**Journal of Technology Informatics and Engineering (JTIE)**
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
DOI: https://doi.org/10.51903/jtie.v5i2.383

# Development and Usability Evaluation of an AI-Powered English Vocabulary Learning Application

Joseflim Marchel*[1], Dendy Kurniawan[1], Azed Yayah Durrotun Nihayah[2]
Email: josef@stekom.ac.id, dendy@stekom.ac.id, azed@stekom.ac.id
Orcid: https://orcid.org/0009-0005-3850-5842, https://orcid.org/0009-0009-2667-7136, https://orcid.org/0009-0000-2388-1160
[1,2]*Dept. of Computer Science. Faculty of Academic Study. Universitas Sains dan Teknologi Komputer, Semarang, Indonesia*
[3]*Dept. of Information System. Faculty of Information Technology. Satya Wacana Christian University. Salatiga. Indonesia*
*Corresponding Author

**Abstract**
*Learning vocabulary is a major challenge for beginner foreign-language learners, as repetition-based methods can be tiring and often lead to rapid forgetting. Generative AI, particularly Large Language Models (LLMs), enables the automatic creation of large amounts of learning content. However, without appropriate pedagogical guidance, AI-generated materials may be inconsistent and provide limited support for effective learning progress. This study therefore developed and evaluated an adaptive application for beginner-level English vocabulary practice. The application uses an LLM to provide contextual assistance and combines it with a modified spaced repetition mechanism. Its vocabulary and learning activities follow the appropriate CEFR level, while the difficulty is adjusted based on the accuracy of each learner's answers. The application was evaluated through a mixed-methods usability study involving 72 learners with beginner-level English proficiency. User experience and acceptance were measured using the System Usability Scale (SUS) and selected measures from the Technology Acceptance Model (TAM). Usability was rated highly. Although the participants were beginners with limited English, they generally found the interface comfortable and manageable. Their SUS responses produced a mean of 81.46 (SD = 7.18), a result classified as "Excellent." They also viewed the application as useful and relatively easy to operate. The respective scores were M = 4.38 (SD = 0.42) for Perceived Usefulness and M = 4.29 (SD = 0.45) for Perceived Ease of Use.*

**Keywords**: *Adaptive Learning Architecture; Artificial Intelligence; Computer-Assisted Language Learning; Human-Computer Interaction; System Usability Scale.*

## I. INTRODUCTION

At an early stage of second-language learning, vocabulary gaps become noticeable quite quickly. Beginners who know relatively few words may lose the meaning of a text or struggle to put a simple idea into words. Part of the problem comes from limited exposure (Ganesan et al., 2025; Chandy et al., 2024). Classroom lessons may be the only regular contact they have with the language. New vocabulary is then learned with little sense of where it belongs in actual communication. A word remembered during practice may not even be noticed when it turns up later in a conversation or reading passage. Digital flashcards and scheduled review tasks, both familiar features of Mobile-Assisted Language Learning (MALL), are intended to keep these words in memory (Lee & Baek, 2023; Shortt et al., 2023). This support is useful, but it mainly deals with recall. Learning a word also involves noticing which words naturally appear beside it and how the surrounding context affects its meaning (Chandy et al., 2024; van den Broek et al.,

*Received on May 2026; Revised on June 2026; Accepted on July 2026; Published on August 2026*

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

2022). Someone may therefore give the correct definition but misread the same word in a fresh sentence, use an odd combination, or overlook a shift in meaning from one situation to another.

Natural Language Processing (NLP) has changed considerably since Large Language Models (LLMs) became available. Their arrival has also given Computer-Assisted Language Learning (CALL) new ways to assist learners (Bahari et al., 2025; Mannekote et al., 2024; Wang et al., 2024). Earlier CALL systems tend to work with fixed rules and content written in advance. An LLM is less restricted in this respect. When support is required, it can build an example around the vocabulary item, point out its different forms, or explain the grammar used with it (Wei, 2023; Zhao, 2025). Using generative AI with beginner learners, however, raises several technical and instructional concerns. If its output is not carefully controlled, an LLM may provide inaccurate information, introduce language beyond the learner's current Common European Framework of Reference for Languages (CEFR) level, or present more information than the learner can reasonably process (Karataş et al., 2024; Wang & Reynolds, 2024). Generative capabilities therefore need to be placed within a structured support system and combined with planned vocabulary-retrieval activities. Without these safeguards, an AI-based learning tool may add unnecessary complexity to the learning experience instead of helping beginners develop vocabulary that they can understand, retain, and use.

A review of the literature points to an important gap linking educational software development, adaptive AI, and usability research. Previous studies have examined gamification and general-purpose LLM chat interfaces in considerable detail (Shortt et al., 2023; Lee & Baek, 2023; Ge et al., 2025). However, only a limited number describe how an adaptive vocabulary-learning system can be built to keep AI-generated material suitable for beginners while also adjusting spaced-repetition activities to learners' responses (Sayed et al., 2023; Naseer et al., 2025). There is also relatively little evidence on how beginner learners actually experience and accept systems that combine these features. Aspects related to human-computer interaction (HCI), overall usability, and technology acceptance have not been studied extensively with genuine beginner groups. Existing projects often focus on demonstrating what generative AI can produce rather than evaluating the resulting system with standardized usability measures. This leaves an important question about interface design and adaptive prompt management. Their effects on learners' mental effort and on how efficiently learners can operate the system are still not fully understood.

This study responds to the identified limitations by reporting the design, technical development, and usability testing of an AI-supported English vocabulary application for beginners. Most of the system's work is divided between two components. The first is a locally managed LLM that supplies language support through prompts. Difficulty is handled separately by a rule-based

component. Together, they form a hybrid architecture that keeps examples and feedback within the linguistic requirements of CEFR A1-A2. We evaluated this design with a validated group of 72 beginner learners. Their responses to the System Usability Scale (SUS) and the Technology Acceptance Model (TAM) provide empirical reference points for the system's usability and the acceptance of AI-assisted educational software.

