

## Article

# A Multimodal Human–AI Instructional Framework for Productive Vocabulary Development: A Classroom Evaluation of a Coordinated LLM–ASR System

Shivan Mawlood Hussein \* and Mustafa Kurt 

Department of English Language Teaching, Research Center for Applied Linguistics, Near East University, Nicosia 99138, North Cyprus, Turkey; mustafa.kurt@neu.edu.tr

\* Correspondence: shivan.mawlood.h@gmail.com

## Abstract

This study examined the implementation and instructional effectiveness of a multimodal AI-supported instructional framework integrating a generative AI assistant (Microsoft Copilot) with a speech-recognition-based mobile learning application (Mondly) to support productive vocabulary development in EFL higher education. Unlike studies focusing on single AI tools, this study evaluates a coordinated dual-module instructional configuration combining LLM-based lexical support with ASR-based spoken retrieval practice within a structured classroom routine. The proposed framework can be viewed as a lightweight socio-technical instructional arrangement in which learners engage with complementary AI components through guided feedback and repeated practice. A quasi-experimental pretest–post-test control group design was conducted over an eleven-week semester with 64 first-year EFL students at an Iraqi university. Productive vocabulary knowledge was measured using the Productive Vocabulary Levels Test (PVLTL), and data were analyzed using mixed-design ANOVA. Results revealed a statistically significant Time  $\times$  Group interaction with a large effect size, indicating greater productive vocabulary gains in the AI-supported condition compared with traditional instruction. Qualitative findings further suggested perceived improvements in lexical retrieval, sentence construction, pronunciation accuracy, and learner engagement. From an instructional perspective, the findings suggest that learning gains were associated with the coordinated use of complementary AI tools within a structured classroom workflow. This study provides a practical instructional model that may be adaptable to comparable resource-constrained higher-education contexts.

**Keywords:** socio-technical systems; human–AI interaction; instructional system design; coordinated AI learning; generative AI; large language model (LLM); automated speech recognition (ASR); productive vocabulary; EFL higher education; classroom-based intervention



Academic Editor: Vladimír Bureš

Received: 19 March 2026

Revised: 21 April 2026

Accepted: 24 April 2026

Published: 27 April 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

Artificial intelligence (AI) technologies are increasingly integrated into higher-education learning environments to support adaptive feedback, intelligent interaction, and data-driven instructional approaches [1,2]. In recent years, advances in generative AI and speech processing technologies have enabled the development of intelligent learning platforms capable of providing contextualized explanations, automated language feedback, and interactive human–computer communication within digital learning environments [3,4]. In the domain of language education, these technologies have created

new opportunities to design AI-supported instructional approaches that combine large language models with speech-enabled applications to facilitate vocabulary learning, sentence generation, pronunciation practice, and iterative retrieval processes [3,5]. This is particularly evident when multimodal input is employed, as research has shown that combining visual and textual information can significantly enhance vocabulary acquisition and comprehension in L2 contexts [6]. Despite these technological advances, empirical classroom-based research remains limited regarding how multiple AI components can be systematically integrated into coherent instructional configurations that support productive lexical development in EFL higher-education contexts [2,7,8].

These developments are particularly relevant in resource-constrained EFL learning contexts, where limited instructional time, large class sizes, and restricted access to individualized feedback may constrain learners' opportunities for sustained lexical production [9–11]. In Iraq and similar EFL settings, university students may demonstrate receptive recognition of vocabulary yet still experience difficulty retrieving and accurately producing lexical items in speaking and writing. Such conditions highlight the need for instructional approaches that intensify structured productive practice within scheduled classroom time while remaining feasible, efficient within existing instructional time, and adaptable to technology-supported learning environments [1,12].

From an instructional perspective, lexical processing efficiency is central to effective second language development because it supports accurate communication and academic expression [9]. Productive vocabulary knowledge reflects learners' ability to retrieve and use lexical items accurately in contextualized language production [9,10]. However, vocabulary instruction in many EFL classrooms continues to rely heavily on memorization-oriented techniques and limited opportunities for active language production, which may not consistently promote noticing, repeated retrieval, pushed output, or long-term retention [10,13]. These challenges are further amplified in low-exposure learning environments where opportunities for communicative practice and productive language use are constrained [9,10].

AI-supported instructional approaches offer a potential response by enabling structured interaction processes and feedback-driven learning cycles [3,4]. Generative AI assistants such as Microsoft Copilot can provide contextualized explanations, sentence modeling, corrective feedback, and dialogic scaffolding that encourage active lexical production [3,14]. Similarly, speech-based mobile applications such as Mondly integrate listening and speaking tasks supported by automated speech recognition, promoting pronunciation-focused rehearsal and repeated spoken retrieval [5]. When implemented as a structured classroom-embedded instructional routine rather than an optional supplement, the complementary use of these tools may support active vocabulary use through guided output, repeated reformulation, and sustained engagement during scheduled lessons [3,10].

AI methodologies increasingly extend beyond purely data-driven models to include emerging data-physics and hybrid AI models that combine learned patterns with prior knowledge, structured constraints, or domain-informed guidance to improve robustness, interpretability, and efficiency in complex systems [15,16]. While such developments are especially prominent in engineering and computational modeling domains, classroom language learning environments require interactive, data-driven tools capable of providing real-time feedback, dialog generation, adaptive support, and pronunciation practice. Accordingly, the present study focuses on AI tools whose functions are closely aligned with pedagogical interaction and productive vocabulary development.

Although AI-assisted language learning has expanded rapidly, much of the existing literature has focused on individual tools, broad language learning outcomes, or learner perceptions rather than on integrated instructional configurations examined under controlled

classroom conditions [2,7,8]. Consequently, limited evidence exists regarding how complementary AI technologies can be coordinated within fixed instructional time to support measurable, productive vocabulary outcomes in EFL classrooms. Recent studies highlight the potential of combining language models, speech technologies, and adaptive learning features; however, classroom-based evaluations of such coordinated implementations remain scarce. Collectively, prior studies indicate promise for AI tools but provide relatively little controlled evidence regarding coordinated use for productive vocabulary development. The present study addresses this gap by evaluating an integrated Microsoft Copilot–Mondly instructional configuration in an authentic university EFL classroom context.

Accordingly, this study investigates the integrated use of Microsoft Copilot and Mondly as a coordinated AI-supported classroom approach for enhancing productive lexical performance in the Department of English Language Teaching at a university in Erbil, northern Iraq. Employing a mixed-methods quasi-experimental design, the study compares an AI-integrated instructional group embedded within regular classroom practice with a control group receiving conventional instruction without AI support. Productive vocabulary development was assessed using the Productive Vocabulary Levels Test (PVL) [17], and the quantitative findings were complemented by qualitative analyses examining learners' perceptions of AI-supported practice and their experiences with the integrated instructional configuration.

The Copilot–Mondly integration is conceptualized as an applied AI-supported instructional configuration in which a generative AI tool and a speech-based mobile learning tool operate jointly to support contextualized vocabulary production, feedback, and repeated practice during classroom instruction [18]. Specifically, Microsoft Copilot provides lexical explanation, sentence generation, and corrective feedback, whereas Mondly supports pronunciation rehearsal and repeated spoken retrieval through automated speech recognition (ASR). Together, these components form a structured, multimodal instructional workflow that links explanation, guided production, rehearsal, and retrieval within scheduled classroom practice. Rather than treating AI tools as separate interventions, the present study evaluates their coordinated use within a functional classroom workflow.

By examining these components as a coordinated instructional arrangement, the study explores whether the structured integration of complementary AI functions is associated with improved learning outcomes. This perspective emphasizes that pedagogical sequencing, rather than tool availability alone, may influence effectiveness. Accordingly, AI-supported vocabulary learning is examined here as a structured classroom routine involving input, guided practice, feedback, and repeated retrieval.

The novelty of the present study lies not in the invention of new tools, but in the classroom-based evaluation of a fixed Copilot–Mondly instructional routine specifically designed to improve productive vocabulary under equivalent instructional time conditions.

### *Research Questions*

1. What is the effect of the integrated Copilot–Mondly AI-assisted instructional configuration on EFL students' productive vocabulary development?
2. What are EFL students' perceptions of AI-assisted vocabulary learning when using the integrated Copilot–Mondly instructional configuration?

## **2. Related Work and Conceptual Foundations**

Recent developments in artificial intelligence (AI) have accelerated the adoption of AI-supported language learning tools in vocabulary instruction by expanding opportunities for adaptive practice, immediate feedback, and structured learning support within formal educational settings [1,2]. Research in vocabulary acquisition consistently indicates

that sustained lexical development depends on repeated retrieval, meaningful output, and contextualized use of target items, which are associated with stronger long-term retention and productive knowledge [9,10]. Advances in conversational and generative AI have extended these possibilities by enabling interactive language production, immediate corrective feedback, and personalized scaffolding that may support learners' ability to retrieve and produce lexical items accurately in communicative tasks [3]. Studies on multimedia-based input further indicate that combining visual and textual information can strengthen form–meaning connections and increase exposure to lexical items, thereby facilitating vocabulary learning [6]. However, limited classroom-based research has examined how generative language models and speech recognition technologies can be pedagogically coordinated within structured instructional routines that support both contextualized lexical generation and pronunciation-focused retrieval practice [2,7,8].

