1.2 Experimental Group (GB) Perspectives

The treatment cohort similarly articulated a dichotomous assessment of the AI-enhanced contextual approach. Affirmative responses centered on three affordances: (1) enhanced lexical comprehension through contextual inferencing and definitional verification, (2) accelerated acquisition via authentic usage exemplars, and (3) superior engagement compared to rote memorization paradigms. Several participants explicitly endorsed the systematic nature of the methodology.

Conversely, reservations emerged regarding long-term sustainability and learner autonomy. A subset of students expressed preference for self-directed learning strategies that afforded greater lexical

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

selection freedom, while others voiced concerns about potential "chore-like" routinization if implemented as compulsory daily practice. Suggestions for optimization included: (i) integrating visual scaffolds (e.g., illustrative images) to augment motivational appeal, (ii) embedding pre-test components more integrally within the learning sequence rather than as isolated evaluative measures, and (iii) providing enhanced phonetic transcription support to address pronunciation difficulties. In summary, GB participants recognized the immediate pedagogical efficacy of the AI-mediated approach but questioned its durability and flexibility as a sustained intervention model.

3.2 Quantitative Data Analysis

Statistical analysis was conducted using IBM SPSS Statistics 26.0. Pre-test scores between GA and GB were subjected to independent samples t-test to ascertain inter-group equivalence in baseline lexical proficiency prior to intervention. The analytical results are presented in Tables 1 and 2.

Table 2: Descriptive Statistics of Pre-test Scores.

Group		N	Mean	Std. Deviation	Std. Error Mean
Pre1	A	8	5.00	1.309	0.463
	B	8	4.88	1.959	0.693
Pre2	A	8	4.13	2.357	0.833
	B	8	4.25	1.035	0.366
Pre3	A	8	4.88	2.232	0.789
	B	8	5.00	1.604	0.567

As presented in Table 2, participants' mean scores across all three pre-test administrations remained at or below 5 points, thereby satisfying the baseline prerequisite established in Section 2.1 (performance below 50% of the total possible score of 15 points). This confirms that participants were appropriately unfamiliar with the target lexical items prior to intervention.

Table 3: Independent Samples Test.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Pre1	Equal variances assumed	2.244	0.156	0.150	14	0.883	0.125	0.833	-1.662	1.912
	Equal variances not assumed			0.150	12.212	0.883	0.125	0.833	-1.687	1.937
Pre2	Equal variances assumed	4.564	0.051	0.137	14	0.893	-0.125	0.910	-2.077	1.827
	Equal variances not assumed			-0.137	9.604	0.894	-0.125	0.910	-2.164	1.914
Pre3	Equal variances assumed	2.133	0.166	-0.129	14	0.899	-0.125	0.972	-2.209	1.959
	Equal variances not assumed			-0.129	12.706	0.900	-0.125	0.972	-2.229	1.979

As presented in Table 3, Levene's Test for Equality of Variances yielded p-values of .156, .051, and .166 across the three pre-test administrations. All values exceed the .05 significance threshold, thereby confirming homogeneity of variance between groups. Independent samples t-tests were subsequently conducted to compare mean scores. The resulting p-values for Equality of Means [Sig. (2-tailed)] were .883, .893, and .899, respectively—all non-significant at $\alpha = .05$. This indicates no statistically significant difference in pre-test performance at the 95% confidence level, thus establishing baseline equivalence. In other words, participants' prior knowledge of target vocabulary was statistically comparable across groups, satisfying a critical prerequisite for subsequent experimental comparisons.

Tables 4 and 5 present the post-test data analysis results.

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

Table 4: Descriptive Statistics of Post-test Scores

Group		N	Mean	Std. Deviation	Std. Error Mean
Post1	A	8	10.50	3.505	1.239
	B	8	9.13	2.997	1.060
Post2	A	8	10.13	3.563	1.260
	B	8	9.38	3.378	1.194
Post3	A	8	13.50	1.927	0.681
	B	8	10.63	3.583	1.267

Table 4 reveals that Group A's mean scores on the three post-tests (10.5, 10.13, and 13.5, respectively) consistently exceeded those of Group B (9.13, 9.38, and 10.63, respectively).

