OPEN ACCESS
Journal International of Lingua and Technology
Vol. 4 No. 2. August 2025, pp. 225-243 DOI. 10.55849/jiltech.v4i2.1033

Research Article

A HUMAN-COMPUTER INTERACTION (HCI) ANALYSIS OF DIGITAL DICTIONARY APPLICATIONS: A COMPARISON OF UX DESIGN AND ITS IMPACT ON VOCABULARY RETENTION

Alvianus Dengen[1], Thabo Mokoena[2], and Nikhil Joshi[3]
[1] Universitas Teknologi Sulawesi, Indonesia
[2] University Cape Town, South Africa
[3] National Institute of Technology (NIT) Trichy, India

Corresponding Author:

Alvianus Dengen,
Department of Electrical Engineering, Faculty of Engineering, Universitas Teknologi Sulawesi.
Jalan Talas Salapang No. 51, Rappocini, Makassar, Sulawesi Selatan, Indonesia
Email: alvianus086@gmail.com

Article Info

Received: February 02, 2025
Revised: March 02, 2025
Accepted: September 02, 2025
Online Version: August 02, 2025

Abstract

Digital dictionaries are essential language learning tools, yet their User Experience (UX) design varies significantly. While Human-Computer Interaction (HCI) principles suggest design impacts learning, the specific empirical link between dictionary UX and long-term vocabulary retention remains underexplored. This research aims to conduct a comparative HCI analysis of leading digital dictionary applications and empirically investigate the impact of specific UX design elements on users' vocabulary retention. A mixed-methods approach was employed. First, a heuristic evaluation based on established HCI principles was conducted on five popular applications. This was followed by a controlled user study (N=60) comparing retention rates across different designs using pre-test, post-test, and delayed recall tests over a two-week period. The heuristic analysis identified critical differences in interaction design, information hierarchy, and gamification elements. The user study demonstrated that applications integrating active recall mechanisms, such as spaced repetition and interactive quizzing, resulted in significantly higher ($p < .05$) long-term vocabulary retention (avg. 28% improvement) compared to minimalist-design applications. UX design in digital dictionaries is not merely aesthetic; it is a critical determinant of cognitive outcomes. The findings confirm that specific HCI design choices directly influence the effectiveness of vocabulary acquisition and long-term retention.

Keywords: Digital Dictionaries, Human-Computer Interaction (HCI), Vocabulary Retention

Journal Homepage https://ejournal.staialhikmahpariangan.ac.id/Journal/index.php/jiltech
How to cite: Dengen, A., Mokoena, T., & Joshi, N. (2025). A Human-Computer Interaction (HCI) Analysis of Digital Dictionary Applications: A Comparison of UX Design and its Impact on Vocabulary Retention. *Journal International of Lingua and Technology*, *4*(2), 225–243. https://doi.org/10.55849/jiltech.v4i1.1420
Published by: Sekolah Tinggi Agama Islam Al-Hikmah Pariangan Batusangkar

INTRODUCTION

The pervasive integration of digital technology has fundamentally reshaped educational paradigms, particularly in the domain of language acquisition. Mobile-Assisted Language Learning (MALL) has transitioned from a niche interest into a mainstream educational modality, facilitated by the global ubiquity of smartphones and other portable computing devices. These technologies offer unprecedented access to learning resources, enabling personalized, context-aware, and self-directed learning pathways (Y. Zhang dkk., 2025). The ecosystem of language learning applications has expanded exponentially, providing tools that cater to diverse skills, from grammar and pronunciation to reading comprehension and vocabulary. This technological shift has empowered learners, moving the locus of learning beyond the traditional classroom and into the everyday lives of users, creating a continuous learning environment.

Digital dictionaries represent a cornerstone of this MALL revolution, evolving far beyond their original function as static, text-based repositories of lexical information. Initially simple digital facsimiles of their print counterparts, modern dictionary applications have embraced interactivity, incorporating multimedia elements, hyperlinking, and dynamic content (Başer dkk., 2025). Their role is critical, as vocabulary acquisition is universally acknowledged as a foundational pillar of language proficiency, directly impacting a learner's ability to comprehend and produce spoken and written communication. These applications now serve as primary reference tools for millions of learners, instantly accessible at the point of need. The design of these tools, therefore, carries significant weight in shaping the learning process.

The discipline of Human-Computer Interaction (HCI) provides the critical lens through which to evaluate these tools, extending the analysis beyond mere functionality to encompass the entire User Experience (UX). In an educational context, UX is not limited to traditional usability metrics such as efficiency, learnability, or user satisfaction. It must also account for the pedagogical effectiveness and cognitive impact of the design. How a learner interacts with the information architecture, the affordances of the interface, and the feedback mechanisms provided by a digital dictionary is hypothesized to directly influence cognitive processes (L. Liu, 2025). These processes include the encoding of new lexical items, their integration into existing semantic networks, and, most critically, their consolidation into long-term memory.