Within recent educational technology discussions, generative AI has been recognized as a resource for widening access to guided learning support and improving instructional responsiveness in higher education [12,19]. This is especially relevant in resource-constrained EFL contexts such as Iraq, where students often have limited opportunities to actively use English and may not receive sufficient individualized support to develop productive lexical competence [9,10]. Research in conflict-affected EFL settings further suggests that AI tools can enhance engagement and vocabulary outcomes, although effectiveness depends on context-sensitive implementation and infrastructure conditions [20]. In this respect, classroom-embedded AI-supported frameworks may help extend structured lexical retrieval and productive language use within existing instructional constraints [21].

### *2.1. Productive Vocabulary Development in EFL Learning*

Productive vocabulary knowledge is widely regarded as a central component of second language development because it directly influences learners' ability to express ideas accurately and participate effectively in academic communication [9]. Developing productive lexical competence requires repeated retrieval and meaningful use of target words in spoken and written contexts over time [10]. However, vocabulary learning in many EFL environments remains heavily teacher-centered, with limited opportunities for sustained communicative production during scheduled lessons [22]. Such constraints may restrict productive vocabulary growth and weaken long-term retention, particularly where structured speaking and writing practice is limited [9,13]. Technology-supported instructional tools may therefore represent a practical approach by intensifying lexical retrieval and guided production within formal instruction while promoting equitable access under institutional time and resource constraints [12,19,23]. Taken together, these findings suggest that productive vocabulary development depends not only on exposure but on structured opportunities for repeated retrieval and meaningful output, which remain limited in many classroom contexts.

### *2.2. Generative AI Assistants and Vocabulary Learning (Microsoft Copilot)*

Generative AI assistants have gained increasing attention in language education because they provide interactive support, contextualized explanations, and dialog-based practice that encourage active language production [4]. Such tools may facilitate productive vocabulary development by generating example sentences, prompting contextualized use of target lexical items, and offering personalized scaffolding during real-time interaction [3]. Microsoft Copilot, embedded within Microsoft applications, can provide lexical explanations, contextual examples, and guided dialogue tasks that assist learners in retrieving and producing vocabulary meaningfully [3,24].

Chatbot-based research further indicates that conversational AI may support productive vocabulary acquisition by encouraging repeated lexical retrieval and contextualized output practice [3,14]. For instance, Zhang and Huang [14] reported that interaction with large language model-based chatbots promoted repeated retrieval and meaning-focused production tasks, contributing to improvements in learners' ability to use target lexical items accurately. In addition, learner-centered research suggests that generative AI tools may enhance perceived usefulness, motivation, and learner engagement by providing structured, responsive support within instructional settings, thereby facilitating sustained, productive language practice [25–27].

However, concerns remain regarding the reliability of AI-generated content, including potential inaccuracies that may affect productivity accuracy if outputs are not critically evaluated [28,29]. Issues related to data privacy, learner dependency, and bias also require careful pedagogical consideration [4]. These concerns highlight the importance of examining generative AI tools such as Copilot within structured classroom-based interventions. At the same time, most existing studies have examined generative AI tools in isolation, with limited controlled evidence on how they can be integrated with complementary instructional components to support sustained productive vocabulary development. These affordances further suggest that conversational AI may support vocabulary learning through iterative learner–AI interaction rather than one-directional content delivery.

### *2.3. Speech-Based Mobile Learning and Gamified Vocabulary Practice (Mondly)*

Mobile-assisted language learning has been shown to enhance accessibility and learner engagement by supporting structured practice within instructional contexts [22,30]. Meta-analytic evidence indicates that mobile learning interventions positively influence language learning outcomes, including vocabulary development, through sustained engagement and repeated retrieval [30]. Smartphone applications have also been identified as effective tools for increasing access to structured lexical practice in classroom-based activities [22].

Mondly, an AI-supported mobile language learning application, integrates gamified vocabulary practice with interactive speaking activities [5]. A key feature of Mondly is its automated speech recognition (ASR), which provides pronunciation feedback and repeated opportunities for spoken production [5]. Such speech-based engagement may support lexical retrieval, phonological processing, and more accurate spoken production through repeated opportunities for practice and feedback [31,32]. These effects are further supported by research indicating that interactive digital learning environments enhance engagement and reinforce vocabulary retention through repeated exposure and feedback [27]. Although Mondly shows potential as a vocabulary learning tool, empirical research examining its impact on productive vocabulary knowledge remains limited, particularly in under-resourced university EFL contexts [7]. Despite its potential, existing research has rarely examined how speech-based mobile applications such as Mondly can be integrated with other AI-supported tools within structured classroom instruction to enhance productive vocabulary outcomes.

### *2.4. Research Gap and Theoretical Framework*

Although AI-supported vocabulary learning has received increasing attention, many studies continue to examine single tools or isolated outcomes, leaving limited evidence on how complementary AI applications can work together to enhance productive vocabulary development [7,8]. Prior research has also reported mixed findings regarding the effectiveness of digital learning environments compared with traditional instructional approaches, indicating the need for further classroom-based investigation of integrated AI-supported approaches [33]. In particular, few studies have systematically investigated the combined

use of a generative AI assistant and a speech-based mobile learning application within scheduled classroom instruction, especially in resource-constrained EFL contexts such as Iraq [2,7].

To conceptualize this gap, the present study draws on established second language acquisition theories. Krashen's Input Hypothesis emphasizes the importance of comprehensible input in lexical development, while Schmidt's Noticing Hypothesis highlights the role of conscious attention to linguistic forms during learning [34,35]. Swain's Output Hypothesis further underscores the value of language production in promoting deeper processing and lexical consolidation [36]. Together, these perspectives provide a rationale for integrating Copilot and Mondly to support comprehensible input, noticing, repeated retrieval, and structured output within classroom practice.

From an applied instructional perspective, the coordinated use of these tools may allow learners to move between explanation, guided production, spoken rehearsal, and feedback within a single classroom routine. Collectively, previous studies suggest promise for generative AI and speech-based mobile tools, yet controlled classroom evidence remains limited regarding whether these complementary technologies can be systematically coordinated within fixed instructional routines to improve productive vocabulary development. This unresolved issue motivates the present study.

### *2.5. Technology Acceptance and Implementation Considerations*

The successful implementation of AI-supported vocabulary learning tools depends not only on their pedagogical design but also on learner acceptance, perceived usefulness, and facilitating technological conditions [37,38]. In applied instructional terms, such tools are relevant not only because they expand access to practice, but also because they can be integrated into replicable classroom routines that support measurable learning outcomes [1,4]. In under-resourced contexts, this applied relevance is particularly important because instructional technologies must function efficiently within existing institutional constraints [12,21].

The long-term instructional viability of AI-assisted learning, however, depends on learner engagement and perceived usefulness, which influence sustained instructional integration [21,39,40]. Technology acceptance models suggest that perceived usefulness, ease of use, and facilitating conditions shape learners' willingness to adopt educational technologies [37,38]. Students' perceptions of accessibility and practical value further influence engagement with mobile learning tools [41], while ethical considerations may affect trust in generative AI technologies [42]. Therefore, examining both learning outcomes and learners' perceptions is essential for understanding the instructional effectiveness and practical viability of integrated AI-supported vocabulary learning in real classroom settings [1,21].

## **3. Materials and Methods**

### *3.1. Research Design*

This study employed a mixed-methods quasi-experimental design to examine the effects of an AI-supported instructional approach integrating Microsoft Copilot (web-based version, accessed via browser during the intervention period; accessed on 9 September 2025) and Mondly (iOS version 10.58.0; Android version 10.27.0; accessed on 9 September 2025) for productive vocabulary development. The research design combined quantitative measurement of vocabulary performance with a qualitative investigation of learner perceptions in order to provide both outcome-based and explanatory evidence regarding the instructional configuration.

The quantitative component adopted a pretest–post-test control group design. Two instructional conditions were compared: (1) a conventional reading–writing instruction condition (control group) and (2) an AI-supported instructional condition integrating Microsoft Copilot and Mondly within regular classroom activities (experimental group). The primary dependent variable was productive vocabulary knowledge measured using the Productive Vocabulary Levels Test (PVLTL).