The study addressed one development goal and one evaluation goal. On the development side, we created an adaptive web-based vocabulary system that brings together LLM-generated contextual help and rule-based spaced retrieval. Its difficulty changes gradually in response to learner progress. The evaluation then focused on usability, interface ergonomics, and acceptance within a selected group of beginner language learners. We measured these aspects with standardized psychometric instruments. Section 2 provides the background for the study through earlier research on adaptive Computer-Assisted Language Learning (CALL). Cognitive ergonomics and usability evaluation are also discussed there. Section 3 moves to the research process and describes the system architecture, adaptation models, participant characteristics, and statistical instruments. The usability findings are presented in Section 4 and compared with previous work in human-computer interaction and educational technology. The paper ends with Section 5, which draws together the conclusions, technical limitations, and possible directions for future development.

## II. LITERATURE REVIEW

### *A. Theoretical Foundations: Cognitive Load Theory and Lexical Scaffolding in CALL*

The way learners acquire vocabulary is closely connected to the limits of human memory. Only a small amount of unfamiliar material can be processed in working memory at once, while long-term memory has a much larger capacity. Cognitive Load Theory (CLT) separates the mental demands of learning into three forms. Intrinsic load reflects how difficult the material is. Extraneous load comes from its presentation, including choices made in the learning interface. Germane load is the effort devoted to organizing new knowledge so that its use gradually becomes more automatic (Skulmowski & Xu, 2022; Sweller, 2024). These differences matter for beginners because meeting a new word may require them to process its spelling, pronunciation, and meaning all at once. If a learning application presents the word without a sentence or meaningful situation, memorization becomes the learner's main strategy. Although repetition may support short-term recall, it can also demand unnecessary mental effort and provide limited support for connecting the word with existing knowledge (Ganesan et al., 2025; Krieglstein et al., 2023).

Vocabulary learning is less demanding when support is appropriate to the learner's current ability. This idea is central to the Zone of Proximal Development (ZPD). Learners are given help with

language they cannot yet use independently, but the assistance is intended to be temporary (Gkintoni et al., 2025). In Computer-Assisted Language Learning (CALL), one way to provide this help is to place unfamiliar words in brief and understandable contexts. These contexts should also be consistent with the learner's CEFR proficiency level. Rather than memorizing a word and its definition as separate pieces of information, learners can use the rest of the sentence to interpret what the word means. Previous psycholinguistic research suggests that learning vocabulary in this way supports stronger memory formation and better retention than studying isolated lists (Chandy et al., 2024; van den Broek et al., 2022). Context alone, however, may not ensure that a word is remembered. Learners need to recall it at different points after it is first introduced. Spaced Retrieval Practice provides these opportunities by gradually extending the period between reviews. Repeated recall across time can limit forgetting and help recently acquired words become more firmly established in long-term memory.

### *B. Large Language Models and Adaptive Architectures in Language Acquisition*

Digital language learning has gradually moved beyond early systems that followed fixed rules. The use of Large Language Models (LLMs) has changed the kind of support available in AI-Assisted Language Learning (AIALL) (Bahari et al., 2025). Many Mobile-Assisted Language Learning (MALL) applications still work differently. Their exercises and flashcards are prepared beforehand and stored in a fixed database. Although this keeps computing requirements low, the application has limited room to respond when one learner struggles with a particular word. Repeating the same stored activities can also become predictable and may gradually weaken motivation (Lee & Baek, 2023; Shortt et al., 2023). Generative AI offers the possibility of responding more directly to those individual difficulties. Using transformer-based architectures trained on extensive language data, these models can produce text in real time, explain how words are formed, and create contextual examples based on the information provided by each user (Mannekote et al., 2024; S. Wang et al., 2024).

Recent studies suggest that language tutoring supported by LLMs can provide meaningful benefits for learners. (Zhao, 2025) and (Wei, 2023), for example, found that learners who practiced with generative dialogue systems developed stronger vocabulary knowledge and showed greater independence in managing their learning than those who studied through fixed word lists. Even so, using an LLM without clear controls can create problems in educational settings. A generative model may produce inaccurate information, use idiomatic expressions or sentence structures that are too difficult for beginners, and take an inconsistent amount of time to respond (Karataş et al., 2024; X. Wang & Reynolds, 2024). Effective use of generative AI in primary and secondary vocabulary instruction therefore requires a carefully designed prompt-

management process. This process should keep the generated language within the appropriate CEFR level and connect the LLM to a rule-based adaptive system (Naseer et al., 2025; Sayed et al., 2023). By combining controlled content generation with performance tracking based on defined thresholds, a learning system can provide support suited to each learner's needs without causing unnecessary confusion or reducing their interest in learning (Ge et al., 2025; Zhang & Huang, 2024).

### *C. Human-Computer Interaction, Usability Engineering, and Technology Acceptance*

Advanced algorithms alone do not show whether educational software works well. The experience of the people using the system matters just as much. This makes empirical testing based on Human-Computer Interaction (HCI) and usability principles necessary (Bangor et al., 2008). According to (Bangor et al., 2008), usability concerns whether particular users can achieve their goals effectively, efficiently, and satisfactorily in the setting where the system is used. The consequences are especially important for learning applications. A confusing interface directs learners' attention toward operating the system when that attention should be available for the lesson. Some of their limited mental capacity is therefore spent on navigation, leaving less attention for understanding and remembering unfamiliar vocabulary (Skulmowski & Xu, 2022).

User experience needs to be measured in a consistent and objective way. Software engineering researchers often do this with instruments that have already been tested for psychometric quality. The System Usability Scale (SUS) is among the best-known instruments used for this purpose (Brooke, 1996). It contains ten statements that alternate between positive and negative wording, with each response recorded on a five-point Likert scale. The responses are then combined to produce a single usability score ranging from 0 to 100. Based on extensive evaluations of thousands of software systems, (Bangor et al., 2009) reported that scores above 68.0 indicate better-than-average usability. A score higher than 80.3 falls within the "Excellent" range, suggesting that the interface is highly intuitive and easy to operate.