Table 5: Independent Samples Test of Post-test Scores

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Post1	Equal variances assumed	0.689	0.420	0.843	14	0.413	1.375	1.630	-2.122	4.872
	Equal variances not assumed			0.843	13.670	0.414	1.375	1.630	-2.130	4.880
Post2	Equal variances assumed	0.282	0.604	0.432	14	0.672	0.750	1.736	-2.973	4.473
	Equal variances not assumed			0.432	13.960	0.672	0.750	1.736	-2.974	4.474
Post3	Equal variances assumed	3.057	0.102	1.999	14	0.065	2.875	1.438	-0.210	5.960
	Equal variances not assumed			1.999	10.737	0.072	2.875	1.438	-0.301	6.051

As indicated in Table 5, Levene's Test for Equality of Variances yielded p-values of .420, .604, and .102 across the three post-test administrations, all exceeding the .05 significance threshold. This confirms homogeneity of variance between groups, thereby satisfying the assumption for the independent samples t-test application.

The t-test results for Equality of Means produced p-values [Sig. (2-tailed)] of .413, .672, and .065, respectively. All values remain above the .05 criterion, indicating that despite Group A's consistently higher mean scores, the observed differences fail to reach statistical significance at the 95% confidence level. Consequently, there is insufficient empirical evidence to conclude that the AI-mediated vocabulary learning approach produced measurably superior outcomes compared to traditional memorization techniques.

4. Major Findings, Pedagogical Implications, and Future Directions

4.1 Summary of Key Findings

Despite the non-significant statistical outcomes attributable to the limited number of treatment cycles, constrained duration, and modest sample size, the convergent analysis of quantitative performance data and qualitative interview responses yields several pedagogically significant insights for extracurricular vocabulary acquisition.

First, the study demonstrates that Generative AI can produce thematically coherent, pedagogically calibrated learning contexts that embed target vocabulary in alignment with the "i+1" comprehensible input principle. The findings indicate that AI systems can dynamically adjust both lexical sophistication and textual length to match learners' Zone of Proximal Development, thereby optimizing input processing efficiency.

Second, context-embedded vocabulary assessments — specifically post-test tasks requiring productive application in novel discourse contexts — demonstrated efficacy in facilitating lexical retention and deepening semantic encoding. This finding corroborates the Output Hypothesis principle

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

that pushes output to strengthen memory consolidation beyond receptive exposure alone.

Third, AI-generated texts functioned as schematic bridges that integrated newly encountered lexical items with students' existing knowledge frameworks. This feature promoted intensive knowledge construction in working memory, offering a promising solution to the structural limitations of traditional textbook presentations that often divorce vocabulary from authentic usage contexts.

Finally, from a material-design perspective, the AI-generated content exhibited three salient affordances: (a) enhanced motivational appeal through thematic customization, (b) adaptability to diverse proficiency levels, and (c) embedded moral-cultural value systems that stimulate learner interest while maintaining ideological appropriateness.

4.2 Pedagogical Implications

Based on these findings, four evidence-based recommendations can be formulated for senior high school vocabulary instruction:

Implication 1: Hybrid Resource Integration

To address the organizational deficiencies inherent in conventional textbooks, practitioners should strategically deploy AI-generated contextual texts as supplementary after-class resources. This approach must be complemented by explicit phonological instruction during classroom time, equipping students with autonomous decoding strategies that enhance orthographic-phonological mapping and long-term retention.

Implication 2: Standards-Aligned Technology Mediation

Mindful of the time constraints characterizing senior high school curricula, educators should operationalize the New Curriculum Standards by employing GAI as a pedagogical scaffold rather than a replacement for teacher judgment. This involves using AI to optimize instructional design, judiciously select contextually rich content, and differentiate materials according to individual learning trajectories.

Implication 3: Strategy-Based Instruction

In response to documented inefficiencies in current vocabulary learning approaches, teachers are advised to systematically impart metacognitive strategies, including contextual inference techniques and morphological analysis skills, within classroom instruction. Subsequently, AI-generated materials can serve as structured practice venues where students apply these strategies independently, thereby fostering learner autonomy and accelerating acquisition efficiency.