Because the study was conducted in an authentic classroom context, using intact student groups rather than random assignment, the design is classified as a quasi-experimental pretest–post-test control group design. Baseline equivalence between groups was examined through pretest performance. Instructional conditions were controlled by maintaining identical syllabus content, instructional time, course materials, and instructor supervision across both groups. Accordingly, the findings should be interpreted as reflecting associations between the instructional conditions and observed learning outcomes rather than strict causal effects.

The intervention evaluated the combined effect of Microsoft Copilot and Mondly as an integrated AI-supported instructional configuration for productive vocabulary learning, rather than isolating the individual impact of each tool [3,4]. Because both tools were implemented simultaneously within a structured instructional routine, the study does not allow attribution of observed effects to each component separately. The findings should therefore be interpreted as reflecting the combined influence of the integrated instructional approach.

### 3.2. AI-Supported Instructional Configuration

This study implemented an AI-supported instructional framework consisting of two complementary components: a large language model (LLM)-based interaction module (Microsoft Copilot) and a speech-recognition-supported mobile learning module (Mondly). The configuration was designed to support contextualized vocabulary production, repeated spoken retrieval, and feedback-driven language practice within scheduled classroom activities. In the present study, the term “architecture” refers to a functional instructional design rather than a standalone software platform specifically developed for this research.

From an instructional perspective, the instructional configuration can be understood as a structured multimodal learning workflow linking lexical explanation, contextualized sentence generation, pronunciation-supported rehearsal, and repeated retrieval practice. In this sense, the configuration coordinates learner input, AI-supported processing, and feedback-oriented output refinement within scheduled classroom activities [18,43]. Microsoft Copilot supported lexical explanation, sentence generation, and guided reformulation, while Mondly supported pronunciation rehearsal and repeated spoken retrieval through automated speech recognition. However, the current implementation did not include backend logging or automated capture of user–AI interaction data. Accordingly, the instructional configuration was evaluated at the classroom level through learning outcomes and learner perceptions rather than computational interaction analytics.

During each instructional session, students engaged in a fixed 20 min AI-supported routine consisting of two sequential stages: (1) approximately 10 min of Mondly-based vocabulary practice involving listening and speaking tasks supported by ASR feedback, and (2) approximately 10 min of Copilot-supported productive vocabulary activities, including sentence construction, short dialog generation, and lexical reformulation tasks. This configuration enabled the integration of generative language interaction with speech-based rehearsal within a structured classroom learning cycle. The functional structure of this instructional design is presented in Figure 1, and the specific roles of the two digital components are summarized in Table 1. During classroom implementation, AI-generated

responses were monitored by the instructor. In cases where inaccurate or inappropriate outputs were identified, corrective guidance was provided to ensure that learners did not internalize incorrect lexical or grammatical forms.

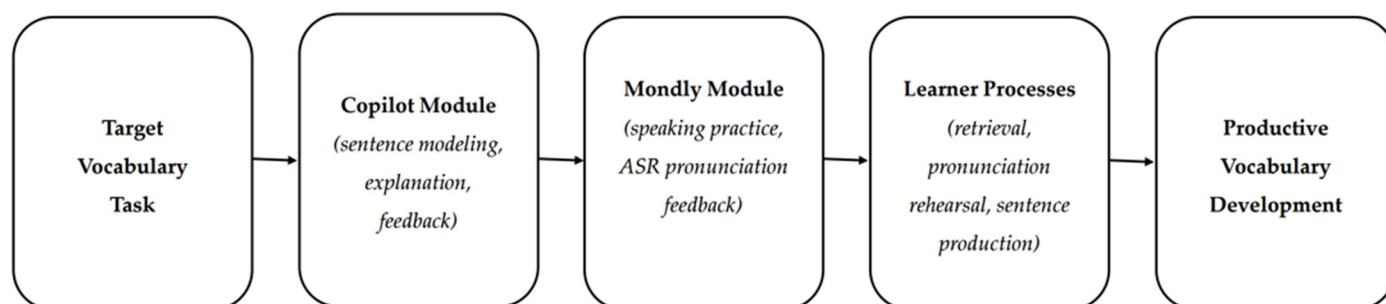


**Figure 1.** Instructional configuration and feedback loop representation of the Copilot–Mondly model for productive vocabulary development.

**Table 1.** Functional roles of the Copilot and Mondly modules within the AI-supported instructional configuration.

| Tool              | Technology Type                   | Main Classroom Function                                             | Targeted Vocabulary Process                                 |
|-------------------|-----------------------------------|---------------------------------------------------------------------|-------------------------------------------------------------|
| Microsoft Copilot | Generative AI/NLP                 | Sentence modeling, explanation, dialog prompts, corrective feedback | Contextualized production, lexical retrieval, reformulation |
| Mondly            | Mobile language learning with ASR | Speaking practice and pronunciation feedback                        | Repeated spoken retrieval, phonological reinforcement       |

Figure 1 presents the instructional configuration as a feedback-driven learning process, illustrating how vocabulary production, feedback, and refinement occur within a structured instructional sequence. From a conceptual perspective, Figure 1 can also be interpreted as an interaction sequence in which the learner, language model, and speech recognition components exchange information through structured feedback cycles. The cyclic structure of this process reflects the recursive nature of learning, in which learner output is reintroduced as input for subsequent stages of processing and refinement. This representation supports a conceptual understanding of how feedback processes and interaction sequences operate within the instructional workflow. In each session, learners first practiced target vocabulary in Mondly through listening, repetition, and speaking recognition tasks. They then transferred selected items to Copilot activities involving sentence creation, short dialogs, lexical reformulation, and correction of self-produced sentences.

Table 1 summarizes the functional roles of the two AI modules within the AI-supported instructional configuration. Together, these components operationalized the instructional configuration as a coordinated multimodal workflow for contextualized production, spoken rehearsal, and feedback-supported retrieval. This modular decomposition clarifies how distinct AI functions were distributed across the instructional workflow, allowing complementary support for lexical generation, spoken rehearsal, and feedback processing during classroom learning.

In operational terms, learners produced spoken or written lexical responses that were followed by AI-supported feedback, reformulation, or pronunciation guidance. Through repeated guided practice, the instructional sequence was intended to support improved lexical accuracy, retrieval fluency, and pronunciation performance within regular classroom instruction.

From an instructional perspective, the proposed configuration can be defined as a coordinated AI-supported instructional framework consisting of three interacting components: (i) learner input (lexical production tasks), (ii) AI-mediated processing (LLM-based

generation and ASR-based feedback), and (iii) output transformation (refined lexical production and pronunciation performance). The process operates through iterative feedback loops in which learner output is continuously evaluated and refined through AI-supported interaction. This formulation enables the instructional process to be understood as a feedback-driven learning sequence rather than as a sequence of isolated tool interactions.

To provide a conceptual representation of the instructional process, the proposed configuration can be expressed as a feedback-driven transformation model:

$$S_{t+1} = f(L_t, A_t, F_t) \quad (1)$$

where  $L_t$  represents learner input (lexical production),  $A_t$  represents AI-supported processing (LLM generation and ASR feedback),  $F_t$  represents feedback-oriented refinement processes at time  $t$ , and  $S_{t+1}$  represents the updated learner state after interaction. This formulation is heuristic and illustrative rather than an empirically estimated computational model. It is intended to conceptually represent how learning may develop through repeated cycles of input, feedback, and refinement during classroom practice.

### 3.3. Participants and Context

The study was conducted in the Department of English Language Teaching (ELT) at Knowledge University in Erbil, Kurdistan Region of Iraq. Participants were first-year undergraduate EFL students enrolled in a reading–writing skills course. This instructional context represents a typical Iraqi university EFL environment in which opportunities for sustained productive language use during classroom instruction are often limited.

All eligible students enrolled in the course were invited to participate in the study. At the beginning of the semester, the instructor explained the research objectives, study procedures, and ethical safeguards. Students were informed that participation was voluntary, that non-participation would not affect their academic standing or course grades, and that they could withdraw at any stage without penalty. Written informed consent was obtained prior to data collection.

A total of 64 first-year undergraduate EFL students (36 females and 28 males) from the ELT Department participated in the study after providing informed consent. Participants were organized into two intact groups formed by the department at the beginning of the academic year. Group assignment was determined by departmental scheduling rather than researcher allocation. One group served as the control group ( $n = 31$ ), receiving traditional reading–writing instruction, while the other served as the experimental group ( $n = 33$ ), receiving traditional reading–writing instruction integrated with Microsoft Copilot and Mondly within regular classroom time. Both groups were taught by the same instructor and followed the same syllabus, course schedule, and instructional materials to ensure instructional consistency. Participants' English proficiency ranged from beginner to pre-intermediate, reflecting early-stage language development and emerging productive lexical competence in the first year of the ELT program. The sample reflected the intact cohort available within the department and is consistent with classroom-based quasi-experimental research in EFL settings.