Alongside usability measures, the Technology Acceptance Model (TAM) provides a well-established framework for understanding why users choose to adopt a new information system (Davis, 1989; Venkatesh & Bala, 2008). According to TAM, actual use is largely shaped by Behavioral Intention (BI). This intention is influenced by two main factors: Perceived Usefulness (PU), or the extent to which users believe that a system can improve their performance, and Perceived Ease of Use (PEOU), which refers to how little effort they expect its use to require. In AI-supported language learning, a high PEOU score suggests that learners can interact with the platform without facing unnecessary interface difficulties. A high PU score, meanwhile, indicates that learners recognize the value of the contextual support provided by the AI (Karataş et al.,

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

2024; Shank et al., 2025). Using TAM together with the System Usability Scale (SUS) therefore allows the system to be assessed from several perspectives, including ease of interaction, cognitive accessibility, and the likelihood that learners will continue using it over time.

## III. RESEARCH METHOD

### *A. Research Design*

The study followed the Design Science Research Methodology (DSRM) and included a cross-sectional usability evaluation. The work was carried out in five stages. These stages covered problem identification and requirements engineering, architecture and algorithm modeling, full-stack implementation with LLM prompt orchestration, usability testing and data collection, and psychometric validation of the model. The engineering work focused on an adaptive web platform that brings together generative AI and deterministic spaced repetition algorithms. After the platform was completed, actual target users took part in a structured post-interaction evaluation. Their responses were used to examine interface ergonomics, cognitive accessibility, and user acceptance.

### *B. Population and Sample*

We recruited participants from the undergraduate introductory English courses at Universitas Sains dan Teknologi Komputer, Indonesia. Enrollment in these introductory English courses was compulsory. Since the research required beginner learners, participants were selected purposively instead of being drawn from several proficiency levels. Each candidate completed the institution's standardized placement test before entering the study. This screening gave the group a comparable starting point. Only candidates classified as A1 (Beginner) under the Common European Framework of Reference for Languages (CEFR) were included.

Table 1. Demographic Profile and Technical Baseline of Participants (N = 72)

| Demographic Attribute | Classification / Range | Frequency (n) | Percentage (%) |
|---|---|---|---|
| Gender | Male | 38 | 52.78 |
| | Female | 34 | 47.22 |
| Age (Years) | 18-19 | 42 | 58.33 |
| | 20-21 | 30 | 41.67 |
| Academic Major | Informatics Engineering | 48 | 66.67 |
| | Information Systems | 24 | 33.33 |
| Primary Access Device | Mobile (Smartphone / Tablet) | 51 | 70.83 |
| | Desktop / Laptop | 21 | 29.17 |
| Prior Exposure to AI Tools | None / Infrequent (< 1 hr/wk) | 45 | 62.50 |
| | Moderate (1-3 hrs/wk) | 22 | 30.56 |
| | Frequent (> 3 hrs/wk) | 5 | 6.94 |

Seventy-two learners formed the final cohort, with 38 males and 34 females. They were between 18 and 21 years old (Mean = 19.42, SD = 0.86). All of them finished both the interaction protocol

and the subsequent evaluation battery. The group was also larger than the psychometric threshold discussed by (Faulkner, 2003). A sample above N=60 is reported to provide more than 98% statistical confidence in detecting serious usability problems while producing dependable parametric System Usability Scale (SUS) estimates. Full demographic information and the cohort's technical baseline appear in Table 1.

### *C. System Architecture and Algorithmic Workflow*

To keep learner interactions responsive, the application distributes its work across a modular client-server architecture rather than a single component. Learner actions begin in (1) the Presentation Layer / Client Interface and are sent to (2) the Backend API and State Management Server. A separate component, (3) the Adaptive Spaced Retrieval Engine, determines the review schedule. Contextual language assistance is supplied by (4) the LLM-Assisted Contextual Generation Module. The arrangement of these components and the movement of information between them are shown in Figure 1.

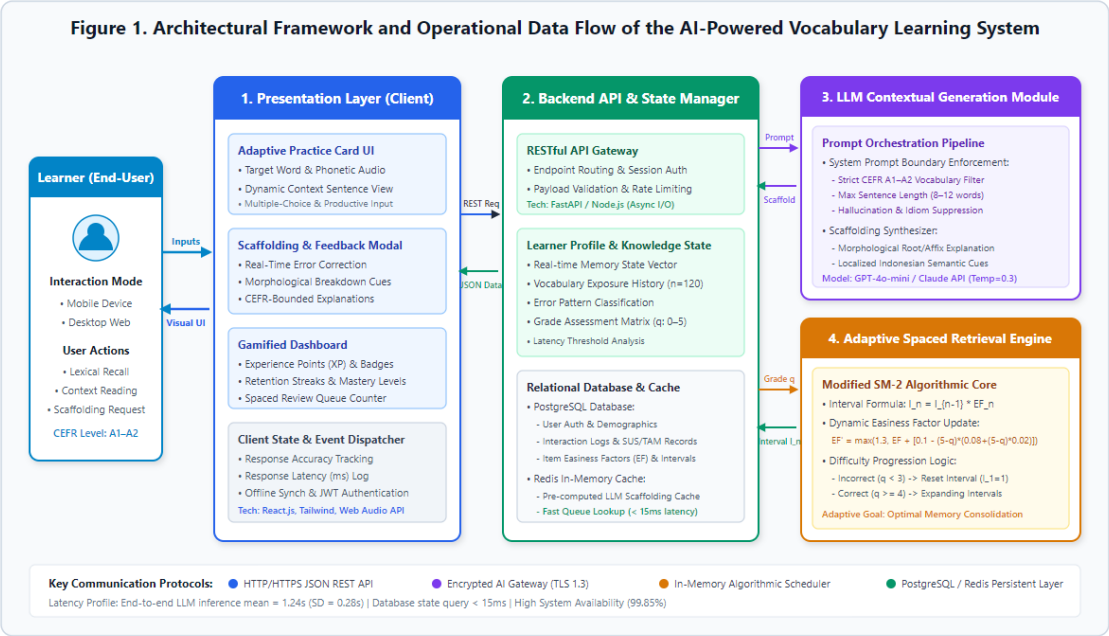


**Figure 1. Architectural framework and operational data flow of the AI-powered vocabulary learning system.**

Students work from a responsive interface developed with React.js and styled through Tailwind CSS. From this interface, they can open vocabulary cards, complete short contextual tasks, and listen to the pronunciation of a target word. Learning progress is represented in several ways rather than through a single score. The interface records experience points, awards mastery badges, and maintains retention streaks as learners continue their practice. FastAPI and a Node.js REST backend handles the server-side processes. Each interaction is written to an encrypted PostgreSQL database, together with the learner state vector, previous error history, and an exact

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

timestamp. Vocabulary review is not arranged at fixed intervals. Instead, a customized SuperMemo (SM-2) algorithm recalculates the next review time from the difficulty of the item, the learner's earlier accuracy, and the time taken to respond. The LLM orchestration pipeline becomes active when a learner struggles with a particular word. A restricted system prompt keeps the grammar within CEFR A1-A2. It asks the model for a simple contextual sentence, a breakdown of the word's morphological root, and localized clues to its meaning. Keeping the request this narrow helps reduce hallucinated information and limits vocabulary that falls outside the permitted level.