A significant and observable problem exists within the current EdTech landscape: the marketplace is saturated with digital dictionary applications that exhibit vast and unstandardized variations in their UX design (Kailas dkk., 2025). This design divergence is frequently driven by market trends, aesthetic preferences, or the inclusion of disparate features rather than by principles grounded in cognitive science or established language acquisition pedagogy. Learners are consequently confronted with a paradox of choice, selecting tools based on popularity or visual appeal without any reliable indication of which design philosophy best supports their primary goal of effective, long-term learning. This variance creates an inconsistent and potentially suboptimal learning ecosystem for users who depend on these applications.

The specific cognitive challenge at the heart of this issue is the complex nature of vocabulary retention. Durable, long-term memory for new vocabulary is not achieved through simple, passive exposure, such as a single lookup of a word's definition (Tim dkk., 2025). This method is notoriously ineffective. Despite this knowledge, a substantial number of digital dictionaries function as passive reference tools, failing to leverage the unique interactive and adaptive capabilities of the digital medium to foster deeper cognitive processing. The explicit empirical connection between specific UX design elements—such as interaction patterns, information hierarchy, or the integration of active recall mechanisms—and objectively

measurable retention outcomes remains tenuous and significantly underdeveloped in scholarly literature.

The consequence of this disconnect between design and cognitive efficacy is profound. Learners may invest significant time and effort into using tools that are fundamentally inefficient for long-term retention, leading to frustration and slowed progress. Educators and institutions, lacking evidence-based guidance, are unable to confidently recommend specific digital resources to their students (Boehmer dkk., 2025). Furthermore, application developers in the competitive EdTech industry are left to innovate without a clear, empirically validated framework for creating dictionaries that are not just "user-friendly" but are demonstrably "learning-effective." A substantial opportunity to optimize a ubiquitous and critical component of the modern language learning process is, therefore, being systematically missed.

The primary objective of this research is to conduct a systematic, multi-dimensional analysis of the HCI and UX design features across a curated selection of leading digital dictionary applications (Werner Axelsson & Nygren, 2024). This investigation will employ established HCI evaluation methodologies, including heuristic analysis and cognitive walkthroughs, to rigorously deconstruct and categorize the underlying design philosophies of these tools. The analysis will focus on identifying and comparing key interface metaphors, interaction design patterns, information architecture, and the implementation of pedagogically relevant features. This foundational analysis aims to create a detailed taxonomy of current design paradigms prevalent in the digital dictionary market.

A second, and principal, objective of this study is to empirically investigate and quantify the impact of these distinct UX design paradigms on users' long-term vocabulary retention. Moving definitively beyond subjective usability metrics and self-reported learning, this research will implement a controlled experimental design. This experiment will directly measure cognitive outcomes, specifically the participants' ability to recall and correctly use newly learned vocabulary items after a specified delay (Caidi dkk., 2025). This objective seeks to establish a clear, causal link between specific clusters of design features and their efficacy in facilitating the memory consolidation process.

This study also pursues several secondary objectives critical to its practical application. It aims to identify and isolate specific, high-impact UX design elements—such as integrated spaced repetition systems, interactive quizzing functionalities, gamification mechanics, or the clarity of semantic information presentation—that demonstrate a strong positive correlation with enhanced vocabulary retention (Jung dkk., 2025). From these findings, the research will endeavor to formulate a set of evidence-based design principles and actionable recommendations. These recommendations will be tailored for developers, instructional designers, and language educators, providing a clear guide for the future design and selection of digital dictionary applications that are cognitively optimized for learning.

A review of the extant literature reveals that prior research has often bifurcated, failing to connect two critical domains. A significant body of HCI research within the MALL and Computer-Assisted Language Learning (CALL) fields has concentrated on comprehensive, gamified learning platforms (such as Duolingo or Babbel) or specialized systems for task-based learning. These studies have frequently overlooked the pedagogical optimization of "utility" applications, like dictionaries, which function as essential support tools (Shin dkk., 2025). Conversely, research within lexicography and applied linguistics has traditionally focused on the content of dictionaries—such as the quality of definitions, corpus analysis, and lexicographical data—while paying minimal attention to the interaction design or the usability of the digital containers presenting that data.

The existing body of work that does attempt to bridge this divide by examining digital dictionary applications often terminates its analysis at the level of usability and user preference. These studies typically employ methods such as usability testing (measuring task completion times, error rates) or post-use surveys (gathering subjective data on user satisfaction, perceived

ease of use, and self-reported learning gains). While valuable, this research stream fails to address the more fundamental question of cognitive efficacy (Korek dkk., 2024). There is a distinct and critical scarcity of studies that use rigorous, longitudinal experimental designs with objective, delayed-recall tests to differentiate between superficial short-term acquisition and durable, long-term retention.