### 3.4. Measurement Instruments

#### 3.4.1. AI Learning Components

The experimental group used two AI-supported digital tools, Microsoft Copilot and Mondly, as integrated components of the instructional configuration implemented during the intervention. Mondly functioned as a mobile-assisted language learning (MALL) platform providing structured vocabulary lessons and gamified practice activities designed to promote repeated lexical retrieval and active use of target items. These activities included

short interactive tasks, such as listening- and speaking-oriented exercises supported by automated speech recognition (ASR) feedback, which facilitated pronunciation-focused rehearsal and productive vocabulary use.

In parallel, Microsoft Copilot was employed as a generative AI assistant to facilitate productive vocabulary development through interactive scaffolding, including providing lexical explanations, generating contextualized example sentences, offering paraphrasing support, and prompting learners to actively produce newly learned vocabulary in short dialog- and writing-based tasks.

Prior research on generative AI in education indicates that such tools may enhance learner engagement and provide responsive instructional support. However, their effectiveness depends on structured pedagogical integration and responsible instructional use [3,4]. The combined use of Mondly and Copilot was therefore designed to provide complementary learning affordances by integrating structured repetition and gamified retrieval practice with contextualized explanation and guided output-oriented activities that promote active lexical production.

#### 3.4.2. Productive Vocabulary Levels Test

The primary quantitative measurement instrument was the Productive Vocabulary Levels Test (PVLTL), originally developed by Laufer and Nation [17]. The study employed the 36-item version consisting of 18 items from the 2000-word frequency level and 18 items from the 3000-word frequency level.

The test was administered in paper-based format under standardized classroom conditions during a regular 60 min class session. Although the full session time was allocated to minimize time pressure and ensure consistent testing conditions, most students completed the PVLTL within approximately 25–35 min. The same test form was administered at both pretest and post-test under identical procedures to allow direct comparison of productive vocabulary growth over time.

Scores were calculated by summing correct responses across the 36 items to obtain a total productive vocabulary score for each participant. Each correctly completed target word received one point. No partial credit was awarded for incorrect spelling, incomplete forms, or semantically inappropriate responses. The PVLTL is widely recognized as a valid and reliable measure of productive vocabulary knowledge and has been extensively used in EFL research [9,17].

#### 3.4.3. Semi-Structured Interview Protocol

Qualitative data were collected through semi-structured interviews conducted after completion of the intervention. The interview protocol explored learners' perceptions of using Microsoft Copilot and Mondly for productive vocabulary development, including perceived benefits, challenges, engagement patterns, and the extent to which the tools supported lexical retrieval, accurate word production, and active use of target vocabulary in speaking and writing.

A purposive subsample of nine students was selected to represent high-, moderate-, and low-improvement profiles based on gain-score patterns in PVLTL performance. Interviewees were selected using ranked PVLTL gain scores: the upper third represented high-improvement learners, the middle third moderate-improvement learners, and the lower third low-improvement learners. This sampling strategy ensured the inclusion of diverse perspectives regarding the intervention's impact on productive vocabulary development. Because the qualitative phase aimed to provide explanatory insight rather than exhaustive thematic coverage, formal saturation was not claimed.

Each interview lasted approximately 10–15 min and was conducted in Kurdish to allow participants to express their views fully and accurately. Interviews were audio-recorded with participants' permission and transcribed verbatim by the researcher. The transcripts were subsequently translated into English for analysis. To enhance translation accuracy and ensure semantic equivalence, an independent bilingual instructor from the department, who was not part of the research team, reviewed the translated transcripts and cross-checked them against the original Kurdish versions. Discrepancies were resolved through discussion. All transcripts were anonymized prior to analysis to ensure participant confidentiality.

### 3.5. Intervention Procedure

The study was implemented over an 11-week semester (Week 1: orientation to course procedures and administration of the pretest; Weeks 2–10: intervention; Week 11: post-test and interviews). During Week 1, all participants were introduced to the study procedures. Following the pretest, students in the experimental group received structured guidance on how to use Mondly for pronunciation-supported vocabulary practice and Microsoft Copilot for output-oriented vocabulary tasks. Clear task instructions and sample prompts were provided to ensure consistent and standardized use during scheduled class sessions.

The PVLIT was administered to both groups in Week 1 as a pretest under standardized classroom testing conditions. The treatment phase was implemented during Weeks 2–10. The course consisted of four sessions per week, each lasting approximately one hour. In the experimental group, a standardized 20 min AI-supported routine was implemented exclusively during scheduled class time in each session. The routine consisted of 10 min of Mondly practice followed by 10 min of Copilot-supported productive vocabulary activities. This routine replaced an equivalent portion of conventional vocabulary-related practice activities, thereby maintaining identical total instructional time across both groups.

To control instructional dosage and minimize uncontrolled exposure, students in the experimental group were instructed to use Mondly and Copilot exclusively during the structured in-class routine. No out-of-class AI practice was assigned, required, or formally encouraged. During the Copilot component, students completed guided output-oriented tasks, including requesting brief lexical explanations, generating contextualized example sentences, producing short dialogs, and revising their own sentences using target vocabulary. Typical Copilot prompts were standardized and aligned with weekly vocabulary targets. Examples included "Explain this word in simple English," "Use this word in a meaningful sentence," "Write a short dialogue using this word," "Give another sentence with the same word in a different context," and "Correct my sentence using this word." In many cases, students first practiced pronunciation and recognition of target vocabulary items in Mondly. Then they transferred the same items to Copilot for sentence construction, dialogue generation, or feedback-supported reformulation. This sequencing was intended to connect spoken retrieval practice with contextualized productive use within the scheduled instructional sequence. The Mondly component focused on structured vocabulary rehearsal through brief listening- and speaking-oriented tasks supported by automated speech recognition (ASR) feedback to promote accurate pronunciation and repeated lexical retrieval.

In contrast, the control group received traditional reading–writing instruction, including textbook-based activities, teacher-led explanation, and in-class vocabulary practice, without the use of AI tools. In Week 11, the PVLIT was administered again as a post-test under identical standardized conditions. Following the post-test, semi-structured interviews were conducted with a purposive subsample of participants from the experimental group only. The interviews aimed to explore learners' perceptions of the structured in-class

AI routine and to provide explanatory insight into the quantitative findings. Because the control group did not receive the AI-supported intervention, qualitative data collection appropriately focused on students who participated in the Copilot–Mondly condition.

To support fidelity of implementation, the instructor followed a fixed routine each session and used the same task prompts and timing sequence across the intervention weeks.

### 3.6. Data Collection

Data collection followed the explanatory sequential mixed-methods procedure. Quantitative data consisted of the following:

- (a) Pretest PVLТ scores;
- (b) Post-test PVLТ scores;
- (c) Gain scores calculated as post-test minus pretest for descriptive comparison and qualitative sampling purposes.

Gain scores were used for descriptive comparisons and to inform purposive sampling for the qualitative phase. In contrast, inferential analyses were conducted using pretest and post-test scores within a mixed-design ANOVA framework.

Qualitative data consisted of semi-structured interview transcripts collected from a purposive subsample of nine students in the experimental group. The interviews explored learners' experiences with Copilot and Mondly, perceived learning benefits, usability challenges, and engagement-related factors influencing productive vocabulary development and lexical retrieval. These qualitative data were used to provide explanatory depth and contextual interpretation of the quantitative findings.

All participants provided written informed consent prior to data collection. Pretest and post-test PVLТ data were collected during scheduled class sessions under standardized administration procedures. Interview recordings were transcribed verbatim, anonymized, and reviewed for accuracy prior to analysis. Participant identifiers were replaced with coded labels, and all data were handled confidentially in accordance with institutional ethical guidelines.

### 3.7. Data Analysis

Quantitative data were analyzed using a two-way mixed-design analysis of variance (ANOVA) with Time (pretest, post-test) as the within-subjects factor and Group (experimental, control) as the between-subjects factor. Analyses were conducted using SPSS (Version 25). This analytic approach enabled simultaneous examination of within-subject change over time and between-group differences across instructional conditions. The dependent variable was the PVLТ productive vocabulary score.

The analysis evaluated (a) the main effect of Time, indicating whether productive vocabulary scores changed significantly from pretest to post-test across participants; (b) the main effect of Group, indicating whether overall performance differed between instructional conditions when averaged across time; and (c) the Time  $\times$  Group interaction, indicating whether the magnitude of change differed between the experimental and control groups.

Assumptions were examined prior to analysis. Levene's tests were conducted to assess the homogeneity of variance at each time point. Because the within-subject factor included only two levels, the assumption of sphericity was not applicable. Effect sizes were reported using partial eta squared (partial  $\eta^2$ ), and statistical significance was evaluated at an alpha level of 0.05.

Qualitative interview data were analyzed using thematic analysis following Braun and Clarke's [44] six-phase framework: familiarization with the data, generation of initial codes, theme development, theme review, theme refinement, and reporting. Two researchers

independently coded the transcripts to enhance credibility and reduce subjective bias. Coding discrepancies were resolved through discussion until consensus was reached. Inter-coder reliability was assessed using Cohen's kappa ( $\kappa = 0.81$ ), indicating almost perfect agreement and supporting the trustworthiness of the analysis.