### *D. Mathematical Formulas and Models*

Three computational models were formally defined within the research framework. They make the mathematical procedures explicit, support replication of the empirical process, and provide a consistent basis for psychometric validation.

#### *1. System Usability Scale (SUS) Calculation Model*

To obtain the composite SUS result, we transformed responses from all ten items with the standardized procedure described by (Brooke, 1996). Scoring depends on the item number because the odd items are positively worded and the even items are negatively worded. The complete calculation appears in Equation 1.

$$\text{SUS} = 2.5 \times \left[ \sum_{i \in \text{Odd}} (R_i - 1) + \sum_{i \in \text{Even}} (5 - R_i) \right] \quad (1)$$

In Equation 1, SUS is the converted composite usability score, expressed on a scale from 0 to 100. The symbol $R_i$ refers to the raw response for questionnaire item $i$, with each response scored from 1 to 5 on a Likert scale. Odd = {1,3,5,7,9} contains the positive items, whereas Even = {2,4,6,8,10} contains the reverse-coded items. The adjusted item scores produce a total between 0 and 40. Multiplying this total by 2.5 converts it to the standard 100-point scale.

#### *2. Internal Consistency Reliability Model*

We examined the psychometric reliability of the measurement instruments using Cronbach's Alpha. The calculation for this coefficient is provided in Equation 2.

$$\alpha = \left( \frac{k}{k-1} \right) \times \left[ 1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2} \right] \quad (2)$$

In this formula, α is the coefficient of internal consistency, while $k$ indicates how many survey items are included in the construct. The variance for an individual item $j$ is represented by $\sigma \frac{2}{j}$,

and $\sigma \frac{2}{x}$ refers to the total variance of the composite scores for all participants. We used $\alpha \geq 0.70$ as the minimum value for acceptable internal consistency.

*3. Adaptive Spaced Repetition Scheduling Algorithm*

We used a modified SM-2 model to determine when each item should be presented again. Its recursive calculation is shown in Equation 3.

$$I_n = \begin{cases} 1, & \text{for } n = 1 \\ 6, & \text{for } n = 2 \\ I_{n-1} \times EF_n, & \text{for } n > 2 \end{cases} \quad (3)$$

In Equation 3, $I_n$ indicates the number of days before repetition n takes place. $EF_n$ is the Easiness Factor, which changes dynamically after each learner interaction. The procedure used to update this value appears in Equation 4.

$$EF_n = \max(1.3, EF_{n-1} + [0.1 - (5 - q) \times (0.08 + (5 - q) \times 0.02)]) \quad (4)$$

In this calculation, $q$ records the quality of the learner's response on a scale from 0 to 5. A score of 0 indicates complete failure or a timeout, while 5 indicates perfect recall with minimal delay. The minimum value of 1.3 keeps the review interval from falling into an infinite cycle when a lexical item is difficult.

*E. Variables and Operational Definition*

The study variables were organized around interface ergonomics and the psychological acceptance of users. Overall Usability was treated as the main dependent measure and represented by a composite SUS score from 0 to 100. We then considered this measure through two separate sub-dimensions. Usability included Items 1, 2, 3, 5, 6, 7, 8, and 9, while Learnability included Items 4 and 10. Four constructs were used to capture psychological acceptance on a 1-to-5 Likert scale. Perceived Usefulness (PU) was represented by three items about whether AI-generated context and adaptive spacing made vocabulary learning more efficient. Another three items addressed Perceived Ease of Use (PEOU), including the effort and clarity involved in the interface, navigation, and feedback. Attitude Toward Using (ATU) was examined with three items concerning users' cognitive and emotional responses to the system. The remaining three items focused on Behavioral Intention (BI) and asked whether learners were prepared to make the application part of their regular study routine.

*F. Measurement Instruments, Validity, and Reliability Testing*

Evaluation data came from two questionnaires. We used the standard ten-item System Usability Scale (SUS) and a 12-item version of the Technology Acceptance Model (TAM) adapted from (Davis, 1989) and (Venkatesh & Bala, 2008). For every TAM statement, participants selected a

value from 1, Strongly Disagree, to 5, Strongly Agree. Before data collection, the questionnaire was reviewed by three specialists representing Computer-Assisted Language Learning (CALL), Software Engineering, and Educational Measurement. Their review addressed content and construct validity. Because the participants were Indonesian speakers, the items underwent forward and backward translation into Bahasa Indonesia. We compared the translated versions to preserve psychological meaning and avoid unnatural expressions. Preliminary testing produced Cronbach's alpha values of $\alpha = 0.88$ for SUS and $\alpha = 0.91$ for the combined TAM items, indicating strong internal consistency.

### *G. Data Sources, Collection Procedures, and Data Analysis Techniques*

We collected the data in two stages. During the learning stage, all 72 participants worked through the same 40 standardized vocabulary modules for beginners. The modules were distributed across three successive days, with a 45-minute session scheduled on each day. Once the last session was complete, participants logged into the online portal to finish the computerized evaluation battery. Quantitative analysis was performed in Python with SciPy and Statsmodels, as well as in IBM SPSS Statistics v27. We summarized the demographic information and questionnaire constructs by reporting Means, Standard Deviations, Medians, and Interquartile Ranges. The Shapiro-Wilk test was used to assess normality at $p > 0.05$. Composite SUS scores were interpreted against the percentile curves and adjective ratings provided by (Bangor et al., 2009). We also examined the TAM sequence $PEOU \rightarrow PU \rightarrow ATU \rightarrow BI$ through Pearson product-moment correlation and multiple linear regression. Statistical significance was evaluated at $\alpha = 0.05$.