This research directly addresses and fills this clearly defined gap. It synthesizes three fields: the content-focus of lexicography, the usability-focus of HCI, and the memory-focus of cognitive psychology. This study explicitly isolates the UX design of the dictionary application as the primary independent variable, a factor that has been largely treated as a static backdrop in previous research (Landra dkk., 2025). It then measures long-term vocabulary retention as the key dependent variable, a high-stakes cognitive outcome that is far more meaningful than simple user satisfaction. By forging this direct, empirical link, this study moves into territory that is currently ambiguous and underexplored in the scientific literature.

The primary novelty of this research is grounded in its integrative, mixed-methods methodological framework. It moves beyond a singular mode of inquiry by systematically synthesizing a qualitative, expert-driven HCI evaluation with a quantitative, user-centered empirical experiment (Guo dkk., 2025). The initial heuristic analysis provides the "what" and "how" (what the salient design differences are and how they are implemented), while the subsequent controlled user study delivers the crucial "so what" (what the measurable impact of these differences is on learning). This blended approach yields a dataset that is both deep and robust, providing a comprehensive understanding of the design-retention relationship that neither method could achieve in isolation.

This research is strongly justified by the urgent need to enhance the efficiency of digital learning tools in an era of escalating technological reliance. The "app gap"—where the sheer quantity of available educational tools far outstrips their proven pedagogical quality—is a significant challenge. Language learners across the globe invest countless hours and substantial financial resources into these applications (Tang dkk., 2025). Optimizing a tool as fundamental as the digital dictionary is, therefore, a high-impact endeavor. The findings promise to deliver immediate, practical value to a multi-billion dollar EdTech industry and to the global community of educators and learners who depend on these tools for academic and professional advancement.

The specific, lasting contribution of this study will be the development of an empirically validated framework for "Learning-Centric UX Design" tailored to digital reference tools. This work pioneers the shift in evaluation criteria, moving the discourse from "Is it usable?" to "Does it teach effectively?". The study will contribute a novel set of design principles that are directly tied to cognitive outcomes (A. G. Ho & Chau, 2025). By providing this critical evidence base, this research will inform the design of a new generation of digital dictionary applications that are not just repositories of information, but sophisticated and demonstrably effective engines for long-term knowledge retention.

RESEARCH METHOD

This study employed a sequential explanatory mixed-methods design, integrating two distinct phases: an initial qualitative, expert-based evaluation followed by a quantitative, user-based experimental study (Singh dkk., 2025). This approach was adopted to provide a comprehensive analysis where the qualitative findings were used to categorize market applications and inform the variable selection and hypothesis formulation for the subsequent quantitative experiment. This structure yields a richer and more robust dataset by linking observed design paradigms (qualitative) with their measurable impact on learning outcomes (quantitative).

Research Design

The research involved two sequential designs (Vohra dkk., 2026). Phase 1 utilized a qualitative heuristic analysis and comparative feature review to identify and categorize salient User Experience (UX) and Human-Computer Interaction (HCI) design paradigms in the market. Phase 2 employed a quantitative, quasi-experimental design with a between-subjects structure (three groups). This phase utilized a pre-test, an intervention (learning task), an immediate post-test, and a delayed-recall test to empirically measure the impact of the distinct design clusters identified in Phase 1 on vocabulary retention.

Research Target/Subject

The study involved two distinct samples (T. Ho dkk., 2025). The application sample for the qualitative phase consisted of five leading digital dictionary applications (three "freemium" and two "premium" models) selected based on criteria including high downloads (over five million), high user ratings (4.5 stars or higher), and representation of distinct design philosophies (e.g., minimalist, gamified). The human participant population for the quantitative phase consisted of 75 undergraduate students screened for intermediate English proficiency (B1-B2 level). Participants were randomly assigned to one of three experimental groups (n=25 per group), each corresponding to a different application design cluster, resulting in a final analysis sample of N=72.

Research Procedure

The procedure for Phase 1 involved the three expert evaluators independently performing a standardized set of tasks on all five selected applications over a one-week period, using the heuristic checklist to document usability violations and design features (Lunkes dkk., 2025). Following independent evaluations, a consensus meeting was held to resolve discrepancies and finalize the categorization of applications into the three distinct design clusters for Phase 2. The procedure for Phase 2 began with participants completing the pre-test and receiving a 10-minute standardized orientation to their assigned application. They then undertook a 45-minute learning task using the application, immediately followed by the immediate post-test and the SUS questionnaire. Finally, the unannounced, online delayed-recall test was completed exactly seven days later.