Implication 4: Contextualized Practice Design

To consolidate lexical gains, practitioners should adapt AI-generated texts into varied, context-embedded productive exercises (e.g., cued recall tasks, semantic mapping, collocational restructuring). This transforms passive consumption of AI output into active knowledge construction, reinforcing the form-meaning mapping essential for depth of processing.

4.3 Limitations and Future Research

Several limitations of the present study warrant acknowledgment and inform future research agendas. First, the small-scale, short-duration design limited statistical power and ecological validity, constraining the generalizability of findings. Second, the absence of delayed post-tests precluded investigation of long-term retention effects. Third, the teacher-mediated AI application model, while economically practical, restricted opportunities to examine direct student-AI interaction patterns.

Future research should address these gaps through longitudinal designs extending over academic semesters, incorporating delayed post-tests to track lexical retention trajectories, and exploring learner-controlled AI interaction models. Additionally, investigations into the differential effects of AI-generated texts across proficiency bands and the impact of multimedia augmentation (e.g., integration of visual scaffolds) would further elucidate the pedagogical potential of this emerging technology. Large-scale quasi-experimental studies employing cluster randomization are needed to establish robust effect sizes and confirm the scalability of this approach in diverse educational contexts.

ISSN 2522-6398 Vol. 8, Issue 10: 208-218, DOI: 10.25236/FER.2025.081031

Acknowledgements

This work was supported by Research and Practice on Senior High School English Teacher Education Based on Academic Quality Monitoring (Yue Jiao Gao Han [2023] No. 29), Research and Practice Project on New Normal Education Construction Promoting High-Quality Development of Basic Education; Research on the Adaptability of English Teachers in Western Guangdong to the New Curriculum Standards for Compulsory Education (2023GXJK382), Guangdong Provincial Education Science Planning Project; Research on the Thresholds in Alice Munro's Works (ZW2021016), Humanities and Social Sciences Talent Program of Lingnan Normal University.

References

[1] Gui, T., Xi, Z., Zheng, R., Wang, J., Zhang, Y., & Wang, X. (2024). A survey on the robustness of deep learning-based natural language processing. Chinese Journal of Computers, 47(1), 90–112.
[2] Laufer, B., & Nation, I. P. (1999). A vocabulary size test of controlled productive ability. Language Testing, 16(1), 36–55.
[3] Leung, J., Pataranutaporn, P., Danry, V., & Zhu, J. (2024). Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In Proceedings of the CHI conference on human factors in computing systems (pp. 11–16). Association for Computing Machinery.
[4] Liang, E. (2018). A study on the effect of vocabulary memorization under the background of senior high school English morning reading (Unpublished master's thesis). Central China Normal University.
[5] Ministry of Education of China. (2022). Digital literacy for teachers (p. 3). China Education Publishing & Media Group.
[6] Ministry of Education of the People's Republic of China. (2020). Senior high school English curriculum standards (2017 edition, 2020 revision). People's Education Press.
[7] Nation, I. S. P. (1990). Teaching and learning vocabulary. Heinle & Heinle Publishers.
[8] Nation, P., & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7), 9–12.
[9] Shen, X. (2022). A comparative study of reading texts in PEP and FLTRP senior high school English textbooks (Unpublished master's thesis). Huaibei Normal University.
[10] Vu, D. V., & Peters, E. (2022). Learning vocabulary from reading-only, reading-while-listening, and reading with textual input enhancement: Insights from Vietnamese EFL learners. RELC Journal, 53(1), 85–100.
[11] Wang, M. (2011). A survey of English teachers' textbook use in senior high schools. Journal of Foreign Languages School of Shandong Normal University (Basic English Education), 13(5), 34–37.
[12] Wang, X. (2022). A study on the application of the palace memory strategy in senior high school English vocabulary teaching (Unpublished master's thesis). Southwest University.
[13] Wen, Y. (2022). Design of an AI-powered seamless vocabulary learning for young learners. In Proceedings of the 30th International Conference on Computers in Education (pp. 659–661). Asia-Pacific Society for Computers in Education.
[14] Wilkins, D. A. (1972). Linguistics in language teaching. MIT Press.
[15] Zhou, H. (2023). A study on the auxiliary application of ChatGPT in college English vocabulary teaching. Journal of Social Science of Jiamusi University, 41(6), 192–194.