### *H. Ethical Considerations*

Formal approval was obtained from the institutional review board before the study began, in line with the institution's requirements for human-participant research. Every candidate received an information sheet describing the study's purpose, scope, procedures, and voluntary nature. Participation proceeded only after the student submitted written informed consent digitally. For confidentiality, identifying information in the survey records and interaction logs was replaced with a unique cryptographic ID. The resulting anonymous records were stored on encrypted local servers that could be accessed only by authorized personnel.

## IV. RESULT AND DISCUSSION

### *A. Result*

Evidence from the system rollout is reported in this section. It includes technical performance measurements and psychometric responses from N=72 beginner learners. The results are described first, while their interpretation is left for the discussion that follows. Materials reported

here include the main interface outputs, the distribution of System Usability Scale (SUS) scores, individual Technology Acceptance Model (TAM) constructs, relationships among the variables, and the multiple linear regression models. .

### *1. System Implementation and Technical Performance*

Deployment took place under live network conditions, with the responsive layout checked on desktop and mobile devices. Figure 2 documents the three interface views used by learners. Panel (a) presents the Adaptive Vocabulary Practice Screen, including the target word, a CEFR-limited context sentence generated automatically, and audio synthesis. When a learner answers incorrectly, the Dynamic Feedback Modal in panel (b) opens with an immediate explanation and morphological analysis. Panel (c) contains the Learner Dashboard, which reports gamified performance, upcoming spaced repetition reviews, and mastery streaks.

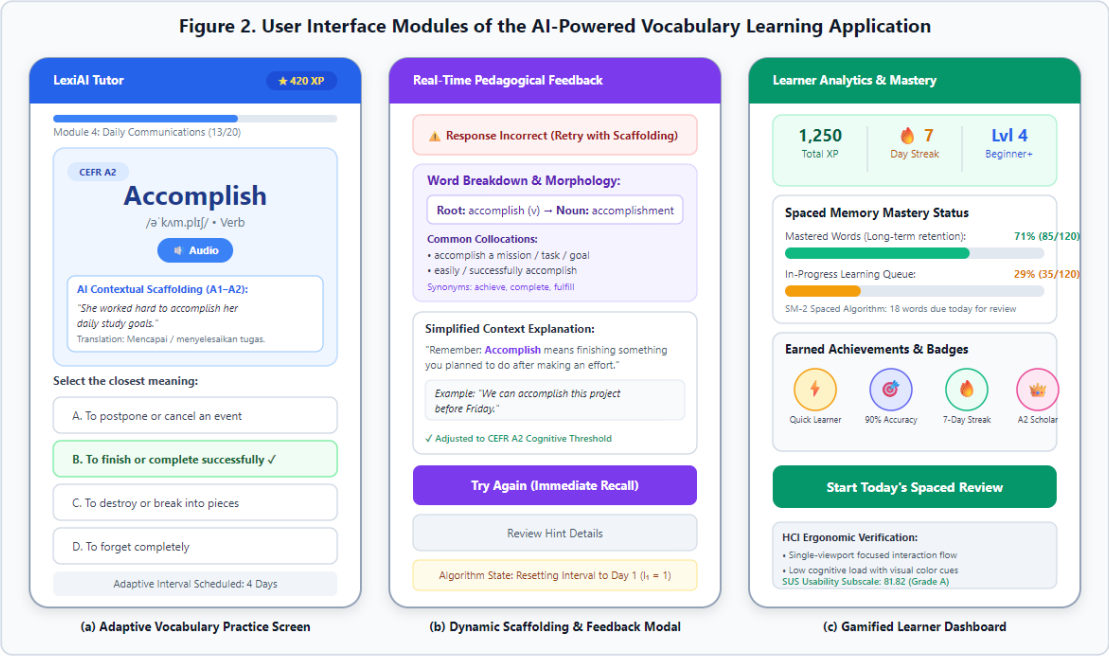


**Figure 2. User interface modules of the AI-powered vocabulary learning application: (a) adaptive vocabulary card with LLM-generated contextual sentence, (b) real-time scaffolding feedback modal, and (c) gamified learner dashboard**

No critical software exception or client-side rendering crash occurred during the interaction sessions. Monitoring across the evaluation with 72 participants also recorded a backend API uptime of 99.85%. The LLM prompt orchestration pipeline completed end-to-end inference in $M = 1.24\,seconds$ on average $(SD = 0.28\,s, Range = 0.82 - 1.95\,s)$. This latency remained within the accepted response range for interactive computer-assisted instructional systems.

### *2. System Usability Scale (SUS) Evaluation*

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The 10-item System Usability Scale yielded an overall mean composite score of $M = 81.46$ ($SD = 7.18$, Median $= 82.50$, Range $= 65.00 - 95.00$) across the N = 72 participants. The Shapiro-Wilk result was $W = 0.974\, with\, p = 0.142$, providing a basis for treating the distribution as normal. Item-level SUS results are reported in Table 2. Each row contains the mean of the raw Likert responses, the corresponding standard deviation, and the adjusted score contribution obtained with Equation 1.

**Table 2. Item-Level Descriptive Statistics for the System Usability Scale (N = 72)**

| Item No. | Item Description | Mean (SD) | Adjusted Contribution (0–4) |
|---|---|---|---|
| SUS 01 | I think that I would like to use this system frequently. | 4.32 (0.62) | 3.32 |
| SUS 02 | I found the system unnecessarily complex. | 1.72 (0.65) | 3.28 |
| SUS 03 | I thought the system was easy to use. | 4.40 (0.55) | 3.40 |
| SUS 04 | I think that I would need the support of a technical person to use it. | 1.58 (0.64) | 3.42 |
| SUS 05 | I found the various functions in this system were well integrated. | 4.25 (0.60) | 3.25 |
| SUS 06 | I thought there was too much inconsistency in this system. | 1.68 (0.67) | 3.32 |
| SUS 07 | I would imagine that most people would learn to use this system quickly. | 4.44 (0.58) | 3.44 |
| SUS 08 | I found the system very cumbersome to use. | 1.63 (0.61) | 3.37 |
| SUS 09 | I felt very confident using the system. | 4.22 (0.63) | 3.22 |
| SUS 10 | I needed to learn a lot of things before I could get going with it. | 1.74 (0.69) | 3.26 |
| Overall | Composite SUS Score (Normalized Scale 0–100) | 81.46 (7.18) | -- |