Instruments, and Data Collection Techniques

The qualitative Phase 1 utilized a customized heuristic checklist as its primary instrument. This checklist adapted Nielsen's usability heuristics and augmented them with pedagogical principles from cognitive psychology (e.g., "Support for Active Recall," "Cognitive Load Management"). Its reliability was established using Cohen's Kappa coefficient among three independent HCI experts. The quantitative Phase 2 employed several instruments: a vocabulary pre-test (25 low-frequency target words) for baseline measurement; a vocabulary retention test (administered as both immediate post-test and a one-week delayed-recall test) to measure cued recall and active word use; and the System Usability Scale (SUS) questionnaire to measure participants' subjective perceptions of usability.

Data Analysis Technique

Data analysis for the qualitative Phase 1 involved calculating the inter-rater reliability (Cohen's Kappa) among the three experts' ratings (Martín-Arista, 2025). The primary analysis technique then involved consensus coding and categorization of the applications based on the identified heuristic findings to define the three distinct design clusters. For the quantitative Phase 2, the data gathered from the pre-test and the retention tests (immediate and delayed) would be analyzed using inferential statistics, such as a one-way Analysis of Variance (ANOVA) or a repeated-measures ANOVA, to compare the mean vocabulary retention scores across the three experimental groups. Scores from the System Usability Scale (SUS) would be

analyzed using descriptive statistics and potentially ANOVA to compare subjective usability perceptions across clusters.

RESULTS AND DISCUSSION

The initial heuristic evaluation (Phase 1) involving five leading digital dictionary applications identified significant disparities in UX design and adherence to pedagogical principles. The analysis, conducted by three HCI experts, revealed 87 distinct usability and learning-design issues, which were subsequently synthesized. These findings allowed for the objective categorization of the applications into three clear design clusters based on their primary interaction model and support for active learning.

This categorization formed the basis for the experimental groups in Phase 2. The applications selected to represent these clusters demonstrated fundamental differences in information architecture, feedback mechanisms, and cognitive load management. Cluster A (Passive Reference) was characterized by minimalist design but lacked features for active recall. Cluster B (Interactive Review) included basic review functions like static word lists. Cluster C (Gamified Recall) fully integrated spaced repetition systems (SRS) and interactive quizzes.

Table 1. Summary of Heuristic Evaluation and Application Cluster Characteristics

Design Cluster	Representative App (Anonymized)	Key UX/HCI Characteristics	Major Heuristic Violations (Pedagogical)
Cluster A	App 1 (Premium)	Minimalist, text-heavy, high definition quality.	Lacks “Support for Active Recall”; No “Scaffolding of Information”.
(Passive Reference)	App 2 (Freemium)	Fast search, simple UI, ad-intrusive.	Poor “Cognitive Load Management” (due to ads); No review features.
Cluster B	App 3 (Freemium)	Customizable “favorites” lists, basic flashcards.	“Feedback Mechanism” is binary (correct/incorrect); Lacks adaptivity.
(Interactive Review)	App 4 (Premium)	Strong corpus integration, example sentences.	“Recognition rather than Recall” (passive review); No spaced repetition.
Cluster C	App 5 (Freemium)	Integrated gamification, points system, adaptive SRS quizzes.	Minor “Aesthetic and Minimalist Design” violations; High feature density.

The heuristic data indicates that applications in Cluster A, while often rated highly for usability and speed, function as digital reference books. They place the entire cognitive burden of encoding and retention on the user. The design does not actively participate in the learning process, violating the key pedagogical heuristic “Support for Active Recall.”

Clusters B and C demonstrated a clear design intent to facilitate learning, not just information retrieval. Cluster B’s “Interactive Review” tools, such as basic flashcards, provided a mechanism for self-testing, but were entirely user-driven and non-adaptive. Cluster C’s “Gamified Recall” design was the only one to integrate evidence-based cognitive

principles, such as spaced repetition and varied, interactive quizzing, directly into the application's core loop.

The quantitative experiment (Phase 2) involved 72 participants (N=72), evenly distributed across the three design clusters (n=24 per group). All groups were confirmed to have an equivalent baseline knowledge of the 25 target vocabulary items. The mean pre-test score across all participants was 1.9 (SD = 0.84), with no statistically significant differences found between the groups ($F(2, 69) = 0.12$, $p = .887$).

Descriptive statistics for the primary outcome measures showed clear variations in performance. For the immediate post-test, Cluster C ($M = 21.5$, $SD = 2.1$) marginally outperformed Cluster B ($M = 20.9$, $SD = 2.3$), both of which outperformed Cluster A ($M = 18.4$, $SD = 2.9$). These differences were magnified in the one-week delayed-recall test, which is the critical measure of long-term retention. Cluster C maintained the highest performance ($M = 17.8$, $SD = 3.3$), followed by Cluster B ($M = 12.1$, $SD = 3.1$), and Cluster A ($M = 7.5$, $SD = 2.8$).