Finally, qualitative findings were integrated with quantitative results during the interpretation stage to generate meta-inferences and provide explanatory insight into factors that may have contributed to the observed productive vocabulary gains in the AI-supported instructional condition.

#### Exploratory Implementation Indicators

In addition to the primary educational outcome analyses, several exploratory indicators were considered to describe implementation-level characteristics of the coordinated instructional system within the classroom context. These indicators are descriptive implementation-oriented summaries rather than formal engineering performance metrics. Their purpose is to provide an applied systems perspective on how the instructional configuration functioned under authentic classroom conditions.

- (1) Learning Gain: Pretest–post-test improvement in PVLTL scores across conditions;
- (2) Differential Improvement: Relative advantage of the coordinated AI-supported condition compared with the conventional condition;
- (3) Instructional Time Efficiency: Improvement achieved without increasing total scheduled instructional time;
- (4) Implementation Feasibility: Learner perceptions regarding usability and classroom practicality.

These indicators should be interpreted as contextual implementation summaries rather than validated system-performance measures. Future research may incorporate richer systems-level evidence such as usage logs, response latency, interaction frequency, and component-level analytics to support more rigorous evaluation of human–AI instructional systems.

#### 3.8. Ethical Considerations

Ethical approval was obtained in accordance with the research ethics procedures of Knowledge University. All participants provided written informed consent prior to data collection. Participation was voluntary, and students were assured that their academic standing would not be affected. Confidentiality was maintained through anonymization of data and secure storage on password-protected devices accessible only to the researchers. Because the study involved the use of generative AI, students were instructed not to share personal or sensitive information when interacting with Microsoft Copilot, consistent with current ethical recommendations for the responsible educational use of generative AI tools [4].

#### 3.9. Reliability and Validity

Quantitative validity was supported through the use of the Productive Vocabulary Levels Test (PVLTL), a widely used and previously validated measure of controlled productive vocabulary knowledge across frequency levels [9,17]. Standardized administration procedures were applied at both pretest and post-test to ensure measurement consistency. Internal validity was strengthened by maintaining identical syllabus content, instructional schedule, instructor supervision, and total instructional time across both groups. Baseline equivalence was established through an independent samples t-test analysis of pretest PVLTL scores, indicating no statistically significant differences between groups prior to the intervention.

Using the same PVLТ form at pretest and post-test represents a methodological limitation, as it may introduce test familiarity effects. Although the nine-week interval reduces the likelihood of direct recall, practice effects cannot be fully ruled out and should be considered when interpreting the results. In addition, the intervention emphasized classroom-embedded, contextually relevant production and retrieval practice rather than the rehearsal of specific test items, making it unlikely that the gains reflect memorization of test content.

Qualitative trustworthiness was enhanced through purposive sampling of participants representing varied PVLТ gain profiles, verbatim transcription of interviews, and systematic application of thematic analysis procedures. Two researchers independently coded the transcripts, and discrepancies were resolved through discussion until consensus was achieved. Inter-coder reliability was assessed using Cohen's kappa ( $\kappa = 0.81$ ), indicating almost perfect agreement and supporting analytic consistency. Integration of qualitative and quantitative findings followed an explanatory sequential mixed-methods design, allowing interview data to inform interpretations of factors associated with the observed productive vocabulary gains.

## 4. Results

### 4.1. Baseline Equivalence

An independent samples t-test was conducted to determine whether the experimental and control groups differed at pretest on the Productive Vocabulary Levels Test (PVLТ). The analysis indicated no statistically significant difference between the experimental group ( $M = 12.48$ ,  $SD = 4.15$ ,  $n = 33$ ) and the control group ( $M = 12.87$ ,  $SD = 4.60$ ,  $n = 31$ ),  $t(62) = -0.35$ ,  $p = 0.726$ , 95% CI  $[-2.57, 1.80]$ . These findings indicate that the two groups were statistically comparable at baseline. The absence of a significant baseline difference supports the internal validity of the quasi-experimental comparison.

### 4.2. Descriptive Statistics

Table 2 presents descriptive statistics for PVLТ scores across time and instructional condition.

**Table 2.** Descriptive statistics for PVLТ scores by group and time.

| Group        | Time      | N  | Mean  | SD   |
|--------------|-----------|----|-------|------|
| Experimental | Pretest   | 33 | 12.48 | 4.15 |
| Experimental | Post-test | 33 | 14.64 | 4.30 |
| Control      | Pretest   | 31 | 12.87 | 4.60 |
| Control      | Post-test | 31 | 13.61 | 4.36 |

These descriptive results suggest greater improvement in productive vocabulary scores for the experimental group compared with the control group, which was subsequently tested through inferential analysis.

Both groups demonstrated improvement in productive vocabulary scores between the pretest and post-test. The experimental group demonstrated a mean increase of 2.16 points (14.64–12.48), whereas the control group demonstrated a mean increase of 0.74 points (13.61–12.87). Although both groups improved over time, the magnitude of change was descriptively larger in the experimental group than in the control group. To further describe patterns of improvement, additional descriptive indicators were calculated.

### Descriptive Indicators of Learning Improvement

To provide additional descriptive context for interpreting score changes, several simple summary indicators were calculated from the group means. These indicators are descriptive summaries rather than inferential or validated performance metrics. Learning gain (LG) was calculated as the difference between post-test and pretest scores. The learning efficiency index (LEI) was defined as the learning gain normalized by instructional duration (in weeks), representing the rate of vocabulary improvement over time. In addition, relative improvement (RI) was calculated as the percentage increase from baseline performance.

These indicators can be formally expressed as follows:

$$LG = \text{Post-test} - \text{Pretest}$$

$$LEI = LG / \text{Duration}$$

$$RI (\%) = (LG / \text{Pretest}) \times 100$$

As shown in Table 3, the experimental group demonstrated a substantially higher learning gain (2.16) compared with the control group (0.74). When normalized over the instructional duration, the experimental group achieved a higher learning efficiency index ( $LEI = 0.20$ ) than the control group ( $LEI = 0.07$ ), indicating greater vocabulary development per unit of instructional time. Similarly, relative improvement was markedly higher in the experimental group (17.31%) than in the control group (5.75%).

**Table 3.** Descriptive indicators derived from PVLTL scores.

| Group        | Pretest | Post-Test | Gain | LEI  | RI (%) |
|--------------|---------|-----------|------|------|--------|
| Experimental | 12.48   | 14.64     | 2.16 | 0.20 | 17.31  |
| Control      | 12.87   | 13.61     | 0.74 | 0.07 | 5.75   |

These results indicate that the experimental group achieved greater gains in productive vocabulary scores within the same instructional time.

#### 4.3. Mixed ANOVA

A two-way mixed-design ANOVA was conducted with Time (pretest, post-test) as the within-subjects factor and Group (experimental, control) as the between-subjects factor. Results are presented in Table 4. Levene's tests supported homogeneity of variance at both pretest and post-test. Because the within-subject factor included only two levels, the assumption of sphericity was not applicable.

**Table 4.** Mixed-design ANOVA results for PVLTL scores.

| Source              | F     | df      | p      | Partial $\eta^2$ |
|---------------------|-------|---------|--------|------------------|
| Time                | 39.19 | (1, 62) | <0.001 | 0.384            |
| Group               | 0.07  | (1, 62) | 0.788  | 0.001            |
| Time $\times$ Group | 10.35 | (1, 62) | 0.002  | 0.143            |

A statistically significant main effect of Time was observed,  $F(1, 62) = 39.19$ ,  $p < 0.001$ , partial  $\eta^2 = 0.384$ , indicating that productive vocabulary scores increased significantly from pretest to post-test across participants. The effect size suggests a large magnitude of change over time.

The main effect of Group was not statistically significant,  $F(1, 62) = 0.07$ ,  $p = 0.788$ , partial  $\eta^2 = 0.001$ , indicating no overall difference between instructional conditions when averaged across time.

A statistically significant Time  $\times$  Group interaction was found,  $F(1, 62) = 10.35$ ,  $p = 0.002$ , partial  $\eta^2 = 0.143$ , indicating that the magnitude of improvement differed between instructional conditions. Specifically, the experimental group demonstrated greater pre–post gains in productive vocabulary performance than the control group. The interaction effect (partial  $\eta^2 = 0.143$ ) indicates a comparatively substantial difference in productive vocabulary gains between instructional conditions.

#### 4.4. Qualitative Findings

Table 5 summarizes the major themes and representative codes derived from the semi-structured interviews. The analysis focuses specifically on learners' perceptions of productive vocabulary development, including lexical retrieval, sentence construction, spoken production, and contextualized word use.