For the dimensional analysis, the SUS items were divided into the two established components. The Usability dimension included Items 1, 2, 3, 5, 6, 7, 8, and 9, while Learnability was measured through Items 4 and 10. The Usability subscale achieved a mean score of $M = 81.82$ ($SD = 7.34$), while the Learnability subscale yielded a mean score of $M = 80.00$ ($SD = 8.65$). We interpreted the participant scores with the adjective and acceptability bands established by (Bangor et al., 2009). Table 3 shows that 58.33% reached the "Excellent" category by scoring at least 80.3. A further 36.11% received a "Good" classification, with scores starting at 68.0 but remaining below 80.3. Only 5.56% scored from 51.0 to below 68.0 and were classified as "OK / Marginal." No participant recorded a score below 50.0.

Table 3. Distribution of Participant SUS Scores Categorized by Usability Benchmarks (N = 72)

| SUS Score Range | Adjective Rating | Acceptability Range | Grade Scale | Frequency (n) | Percentage (%) |
|---|---|---|---|---|---|
| $\geq 80.3$ | Excellent | Acceptable | Grade A | 42 | 58.33 |
| $68.0 - 80.2$ | Good | Acceptable | Grade B / C | 26 | 36.11 |
| $51.0 - 67.9$ | OK / Marginal | Marginal | Grade D | 4 | 5.56 |
| $< 51.0$ | Poor / Awful | Not Acceptable | Grade F | 0 | 0.00 |

### 3. *Technology Acceptance Model (TAM) Evaluation*

Table 4 summarizes responses to the 12-item Technology Acceptance questionnaire, which covered four theoretical constructs on a 5-point Likert scale. Every construct had a mean above

4.20 out of 5.00, and the scores showed relatively little variation between participants. Along with these descriptive results and measures of dispersion, the table reports internal reliability. Cronbach's alpha ranged from $\alpha = 0.84$ to $\alpha = 0.89$ across the four constructs and reached $\alpha = 0.92$ when the complete set of items was analyzed.

Table 4. Descriptive Statistics and Reliability Indices for TAM Constructs (N = 72)

| Construct | No. of Items | Mean (M) | Std. Dev. (SD) | Median | Min | Max | Cronbach's Alpha ($\alpha$) |
|---|---|---|---|---|---|---|---|
| Perceived Usefulness (PU) | 3 | 4.38 | 0.42 | 4.33 | 3.33 | 5.00 | 0.86 |
| Perceived Ease of Use (PEOU) | 3 | 4.29 | 0.45 | 4.33 | 3.00 | 5.00 | 0.84 |
| Attitude Toward Using (ATU) | 3 | 4.33 | 0.41 | 4.33 | 3.33 | 5.00 | 0.87 |
| Behavioral Intention (BI) | 3 | 4.41 | 0.39 | 4.33 | 3.67 | 5.00 | 0.89 |
| Composite Instrument | 12 | 4.35 | 0.38 | 4.33 | 3.42 | 5.00 | 0.92 |

Perceived Ease of Use and the composite SUS result showed one of the clearest relationships in the analysis ($r = 0.764, p < 0.001$). Participants who reported that the system required little effort also tended to assign higher usability scores. Pearson's product-moment analysis was used for this comparison. The wider pattern in Table 5 was consistent, as every coefficient was positive and statistically significant at $p < 0.001$. The close agreement between ease of operation and overall usability also provides evidence of concurrent validity. Perceived Usefulness showed substantial positive associations with Attitude Toward Using ($r = 0.742, p < 0.001$) and Behavioral Intention ($r = 0.718, p < 0.001$).

Table 5. Pearson Bivariate Correlation Matrix Among Evaluated Constructs (N = 72)

| Construct | (1) SUS | (2) PEOU | (3) PU | (4) ATU | (5) BI |
|---|---|---|---|---|---|
| (1) SUS | 1.000 | -- | -- | -- | -- |
| (2) PEOU | 0.764*** | 1.000 | -- | -- | -- |
| (3) PU | 0.631*** | 0.682*** | 1.000 | -- | -- |
| (4) ATU | 0.685*** | 0.715*** | 0.742*** | 1.000 | -- |
| (5) BI | 0.658*** | 0.694*** | 0.718*** | 0.792*** | 1.000 |

*Note:* *** $p < 0.001$ *(two-tailed).*

A multiple linear regression model was fitted to examine the predictive validity of the acceptance framework. Behavioral Intention (BI) served as the outcome variable, with PEOU, PU, and ATU entered simultaneously as predictors. The overall regression model was statistically significant, $F(3,68) = 50.88, p < 0.001$, accounting for 69.2% of the variance in Behavioral Intention ($R^2 = 0.692$, Adjusted $R^2 = 0.678$). As summarized in Table 6, Attitude Toward Using emerged as the strongest direct predictor ($\beta = 0.482, t = 4.31, p < 0.001$), followed by Perceived Usefulness ($\beta = 0.264, t = 2.48, p = 0.016$) and Perceived Ease of Use ($\beta = 0.185, t = 1.82, p = 0.073$).

### *B. Discussion*

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 5 No. 2 August 2026**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 232-250
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The evidence points to three strengths of the system: technical feasibility, manageable interface ergonomics, and positive user acceptance. Two results are especially relevant, namely the composite System Usability Scale score of 81.46 and the favorable response across all Technology Acceptance Model dimensions. The platform also brings controlled generative AI together with adaptive spaced retrieval. This combination responds to cognitive and usability difficulties commonly encountered in computer-assisted language learning tools for beginners.