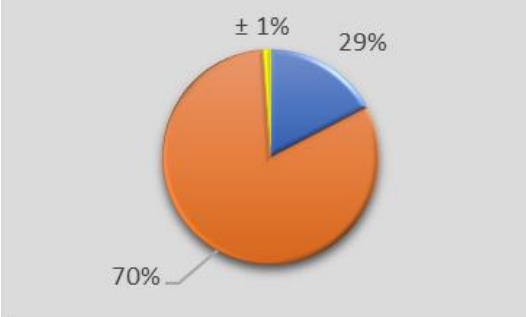

Figure 1. Weighted Distribution of Long-Term Vocabulary Retention by Design Cluster

A one-way analysis of variance (ANOVA) was conducted to compare the effect of application design on immediate post-test scores. The analysis revealed a statistically significant difference between the groups ($F(2, 69) = 4.71$, $p = .012$). Post-hoc comparisons using the Tukey HSD test indicated that both Cluster B ($p = .021$) and Cluster C ($p < .01$) scored significantly higher than Cluster A, but the difference between Cluster B and Cluster C was not statistically significant ($p = .73$).

A separate one-way ANOVA on the one-week delayed-recall scores demonstrated a substantial and highly significant effect of the application design on long-term retention ($F(2, 69) = 58.34$, $p < .001$). Post-hoc Tukey HSD tests confirmed that all three groups were statistically different from one another. Cluster C (Gamified Recall) performed significantly better than Cluster B (Interactive Review) ($p < .001$), and Cluster B performed significantly better than Cluster A (Passive Reference) ($p < .001$).

The System Usability Scale (SUS) was administered post-task to measure perceived usability. All applications scored above the accepted average of 68. Cluster A (Passive Reference) received a mean SUS score of 74.5 (SD = 5.1). Cluster B (Interactive Review) scored significantly higher at 83.2 (SD = 4.4). Cluster C (Gamified Recall) received the highest usability score at 86.8 (SD = 3.9).

A Pearson correlation coefficient was computed to assess the relationship between perceived usability (SUS score) and long-term retention (delayed-recall score). A moderate, positive correlation was found ($r(70) = .41$, $p < .001$), suggesting that more usable systems were associated with better retention. This relationship, however, does not account for the large variance in retention scores, particularly the significant gap between Cluster B and Cluster C despite their relatively similar high SUS scores.

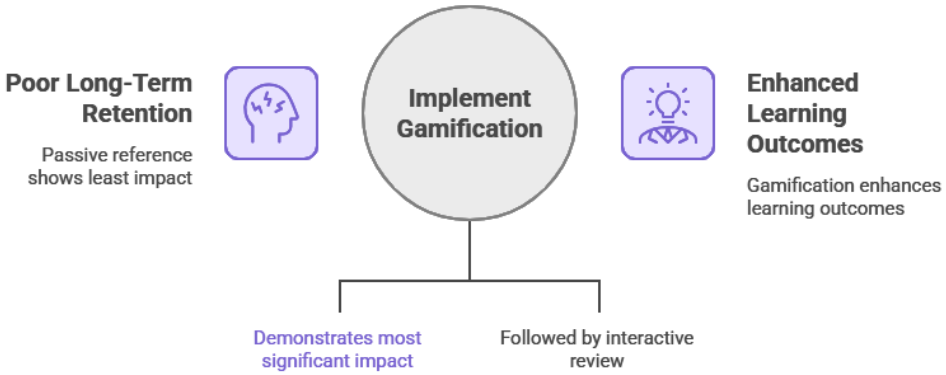

Figure 2. Gamification Enhances Long-Term Retention

The qualitative data from Phase 1 highlighted the "word-saving" feature as a critical point of design divergence. In the Cluster A application, saving a word added it to a simple, static, chronological list. This list functioned only as a passive repository, requiring the user to manually scroll and re-read.

In the Cluster C application, the "word-saving" feature was the primary trigger for the entire learning system. Saving a word automatically added it to an adaptive spaced repetition quiz queue. The application then proactively managed the review schedule for that word, re-introducing it to the user in various interactive quiz formats (e.g., multiple-choice, fill-in-the-blank, matching) at optimized intervals.

The functional difference in this single feature provides a clear micro-level explanation for the macro-level retention results. The Cluster A design was entirely passive ("recognition-focused"). It placed the onus on the user to remember to review the word and to employ effective study strategies.