**Table 5.** Summary of qualitative themes and representative codes identified through thematic analysis of student interviews.

| Theme                                                                                      | Representative Codes                                                                              |
|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Theme 1: Perceived Effectiveness of AI-Supported Vocabulary Learning                       | Example generation, sentence modeling, corrective feedback, dialog creation, contextual variation |
| Theme 2: Reinforced Retention Through Structured Repetition and Pronunciation Practice     | Repetition, pronunciation rehearsal, ASR feedback, structured practice, speaking tasks            |
| Theme 3: Increased Engagement and Active Participation in Productive Vocabulary Practice   | Immediate feedback, guided writing, speaking confidence, active participation                     |
| Theme 4: Perceived Accessibility and Practical Classroom Use in an Under-Resourced Context | Large-class suitability, equal access, structured AI support                                      |

##### 4.4.1. Theme 1: Perceived Effectiveness of AI-Supported Vocabulary Learning

This theme reflects how the integration of Microsoft Copilot and Mondly supported learners in actively producing vocabulary in speaking and writing rather than merely recognizing meanings. Students emphasized that the tools helped them construct sentences, generate dialogs, and apply new words in meaningful contexts during the structured classroom routine. These perceptions suggest that students moved from passive lexical recognition toward active production and contextualized word use.

Participants reported that Copilot modeled sentence construction and provided immediate corrective feedback, enabling them to practice productive use of newly learned words:

“Copilot showed me how to use the word in a sentence, so I could write my own sentence.” (S1)

“It gave simple explanations and examples, and then I tried to make similar sentences.” (S2)

“Sometimes I asked Copilot to create a short dialogue using the new word, and I practiced it.” (S6)

Several students described how corrective feedback strengthened productive accuracy:

“It corrected my sentence when I used the word incorrectly.” (S7)

Learners also reported improved contextual flexibility in word use:

“I liked asking Copilot to help me use a word in different sentences.” (S7)

These findings suggest that AI-mediated scaffolding supported guided lexical retrieval and contextualized word production within classroom instruction. Rather than memorizing isolated items, students engaged in structured sentence generation and dialog practice, which is consistent with the greater pre–post gains observed in the experimental group in the PVLt analysis.

#### 4.4.2. Theme 2: Reinforced Retention Through Structured Repetition and Pronunciation Practice

This theme captures how Mondly reinforced productive vocabulary development through repeated spoken production and pronunciation-supported practice during the structured in-class routine. Students emphasized that actively producing words aloud enhanced both confidence and the efficiency of lexical retrieval in subsequent speaking and writing tasks. These observations suggest that repeated spoken retrieval and pronunciation practice supported more efficient lexical access during productive language use.

Participants frequently referred to repetition as central to remembering new words:

“Repeating the word many times helped me use it more confidently.” (S1)

“When I repeated and practiced speaking, I could use the word more easily later.” (S2)

Pronunciation practice supported phonological accuracy and production confidence:

“The pronunciation practice helped me say words correctly.” (S3)

“Hearing and practicing the word together helped me use it correctly.” (S5)

Students also valued the structured and time-efficient format of the routine:

“Lessons were short, so I could practice the app every day in class.” (S4)

Some students acknowledged minor technological limitations:

“Sometimes the app did not understand my pronunciation, but I tried again.” (S6)

Overall, repeated oral production appeared to support more efficient access to vocabulary knowledge during productive language use. Through structured retrieval and pronunciation rehearsal within scheduled sessions, learners strengthened the connection between meaning, form, and spoken output. These learner perceptions are broadly consistent with the greater pre–post gains observed in the experimental group.

Although overall perceptions were positive, some students also identified minor challenges during implementation. Several participants noted that speech recognition feedback was occasionally inconsistent, requiring repeated attempts before responses were accepted. Others mentioned that unstable internet connectivity sometimes interrupted the smooth use of the applications. A small number of learners also indicated that brief teacher clarification remained helpful when encountering unfamiliar vocabulary or more complex meanings. These comments suggest that the benefits of the AI-supported routine were influenced by technological reliability and continued pedagogical guidance.

#### 4.4.3. Theme 3: Increased Engagement and Active Participation in Productive Vocabulary Practice

This theme reflects how the combined use of Copilot and Mondly increased students’ engagement and active participation in the instructional configuration’s interaction process during classroom instruction. Participants reported that the tools enabled them to practice speaking and writing more intensively within the structured AI routine.

Examples include the following:

“Yes, because Mondly helped me practice speaking and Copilot helped me use the word in sentences.” (S1)

“I practiced in Mondly and then used the word in Copilot writing tasks.” (S2)

Students highlighted the shift from passive recognition to active production:

“The apps helped me not only know the word but also use it correctly.” (S3)

“It was better than only reading from the textbook because I produced sentences.” (S4)

Increased confidence in classroom participation was also noted:

“I felt more confident using vocabulary in class after using the apps.” (S8)

These perceptions indicate that AI-supported tools increased structured opportunities for speaking and writing practice during instructional time. By expanding guided output practice within scheduled sessions, the tools may have supported productive vocabulary development in the experimental condition.

#### 4.4.4. Theme 4: Perceived Accessibility and Practical Classroom Use in an Under-Resourced Context

This theme reflects students' perceptions of the broader applicability and instructional efficiency of integrating AI-supported vocabulary instruction within their university context. Participants emphasized that the structured Copilot–Mondly routine could be effectively implemented in large classes and that multiple students could engage simultaneously during scheduled sessions.

Students highlighted the model's perceived flexibility and accessibility:

"Yes, because students can practice speaking and writing within the apps." (S1)

"It helps students use vocabulary through the app during classroom activities." (S2)

Although the intervention in the present study was implemented exclusively during structured in-class sessions, these comments reflect learners' perceptions of the potential flexibility of AI-supported vocabulary practice rather than documented out-of-class use. Participants appeared to interpret the AI tools as adaptable resources that could, in principle, extend opportunities for productive engagement.

Students also recognized the suitability of the model for large-class contexts:

"It is good for large classes; all the students can ask at the same time." (S7)

"Many students can use it at the same time." (S3)

In addition, learners emphasized perceived independence and continuity of learning:

"It helps students become more independent in using vocabulary." (S6)

"Students can continue practicing speaking and writing after the course." (S8)

These comments suggest that students viewed the AI-supported routine as a structured mechanism that reduced reliance on immediate teacher mediation during classroom activities, while still operating within formal instructional guidance.

Participants also acknowledged the importance of sustained engagement:

"It is helpful if students use it regularly." (S9)

Overall, these findings indicate that learners perceived the AI-supported routine as accessible and instructionally manageable within the present context. The structured 20 min AI routine increased opportunities for productive vocabulary practice during scheduled classroom sessions within existing instructional time. While broader validation is required, these perceptions suggest potential applicability in comparable resource-constrained university EFL settings.

## 5. Discussion

The observed differences between instructional conditions may relate not only to the presence of AI tools but also to how their functions were organized within a structured classroom routine. In the present intervention, Copilot supported lexical modeling, sentence construction, and feedback, while Mondly supported pronunciation rehearsal and repeated spoken retrieval. These complementary functions may have increased opportunities for guided production and feedback during scheduled instructional time. This interpretation is consistent with previous research suggesting that the educational value of digital tools depends not only on technological availability but also on pedagogical integration within classroom practice [33].

The present study primarily assessed outcome-level learning rather than detailed interaction processes. Although the Copilot–Mondly routine was conceptually organized as a feedback-oriented instructional sequence, the absence of interaction-level data limits analysis of how specific patterns of use were associated with learning outcomes. Future research should incorporate log-based evidence, such as prompt–response sequences and usage frequency, to enable more fine-grained analysis of AI-supported classroom processes.

This study evaluated an AI-supported instructional routine integrating Microsoft Copilot and Mondly to support productive vocabulary development in an Iraqi university EFL context. Although both groups improved over time, the experimental group demonstrated significantly greater gains, as reflected in the statistically significant Time  $\times$  Group interaction (partial  $\eta^2 = 0.143$ ). Because baseline equivalence was established and total instructional time was maintained across conditions, the findings suggest that the integrated Copilot–Mondly routine provided added instructional value beyond traditional reading–writing instruction. In the present context, the approach may offer a practical way to increase feedback opportunities and productive practice during scheduled classroom time [1,3].

The quantitative results indicate that the observed improvement should not be interpreted as a general time effect, but as potentially related to increased opportunities for active lexical retrieval and contextualized word production during classroom instruction. The AI-supported routine may have supported productive vocabulary learning through immediate corrective feedback, repeated spoken retrieval, pronunciation-focused rehearsal, and guided sentence construction within existing instructional time [1,2].