Table 6. Multiple Linear Regression Model Summary Predicting Behavioral Intention (BI)

| **Predictor Variable** | **Unstandardized B** | **Std. Error (SE)** | **Standardized Beta ($\beta$)** | **t-value** | **p-value** | **Collinearity (VIF)** |
|---|---|---|---|---|---|---|
| (Constant) | 0.612 | 0.284 | -- | 2.15 | 0.035 | -- |
| Perceived Ease of Use | 0.161 | 0.088 | 0.185 | 1.82 | 0.073 | 2.14 |
| Perceived Usefulness | 0.245 | 0.099 | 0.264 | 2.48 | 0.016 | 2.35 |
| Attitude Toward Using | 0.458 | 0.106 | 0.482 | 4.31 | $< 0.001$ | 2.56 |

*Model Diagnostics:* $R = 0.832$, $R^2 = 0.692$, *Adjusted* $R^2 = 0.678$, $F(3,68) = 50.88$, $p < 0.001$.

### *1. Usability Benchmarking and Interface Ergonomics*

Published evaluations of commercial Mobile-Assisted Language Learning (MALL) applications provide a useful comparison. Reported SUS scores for Duolingo, Memrise, and Busuu generally range from 71.0 to 76.5 (Lee & Baek, 2023; Shortt et al., 2023). The application evaluated here produced a higher mean of $M = 81.46$ ($SD = 7.18$). This value falls in the 91st percentile of global usability benchmarks and is classified as "Excellent" with "Grade A" under (Bangor et al., 2009). Several architectural and interaction design decisions may have contributed to the difference. One design decision was to keep most learning activities inside a single viewport, avoiding unnecessary navigation between screens. The item-level results reflect this choice. SUS Item 3, "easy to use," received $M = 4.40$, while Item 4, "need technical support," received only $M = 1.58$. A Learnability score of $M = 80.00$ further indicates that beginners could understand how to operate the application without a lengthy adjustment period. Visual clarity came from a simple layout and color-coded badges, while immediate pronunciation audio provided direct feedback. These features limited uncertainty during interaction and followed basic heuristic ergonomics (Nielsen, 1994).

### *2. Cognitive Load Regulation Through Bounded LLM Scaffolding*

Attention to cognitive limits may help explain the system's usability results. Under Cognitive Load Theory, beginners have greater difficulty when a learning task includes mental demands that do not contribute directly to learning (Sweller, 2024). Two common approaches can create this situation. Conventional flashcards often repeat definitions without placing the word in a meaningful setting, adding effort without supporting deeper understanding (Ganesan et al., 2025; Krieglstein et al., 2023). Unrestricted generative chatbots create another risk because their

responses may contain advanced vocabulary, idioms, and sentence patterns that beginners are not yet prepared to process (Karataş et al., 2024; X. Wang & Reynolds, 2024).

Rather than allowing the model to respond freely, the prompts were restricted to CEFR A1-A2. When a learner found a word difficult, the generated support used a short contextual example and familiar grammar. The Perceived Usefulness result of $M = 4.38$ ($SD = 0.42$) indicates that participants valued this assistance. An incorrect response also triggered two immediate forms of help: a breakdown of the word's morphology and a localized semantic clue. This encouraged learners to examine how the meaning was formed instead of repeating the answer from memory. From a cognitive load perspective, the arrangement reduces extraneous effort and promotes germane schema construction (Chandy et al., 2024; Skulmowski & Xu, 2022; van den Broek et al., 2022).

*3. Technology Acceptance and Behavioral Adoption Dynamics*

Students' willingness to adopt the educational AI tool can be understood from the pattern among the TAM variables. None of the theoretical pathways reported in Table 5 was negative or statistically insignificant. Perceived Ease of Use was closely associated with the composite SUS result ($r = 0.764, p < 0.001$). In practical terms, learners who felt that the interface required less effort also rated its overall usability more favorably. The relationship highlights the role of interface ergonomics in shaping an effortless interaction experience.

Table 6 indicates that Behavioral Intention was influenced most strongly by Attitude Toward Using ($\beta = 0.482, p < 0.001$). Perceived Usefulness made a smaller but still significant contribution through direct and indirect paths ($\beta = 0.264, p = 0.016$). Together, PEOU, PU, and ATU accounted for 69.2% of the variance in Behavioral Intention, which had a score of 4.41. These results are in line with the acceptance models proposed by (Davis, 1989) and (Venkatesh & Bala, 2008). Ease of use may persuade students to begin using an application, but a visible learning benefit is also important if that use is to continue. The Behavioral Intention score of 4.41 suggests that learners can develop trust in generative AI when the technology operates within clear educational boundaries (Sayed et al., 2023; Shank et al., 2025).

*4. Theoretical and Practical Implications*

At the theoretical level, the findings connect generative AI in Computer-Assisted Language Learning (CALL) with Cognitive Load Theory. The evidence suggests that prompt design can mediate between raw LLM output and the needs of a beginner learner. By limiting and organizing what the model produces, the system turns potentially overwhelming content into language

support that is structured and suitable for the learner's current level. For software engineering, the study contributes an architecture that has been tested for running Large Language Models in mobile and web learning environments. The design links localized prompt orchestration to a modified SM-2 repetition engine, while the API recorded a mean latency of $M = 1.24\ s$. This arrangement offers developers a working reference for preserving pedagogical controls without sacrificing responsive system behavior.

### *5. Limitations and Future Research Directions*

Several boundaries need to be considered when interpreting these results. The participant group consisted of N = 72 undergraduate beginners from a single institution, so the same findings cannot yet be assumed for other populations. Further testing should include secondary school students and adults engaged in lifelong learning. The prompt pipeline also remains limited to CEFR A1-A2. Future development could allow the language level to rise dynamically toward B1-C1 as learner proficiency improves. In addition, the present evaluation concentrated on usability and acceptance rather than long-term learning. Studies extending across several semesters are needed to examine delayed vocabulary retention and whether the pedagogical benefits continue over time.

## V. CONCLUSION AND RECOMMENDATION

This project sought a middle ground between static vocabulary flashcards and unrestricted conversational AI. Its design pairs adaptive spaced retrieval with prompt-controlled LLM sentences kept within CEFR A1-A2. A live trial involving N = 72 beginner learners showed that participants could operate the interface with little difficulty and generally responded well to its ergonomic design. The usability findings supported this positive response. SUS produced $M = 81.46$ ($SD = 7.18$), corresponding to the "Excellent" category, Grade A, and the 91st percentile. TAM results showed a similar pattern, with $M = 4.38$ for Perceived Usefulness, $M = 4.29$ for Perceived Ease of Use, and $M = 4.33$ for Attitude Toward Using. In the acceptance model, these measures accounted for 69.2% of the variance in Behavioral Intention. The mean Behavioral Intention score for continued daily use was $M = 4.41$.