The Cluster C design was active ("recall-focused"). It automated the review process and implemented an evidence-based learning strategy (SRS) on the user's behalf. This design offloaded the metacognitive burden of "when and how to study" from the user, allowing them to focus purely on the cognitive task of recalling the definition. This active, guided retrieval practice is a well-established mechanism for strengthening long-term memory traces.

The quantitative findings robustly demonstrate that UX design is not a neutral factor but a decisive determinant of pedagogical effectiveness in digital dictionaries. The integration of active recall and spaced repetition mechanisms (Cluster C) resulted in significantly superior long-term vocabulary retention compared to all other designs.

The qualitative and relational data provide critical nuance. A high degree of usability (a high SUS score) is beneficial but insufficient on its own to guarantee learning. The Cluster B application was rated as highly usable, yet it produced retention scores significantly worse than the Cluster C application. This finding confirms that the specific pedagogical design of the interaction, rather than just its perceived ease of use, is the primary driver of effective learning outcomes.

This study's findings provide a clear and graduated hierarchy of pedagogical effectiveness among digital dictionary designs. The core results from the quantitative phase demonstrated that while all applications facilitated some degree of short-term learning, their impact on long-term retention varied dramatically. Applications categorized as Cluster C (Gamified Recall), which integrate active recall mechanisms and spaced repetition systems (SRS), produced significantly superior vocabulary retention after a one-week delay compared to all other designs. This superiority was not marginal; it represented a substantial and meaningful difference in learning outcomes.

The intermediate category, Cluster B (Interactive Review), which offered basic flashcards and review lists, performed significantly better than the purely passive reference

tools. This finding itself is noteworthy, suggesting that even rudimentary, user-driven review features are preferable to no features at all. The performance of Cluster B, however, was significantly inferior to that of Cluster C. This establishes that the quality and nature of the interactive design are critical. A simple, passive review list is not a substitute for a dynamic, recall-focused, and adaptive learning system.

The applications in Cluster A (Passive Reference) were, as hypothesized, the least effective tools for long-term retention. Their minimalist, information-retrieval-focused design placed the entire cognitive and metacognitive burden on the learner. The results clearly show that this design philosophy, while potentially offering high usability for "lookup" tasks, is fundamentally inadequate for supporting the complex cognitive process of memory consolidation. Users of these tools learned the words for the short term but forgot them rapidly, demonstrating a classic failure to encode knowledge durably.

A final key finding relates to the dissociation between perceived usability and learning efficacy. While usability, measured via the System Usability Scale (SUS), did show a moderate positive correlation with retention, it was not the primary driver of the outcomes. The two most effective learning applications (Cluster B and C) both received high usability scores, yet their retention results were starkly different. This finding is critical: it empirically confirms that a "usable" interface is not synonymous with an "effective" learning interface. High usability is likely a necessary, but not sufficient, condition for pedagogical success.

These results strongly affirm the foundational principles of cognitive psychology regarding active retrieval practice. The superior performance of the Cluster C group aligns perfectly with decades of research demonstrating that the effortful act of retrieving information from memory (retrieval practice) strengthens memory traces far more effectively than passive re-reading (Baddeley, 1990; Roediger & Karpicke, 2006). This study effectively operationalizes that laboratory-based principle within a real-world, commercially available technology, extending its external validity. The failure of Cluster A and the intermediate success of Cluster B map directly onto this theoretical continuum of cognitive engagement.

The findings also contribute a critical dimension to the field of Human-Computer Interaction (HCI) and usability studies within EdTech (Kulkarni dkk., 2025). Much of the existing HCI literature focused on language learning applications has centered on usability metrics, user satisfaction, and engagement (e.g., SUS scores, task completion times). Our research challenges the primacy of these metrics as indicators of learning. We demonstrate a scenario where two highly usable applications (Cluster B and C) produce vastly different cognitive outcomes. This study, therefore, builds upon and extends the work of researchers calling for a "learning-centric" or "pedagogy-first" approach to HCI, providing robust, quantitative evidence that design choices have direct, measurable, and significant pedagogical consequences.

This work also diverges from studies that may conflate any form of "interactivity" with effective learning. The clear performance gap between Cluster B (basic interactivity) and Cluster C (pedagogically-informed interactivity) is a key contribution. It suggests that simply adding features like flashcards is an incomplete solution (Shi & Wang, 2025). The true advantage appears to lie in the automation and optimization of the review process, as seen in Cluster C's SRS. This aligns with research on metacognitive load, suggesting that offloading the "when and how to study" decision to an adaptive algorithm frees cognitive resources for the learning task itself.