Interaction with Copilot appeared to provide learners with opportunities to generate original sentences, reformulate expressions, and apply target vocabulary in varied contexts. Rather than reproducing fixed textbook responses, students engaged in guided lexical production supported by AI-mediated prompts and feedback. These perceptions suggest that the intervention may have supported productive vocabulary development through contextualized output practice combined with immediate instructional scaffolding [10]. The findings are broadly consistent with the potential value of integrating generative AI into structured classroom routines for language learning [1,3].

Similarly, findings related to Mondly highlight the possible value of structured repetition combined with spoken production and pronunciation-focused retrieval [10,31]. Students emphasized that hearing, repeating, and actively producing target words increased confidence and ease of later use. This pattern is consistent with retrieval-based learning principles and suggests that repeated output practice may have supported more efficient lexical access during production. Qualitative data also indicated that superficial engagement reduced perceived benefit, suggesting that learning outcomes may depend not only on exposure frequency but also on depth of engagement during output-oriented tasks.

Another important dimension was increased engagement and active participation during structured classroom activities [45]. Students reported greater involvement in speaking and writing tasks than during traditional instruction, particularly because the AI-supported routine enabled immediate practice and feedback within scheduled sessions. In contexts where teacher time may be limited, structured AI mediation may help extend feedback opportunities within instructional time [1,3].

The findings support theoretical perspectives emphasizing comprehensible input, noticing, output, and retrieval practice in productive vocabulary development [10,34–36]. Copilot-supported modeling and contextualized explanations may have provided meaningful input, while corrective feedback created opportunities for pushed output. Mondly’s structured speaking tasks may likewise have supported repeated lexical retrieval and phonological reinforcement. The coordinated use of these affordances is broadly consistent with the observed gains. Rather than relying solely on lexical exposure, the intervention provided guided production, feedback, rehearsal, and retrieval opportunities embedded within classroom instruction.

From an applied instructional technology perspective, the AI-supported routine may offer a practical way to improve feedback availability and productive practice within resource-constrained EFL classrooms [1,3]. Participants perceived the model as applicable

in large-class settings, where consistently providing individualized feedback is difficult. By increasing structured output practice within existing instructional time and redistributing some feedback processes through guided human–AI interaction, the intervention illustrates how complementary AI-supported tools can extend classroom practice opportunities. Participants also identified infrastructural constraints such as internet connectivity, indicating that effective implementation depends on technological readiness and institutional support.

Future implementations of the proposed instructional configuration may incorporate interaction-level data collection and analysis of learner–AI engagement patterns. Such approaches could support a more detailed understanding of how instructional processes contribute to learning outcomes and inform the refinement of classroom-based AI-supported practices.

#### *Instructional Design Implications*

From an instructional perspective, the Copilot–Mondly configuration indicates how complementary AI modules can be integrated into a coordinated instructional design. The design combines generative language modeling for contextualized lexical production with automated speech recognition for pronunciation-supported retrieval practice. When embedded within structured classroom workflows, such multimodal AI configurations may support wider practical classroom use by extending feedback availability and production opportunities in large-class environments.

From a broader instructional perspective, the proposed design may be relevant beyond vocabulary learning in contexts that require iterative skill development. Elements of the routine, such as AI-supported explanation, guided production, feedback, and repeated practice, could potentially be adapted for areas such as writing development, pronunciation training, or domain-specific language learning. However, further empirical validation across contexts would be required before broader generalization can be made.

### **6. Practical Implementation Implications for AI-Supported Vocabulary Instruction**

AI-supported instructional approaches for vocabulary learning may be more beneficial when systematically embedded within structured classroom instruction rather than used as optional technological supplements [40]. In the present study, the combination of generative AI-mediated sentence modeling and speech-based retrieval practice was associated with stronger productive vocabulary gains during regular classroom activities. From an applied instructional perspective, the model may offer a practical way to increase feedback opportunities, output practice, and learner engagement without extending total teaching time [1,3]. However, effective implementation depends on reliable infrastructure, digital literacy support, and appropriate pedagogical supervision [12].

The findings also suggest several tentative design principles for AI-supported vocabulary learning. Effective implementation may benefit from coordination of complementary AI functions, structured sequencing of learner–AI interaction, and feasible integration within existing classroom constraints. In particular, combining generative language models with speech recognition technologies may create learning workflows that support contextualized lexical generation alongside pronunciation-focused retrieval practice.

From a design perspective, the present findings suggest that AI-supported vocabulary instruction may benefit from three practical features: modularity, repeatability, and implementation feasibility. Modularity allows complementary tools to perform distinct instructional functions; repeatability supports consistent integration within scheduled classroom time; and implementation feasibility helps ensure that the approach can operate under realistic conditions such as device access, internet stability, and teacher supervision.

Overall, the integrated Copilot–Mondly configuration can be viewed as a classroom-embedded instructional routine that uses guided AI-supported interaction to increase productive vocabulary practice within existing institutional constraints. When supported by structured pedagogical design and reliable technological infrastructure, such AI-assisted routines may have practical relevance for comparable higher-education settings [1,2].

More broadly, this configuration illustrates how AI may function within a classroom socio-technical environment in which human instruction, digital mediation, and learner engagement are coordinated through structured pedagogical routines [4,18]. Rather than replacing teachers, AI in this context appeared to extend feedback opportunities within the learning environment, forming a hybrid pedagogical approach. Such integration reflects classroom practice in which digital tools are embedded within ongoing instruction rather than adopted as isolated technological resources [1].

## 7. Conclusions

This study examined an AI-supported instructional approach integrating a generative AI assistant (Microsoft Copilot) with a speech-based mobile learning application (Mondly) to support productive vocabulary development in an Iraqi university EFL context. Although productive vocabulary scores improved across both instructional conditions, students who participated in the structured Copilot–Mondly routine demonstrated significantly greater gains than those receiving traditional instruction alone. These findings suggest that integrating AI-mediated contextually informed sentence modeling and corrective feedback with structured retrieval-based speaking practice may support measurable improvements in productive lexical performance within existing instructional time. Accordingly, the study contributes pedagogical evidence on productive vocabulary learning and offers a structured example of how LLM- and ASR-based technologies can be coordinated within classroom-based AI-supported instruction.

Qualitative evidence further helped interpret factors associated with this improvement. Learners perceived Copilot as supporting accurate word production through iterative explanation, sentence modeling, and dialog-based scaffolding. In contrast, Mondly was perceived as reinforcing productive control through repeated spoken retrieval, pronunciation-focused rehearsal, and structured classroom practice. Increased engagement and active participation in speaking and writing tasks during scheduled sessions also emerged as potentially important facilitating factors. At the same time, infrastructural constraints such as internet connectivity highlighted the importance of context-sensitive, technology-supported implementation.

From an applied instructional technology perspective, the findings suggest that complementary AI tools can be implemented as a structured classroom routine to improve feedback availability and opportunities for productive output within existing instructional time. However, effective implementation depends on reliable infrastructure, pedagogical guidance, and responsible use of generative AI. Overall, the study suggests that coordinated use of complementary AI tools may enhance productive vocabulary learning when embedded within structured pedagogical routines rather than used as isolated technological supplements. While further replication across contexts is needed, the present findings indicate possible practical value for higher-education EFL settings.

## 8. Limitations

Several limitations should be considered when interpreting the findings. First, this study employed a quasi-experimental design using intact groups rather than random assignment. Although baseline equivalence was established and key instructional variables were maintained across groups, the absence of randomization limits the strength of causal

inference. While the statistically significant Time  $\times$  Group interaction provides evidence of differential improvement between instructional conditions, randomized experimental designs would offer stronger confirmation of causal effects.

Second, productive vocabulary development was assessed using the same PVLТ form at pretest and post-test. Although the instrument has been widely validated for measuring controlled productive lexical knowledge, repeated administration may introduce familiarity effects. Moreover, the PVLТ assesses productive recall primarily at the word level and does not fully capture extended, discourse-level vocabulary use in authentic communicative tasks. Future studies should therefore incorporate delayed post-tests, discourse-based production measures, and performance-oriented assessments to examine long-term retention and transfer to communicative contexts.

Third, the study evaluated instructional effectiveness without collecting detailed interaction-level data on learner–AI engagement. For example, prompt histories, response patterns, and usage frequency were not systematically recorded. This limits the ability to examine how specific interaction patterns were associated with learning outcomes or to identify the relative contribution of different instructional components. Future research should incorporate learning analytics and interaction-level data to enable more fine-grained analysis of instructional processes and to better understand how AI-supported activities influence vocabulary development over time.

Fourth, the intervention examined the combined implementation of Microsoft Copilot and Mondly. While this integrated design reflects authentic classroom practice, it prevents the isolation of the independent contribution of each tool. Comparative studies including multiple experimental conditions would allow clearer attribution of effects to specific components, such as generative AI-mediated feedback versus pronunciation-supported retrieval practice.