The practical implications apply to software engineers, instructional designers, and language teachers in different ways. Developers should avoid allowing AI tutoring systems to depend on unrestricted conversational prompts alone. A rule-based middleware layer can compare generated text with standard proficiency vocabulary before it reaches the learner, reducing the risk of unnecessary cognitive load. Language teachers can use the platform as a pedagogically guided option for independent vocabulary practice outside regular lesson time. The current evidence comes from an initial evaluation involving beginners at a single university. A future version

should allow CEFR difficulty to adjust dynamically for B1-C1 learners and could apply serverless edge-caching to shorten inference time. Longer trials conducted over several semesters would also make it possible to track vocabulary retention and delayed knowledge transfer across more varied learner populations.

## DATA AVAILABILITY

The anonymized data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to participant confidentiality and institutional ethical requirements, the raw interaction logs and questionnaire records are not publicly accessible.

## REFERENCES

Bahari, A., Han, F., & Strzelecki, A. (2025). Integrating CALL and AIALL for an interactive pedagogical model of language learning. *Education and Information Technologies 2025 30:10*, *30*(10), 14305–14333. https://doi.org/10.1007/S10639-025-13388-W

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, *4*, 3-. https://doi.org/10.1145/3472749.3474738

Brooke, J. (1996). *Usability Evaluation In Industry - Google Books: 189.3*. Usability evaluation in industry.

Chandy, R., Serrano, R., & Pellicer-Sánchez, A. (2024). The effect of context on the processing and learning of novel L2 vocabulary while reading. *Applied Psycholinguistics*, *45*(6), 1086–1113. https://doi.org/10.1017/S0142716424000407

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383. https://doi.org/10.3758/BF03195514

Ganesan, P., Binti Abdul Aziz, A., & Ismail, H. H. (2025). Comparing Rote Memorization and Contextual Learning in Vocabulary Acquisition among Upper Primary ESL Students. *International Journal of Academic Research in Progressive Education and Development*, *14*(2). https://doi.org/10.6007/IJARPED/v14-i2/25165

Ge, W., Sun, Y., Wang, Z., Zheng, H., He, W., Wang, P., Zhu, Q., Wang, B., & Wang, P.-H. (2025). SRLAgent: Enhancing Self-Regulated Learning Skills through Gamification and LLM Assistance. *Proceedings of Preprint*, *1*.

Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in

Redefining Learning Efficacy. *Brain Sciences*, *15*(2), 203. https://doi.org/10.3390/brainsci15020203

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, *29*(15), 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and Validation of a Theory-Based Questionnaire to Measure Different Types of Cognitive Load. *Educational Psychology Review*, *35*(1), 9. https://doi.org/10.1007/s10648-023-09738-0

Lee, J.-Y., & Baek, M. (2023). Effects of Gamification on Students' English Language Proficiency: A Meta-Analysis on Research in South Korea. *Sustainability*, *15*(14), 11325. https://doi.org/10.3390/su151411325

Mannekote, A., Davies, A., Pinto, J. D., Zhang, S., Olds, D., Schroeder, N. L., Lehman, B., Zapata-Rivera, D., & Zhai, C. (2024). Large language models for whole-learner support: opportunities and challenges. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1460364

Naseer, F., Khan, M. N., Addas, A., Awais, Q., & Ayub, N. (2025). Game Mechanics and Artificial Intelligence Personalization: A Framework for Adaptive Learning Systems. *Education Sciences*, *15*(3), 301. https://doi.org/10.3390/EDUCSCI15030301/S1

Nielsen, J. (1994). *Usability Engineering | Enhanced Reader*. Morgan Kaufmann Publishers Inc.

Sayed, W. S., Noeman, A. M., Abdellatif, A., Abdelrazek, M., Badawy, M. G., Hamed, A., & El-Tantawy, S. (2023). AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform. *Multimedia Tools and Applications*, *82*(3), 3303–3333. https://doi.org/10.1007/s11042-022-13076-8

Shank, E., Tang, H., & Morris, W. (2025). Motivation in online course design using self-determination theory: an action research study in a secondary mathematics course. *Educational Technology Research and Development*, *73*(1), 415–441. https://doi.org/10.1007/s11423-024-10410-9

Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: a systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*, *36*(3), 517–554. https://doi.org/10.1080/09588221.2021.1933540

Skulmowski, A., & Xu, K. M. (2022). Understanding Cognitive Load in Digital and Online Learning: a New Perspective on Extraneous Cognitive Load. *Educational Psychology Review*, *34*(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Sweller, J. (2024). Cognitive load theory and individual differences. *Learning and Individual Differences*, *110*, 102423. https://doi.org/10.1016/j.lindif.2024.102423

van den Broek, G. S. E., Wesseling, E., Huijssen, L., Lettink, M., & van Gog, T. (2022). Vocabulary Learning During Reading: Benefits of Contextual Inferences Versus Retrieval Opportunities. *Cognitive Science*, *46*(4), e13135. https://doi.org/10.1111/COGS.13135

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large Language Models for Education: A Survey and Outlook. *IEEE Signal Processing Magazine*, *42*(6), 51–63. https://doi.org/10.1109/MSP.2025.3594309

Wang, X., & Reynolds, B. L. (2024). Beyond the Books: Exploring Factors Shaping Chinese English Learners' Engagement with Large Language Models for Vocabulary Learning. *Education Sciences*, *14*(5), 496. https://doi.org/10.3390/educsci14050496

Wei, L. (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1261955

Zhang, Z., & Huang, X. (2024). Exploring the impact of the adaptive gamified assessment on learners in blended learning. *Education and Information Technologies*, *29*(16), 21869–21889. https://doi.org/10.1007/s10639-024-12708-w

Zhao, Y. (2025). The Impact of GPT-based Dialogue Systems vs. Traditional Word-list Memorization on Second Language Vocabulary Acquisition. *Lecture Notes in Education Psychology and Public Media*, *106*(1), 41–50. https://doi.org/10.54254/2753-7048/2025.CB25093