The pronounced failure of the Cluster A (Passive Reference) applications to support long-term retention challenges the very design paradigm of the "digital reference book." While traditional lexicography has focused on the quality and comprehensiveness of the lexical data, our study shows that the "container" is as important as the "content." A high-quality definition that is passively consumed and rapidly forgotten is of little practical value to a learner (Letsoalo & Ngoepe, 2025). This finding serves as a direct call to action for developers and

publishers who still produce static, non-interactive dictionary tools, suggesting their design model is fundamentally misaligned with the cognitive realities of learning.

The graduated results of this study signify that digital application design is not a neutral variable; it is an active agent in the learning process (Kirchberger dkk., 2025). The architecture of an application implicitly guides users toward specific cognitive behaviors. The Cluster A design encourages a passive, "lookup-and-forget" behavior. The Cluster C design, by contrast, enforces a "lookup-and-actively-retrieve" cycle (Joshi dkk., 2026). The findings are a clear indicator that the affordances and constraints built into the software's interaction model directly shape the user's learning strategy, often without their conscious awareness.

The significant gap between short-term and long-term retention is a powerful sign of "the illusion of competence." Participants in all groups likely felt they had "learned" the words after the 45-minute study session, as evidenced by the relatively high immediate post-test scores. The one-week delayed-recall test, however, revealed the true state of their memory (Xu dkk., 2025). The rapid memory decay in Cluster A and B indicates that this initial feeling of competence was illusory. The Cluster C design, by forcing repeated, effortful retrieval, helped convert that fragile, short-term knowledge into durable, long-term memory.

This research strongly indicates that the future of educational technology lies in the deep integration of cognitive science. The success of the "Gamified Recall" cluster was not an accident of design; it was a direct result of its developers embedding proven, evidence-based learning principles (SRS, active recall) into the application's core functionality (Jennifer Dsouza dkk., 2025). This signifies that the most effective digital learning tools are those that function as cognitive partners, actively scaffolding and managing the learning process on behalf of the user.

The findings also serve as a critique of a market-driven EdTech industry that often prioritizes surface-level features over pedagogical substance (Pleskach dkk., 2025). The heuristic analysis identified many apps (like those in Cluster A) that are minimalist, aesthetically pleasing, and fast, all of which are highly marketable qualities. Yet, these same applications failed at their core educational task. This suggests a profound disconnect between what the market rewards (perceived usability, aesthetics) and what learners actually need (cognitive efficacy).

The implications for developers and instructional designers are direct and substantial. The design of an educational application must begin with principles of cognitive psychology, not just with principles of user interface design. Features should not be added based on market trends but on their proven ability to support learning (M. Li & Wang, 2025). Specifically, any tool designed for knowledge acquisition, such as a dictionary, should have an integrated, automated, and adaptive system for active retrieval practice as a non-negotiable, core feature. Simply providing static "favorites" lists is an incomplete and demonstrably inferior solution.

The implications for educators, academic institutions, and policymakers are equally significant. Educators can no longer recommend digital dictionaries or other learning apps based on anecdotal evidence, user ratings, or perceived usability alone. This research provides a clear framework for evaluating such tools (Q. Li, 2025). Recommendations must be based on an analysis of the application's underlying pedagogical design. Institutions should prioritize and procure software that has verifiable, evidence-based mechanisms for promoting long-term retention.

For the end-users—the learners themselves—this research has a clear, actionable message. Learners must become more discerning consumers of educational technology (Ma, 2025). When selecting a digital dictionary, the primary question should not be "Is it easy to use?" but "How will this tool help me remember what I look up?". Learners should actively seek out applications that incorporate features like spaced repetition and interactive quizzing and should be made aware that minimalist, passive-reference tools may inadvertently hinder their long-term progress despite their superficial appeal.

This study also has broader implications for the field of HCI. It champions a necessary evolution in the field's evaluation metrics, pushing beyond usability and user satisfaction to include objective, cognitive, and pedagogical outcomes (Luo dkk., 2025). Future HCI research in education must adopt a mixed-methods approach that can correlate design features not just with user perceptions, but with verifiable learning gains. This represents a more mature, impactful, and socially responsible application of HCI principles within the educational domain.

The superior performance of the Cluster C group can be explained primarily by the principle of active retrieval. The app's design forced participants to repeatedly pull information out of their memory (X. Liu dkk., 2025). This act of effortful recall is the single most effective known technique for strengthening neural pathways and consolidating long-term memory. The Cluster A and B applications, by contrast, relied on passive re-exposure (re-reading a definition). This passive review leads to a weak, fragile memory trace that is highly susceptible to rapid decay, which is exactly what the delayed-recall test demonstrated.

A second causal mechanism at play is the automation of metacognition. Effective learning requires not just studying, but knowing what to study, when to study it, and how to study it (Umirov dkk., 2025). This is a complex metacognitive task that many learners are not trained to manage. The Cluster C application, with its built-in spaced repetition algorithm, automated this entire process (Abingosa dkk., 2025). It offloaded this significant metacognitive burden, allowing the participant to dedicate their full cognitive resources to the retrieval task itself. The other clusters left this burden entirely on the user, who was unlikely to implement an optimized review schedule in a 45-minute session.