Finally, the qualitative findings were based on a relatively small purposive subsample within a single institutional setting, which may limit transferability to contexts with different learner profiles, levels of digital literacy, or technological infrastructures. The practical implementation of AI-supported productive vocabulary instruction also depends on facilitating conditions such as reliable internet connectivity and access to appropriate devices. Where such conditions are unstable, implementation barriers may moderate learning outcomes. Future studies should therefore examine how contextual, infrastructural, and institutional variables influence the effectiveness and long-term sustainability of AI-supported vocabulary interventions across diverse under-resourced educational environments.

**Author Contributions:** Conceptualization, S.M.H.; methodology, S.M.H.; software, S.M.H.; formal analysis, S.M.H.; investigation, S.M.H. and M.K.; resources, S.M.H.; data curation, S.M.H. and M.K.; writing—original draft preparation, S.M.H. and M.K.; writing—review and editing, S.M.H. and M.K.; supervision, M.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study was conducted in accordance with the Declaration of Helsinki and received ethical approval from Knowledge University Scientific Research Ethics Committee (Approval Code: KNU/IRB/2025/348, Approval Date 8 September 2025).

**Informed Consent Statement:** Written informed consent was obtained from all participants involved in the study. Participation was voluntary, and students were informed that they could withdraw from the study at any time without penalty.

**Data Availability Statement:** The data presented in this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Holmes, W.; Bialik, M.; Fadel, C. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*; Center for Curriculum Redesign: Boston, MA, USA, 2019.
2. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [\[CrossRef\]](#)
3. Godwin-Jones, R. Distributed agency in second language learning and teaching through generative AI. *Lang. Learn. Technol.* **2024**, *28*, 5–30. [\[CrossRef\]](#)
4. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [\[CrossRef\]](#)
5. Nushi, M.; Fattahi, N.; Ebn-Abbasi, F. An evaluative review of Mondly: A mobile language learning application. *EuroCALL Rev.* **2023**, *30*, 69–85. [\[CrossRef\]](#)
6. Reynolds, B.L.; Cui, Y.; Kao, C.-W.; Thomas, N. Vocabulary Acquisition through Viewing Captioned and Subtitled Video: A Scoping Review and Meta-Analysis. *Systems* **2022**, *10*, 133. [\[CrossRef\]](#)
7. Bao, W.; Wang, T.; Zhang, L.; Yusop, F.D.; Ruan, X. A systematic review of AI in second language acquisition using the expanded SAMR model (2015–2024). *Discov. Comput.* **2025**, *28*, 292. [\[CrossRef\]](#)
8. Yan, W.; Li, B.; Lowell, V.L. Integrating artificial intelligence and extended reality in language education: A systematic literature review (2017–2024). *Educ. Sci.* **2025**, *15*, 1066. [\[CrossRef\]](#)
9. Nation, I.S.P. *Learning Vocabulary in Another Language*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013. [\[CrossRef\]](#)
10. Webb, S.; Nation, I.S.P. *How Vocabulary Is Learned*; Oxford University Press: Oxford, UK, 2017.
11. Yang, L.; Chen, S.; Li, J. Enhancing sustainable AI-driven language learning: Location-based vocabulary training for learners of Japanese. *Sustainability* **2025**, *17*, 2592. [\[CrossRef\]](#)
12. UNESCO. *Global Education Monitoring Report 2020: Inclusion and Education—All Means All*; UNESCO Publishing: Paris, France, 2020.
13. Schmitt, N. Instructed second language vocabulary learning. *Lang. Teach. Res.* **2008**, *12*, 329–363. [\[CrossRef\]](#)
14. Zhang, Z.; Huang, X. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon* **2024**, *10*, e25370. [\[CrossRef\]](#)
15. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [\[CrossRef\]](#)
16. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **2022**, *55*, 1–37. [\[CrossRef\]](#)
17. Laufer, B.; Nation, P. A vocabulary-size test of controlled productive ability. *Lang. Test.* **1999**, *16*, 33–51. [\[CrossRef\]](#)
18. Zhang, Y.; Dong, C. Exploring the Digital Transformation of Generative AI-Assisted Foreign Language Education: A Socio-Technical Systems Perspective Based on Mixed-Methods. *Systems* **2024**, *12*, 462. [\[CrossRef\]](#)
19. UNESCO. *AI and Education: Guidance for Policy-Makers*; UNESCO: Paris, France, 2021.
20. Hezam, A.M.M.; Mahyoub, R.A.M. AI-enhanced vocabulary acquisition in conflict-affected contexts: A study of Yemeni EFL learners' adaptations, challenges, and outcomes. *Educ. Linguist. Res.* **2025**, *11*, 64. [\[CrossRef\]](#)
21. Roy, S.S.; Gandhimathi, S.N.S. Self-directed learning for optimizing sustainable language learning via mobile assisted language learning: A systematic review. *Front. Educ.* **2025**, *9*, 1463721. [\[CrossRef\]](#)
22. Kacetyl, J.; Klímová, B. Use of smartphone applications in English language learning—A challenge for foreign language education. *Educ. Sci.* **2019**, *9*, 179. [\[CrossRef\]](#)
23. Ferik Savec, V.; Jedrinović, S. The role of AI implementation in higher education in achieving the sustainable development goals: A case study from Slovenia. *Sustainability* **2025**, *17*, 183. [\[CrossRef\]](#)
24. Stratton, J. An introduction to Microsoft Copilot. In *Copilot for Microsoft 365: Harness the Power of Generative AI in the Microsoft Apps You Use Every Day*; Apress: Berkeley, CA, USA, 2024; pp. 19–35. [\[CrossRef\]](#)
25. Aydın Yıldız, T. The impact of ChatGPT on language learners' motivation. *J. Teach. Educ. Lifelong Learn.* **2023**, *5*, 582–597. [\[CrossRef\]](#)
26. Kohnke, L.; Zou, D.; Su, F. Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100371. [\[CrossRef\]](#)
27. Vnucko, G.; Klimova, B. Exploring the Potential of Digital Game-Based Vocabulary Learning: A Systematic Review. *Systems* **2023**, *11*, 57. [\[CrossRef\]](#)
28. Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digit. Med.* **2023**, *6*, 75. [\[CrossRef\]](#)
29. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 248. [\[CrossRef\]](#)

30. Cho, K.; Lee, S.; Joo, M.-H.; Becker, B.J. The effects of using mobile devices on student achievement in language learning: A meta-analysis. *Educ. Sci.* **2018**, *8*, 105. [\[CrossRef\]](#)
31. Derwing, T.M.; Munro, M.J. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*; John Benjamins: Amsterdam, The Netherlands, 2015. [\[CrossRef\]](#)
32. Kang, O.; Rubin, D.; Pickering, L. Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Mod. Lang. J.* **2010**, *94*, 554–566. [\[CrossRef\]](#)
33. Pikhart, M.; Klimova, B.; Ruschel, F.B. Foreign Language Vocabulary Acquisition and Retention in Print Text vs. Digital Media Environments. *Systems* **2023**, *11*, 30. [\[CrossRef\]](#)
34. Krashen, S.D. *The Input Hypothesis: Issues and Implications*; Longman: London, UK, 1985.
35. Schmidt, R. The role of consciousness in second language learning. *Appl. Linguist.* **1990**, *11*, 129–158. [\[CrossRef\]](#)
36. Swain, M. Three functions of output in second language learning. In *Principles and Practice in Applied Linguistics: Studies in Honor of H.G. Widdowson*; Oxford University Press: Oxford, UK, 1995; pp. 125–144.
37. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [\[CrossRef\]](#)
38. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [\[CrossRef\]](#)
39. Al-Araibi, A.A.; Naz'ri Bin Mohd Nawi, A.; Al-Badi, A. A conceptual model of educational technology acceptance and use. *Int. J. Emerg. Technol. Learn.* **2019**, *14*, 4–18.
40. Wu, D.; Zhang, S.; Ma, Z.; Yue, X.-G.; Dong, R.K. Unlocking Potential: Key Factors Shaping Undergraduate Self-Directed Learning in AI-Enhanced Educational Environments. *Systems* **2024**, *12*, 332. [\[CrossRef\]](#)
41. Hayadi, B.H.; Hariguna, T. Determinants of student engagement and behavioral intention towards mobile learning platforms. *Contemp. Educ. Technol.* **2025**, *17*, ep558. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [\[CrossRef\]](#)
43. Yu, J.; Song, J.; Lu, Y. Harnessing Generative Artificial Intelligence to Construct Multimodal Resources for Chinese Character Learning. *Systems* **2025**, *13*, 692. [\[CrossRef\]](#)
44. Braun, V.; Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [\[CrossRef\]](#)
45. Almassaad, A.; Alajlan, H.; Alebaikan, R. Student Perceptions of Generative Artificial Intelligence: Investigating Utilization, Benefits, and Challenges in Higher Education. *Systems* **2024**, *12*, 385. [\[CrossRef\]](#)

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.