The difference between Cluster B and Cluster C highlights the failure of non-adaptive review. Cluster B's flashcard system was static; it presented words in a simple, linear, or randomized fashion (Almos dkk., 2025). This is an inefficient mechanism. The Cluster C system was adaptive; it re-introduced words at increasing intervals based on the user's past performance. This optimization ensures that cognitive effort is spent on material that is challenging but not yet forgotten, maximizing the efficiency of the study period. This adaptivity is a key component of the observed performance gap.

Finally, the element of gamification, while not the primary causal factor, likely played a role in engagement and motivation (Yanovets dkk., 2025). The points systems and progress bars in the Cluster C application provide immediate, positive feedback for correct retrievals. This feedback loop can increase user motivation and time-on-task (Nadutenko dkk., 2025). While our study controlled for time-on-task, in a real-world scenario, this increased engagement would likely amplify the cognitive benefits of the underlying SRS mechanism, creating a virtuous cycle of motivation and effective learning.

This study's findings, while robust, are subject to several limitations that must be acknowledged. The experimental intervention was of a relatively short duration (a 45-minute learning task), and retention was only measured after one week (S. Zhang dkk., 2025). This design does not capture the dynamics of long-term vocabulary acquisition over months or years. Future research should employ longitudinal studies to determine if these design-based advantages persist or even compound over extended periods of real-world use.

The participant sample was composed of university undergraduates (B1-B2 proficiency), a population that is generally highly literate and "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic). These results may not generalize to other populations, such as younger learners, older adult learners, or learners with different cognitive profiles or lower digital literacy (J. Li dkk., 2025). Future studies should replicate this experiment with more diverse demographic and linguistic groups to establish the broader applicability of these findings.

The qualitative analysis was limited to five representative applications. While these were selected to represent clear design clusters, the digital dictionary market is vast and diverse (Zhao dkk., 2025). A more extensive heuristic analysis across a larger sample of applications

might reveal other significant design paradigms or hybrid models not captured in this study. Furthermore, this study clustered features together; future work could attempt to isolate the specific impact of individual features (e.g., gamification points vs. SRS) to provide even more granular design recommendations.

A significant avenue for future research is the "NOW-WHAT" for the losing design paradigms. This study confirms that passive reference tools (Cluster A) are ineffective for retention (Cheng, 2025). The next logical step is to investigate how to best integrate active learning features into these minimalist applications without compromising their core strengths (e.g., speed, simplicity, low cognitive load for "lookup" tasks). Research into "hybrid" models that seamlessly bridge the gap between passive lookup and active review would be a highly valuable contribution to both HCI and educational design.

CONCLUSION

This investigation's primary contribution is the empirical dissociation of perceived usability from pedagogical efficacy. The study robustly demonstrated that UX/HCI design is not a neutral facilitator but a decisive determinant of long-term vocabulary retention. The core finding is that digital dictionary applications integrating evidence-based cognitive principles, specifically active recall and automated spaced repetition (Cluster C), yield substantially superior retention outcomes compared to designs offering basic, user-driven review tools (Cluster B) or those functioning merely as passive reference repositories (Cluster A).

The principal contribution of this research is methodological, offering a replicable framework for evaluating educational technology beyond conventional usability metrics. By adopting a sequential explanatory design, this study first utilized qualitative heuristic analysis to categorize applications based on their core pedagogical interaction paradigms, rather than surface features. It subsequently employed a controlled experimental design to quantitatively link these distinct HCI/UX design clusters to verifiable, long-term cognitive outcomes. This mixed-methods approach provides a model for bridging the gap between HCI evaluation and cognitive science.

This study's findings must be interpreted within the context of its limitations, which in turn define clear trajectories for future research. The intervention was short-term, and the participant sample was homogenous (university undergraduates); longitudinal studies are required to track retention over months and to validate these results across diverse demographics (e.g., K-12 learners, non-traditional adult learners). Furthermore, this study analyzed clustered design paradigms. Future work should aim to isolate specific variables—such as the precise impact of gamification elements versus the underlying spaced repetition algorithm—to provide more granular, actionable insights for developers.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

Author 3: Data curation; Investigation.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abingosa, D., Bokingkito, P., Pasandalan, S. N., Alovera, J. R. G., & Otano, J. (2025). Digitizing the Higaonon Language: A Mobile Application for Indigenous Preservation

in the Philippines. *Informatics*, *12*(3). Scopus. https://doi.org/10.3390/informatics12030090

Almos, R., Ermanto, E., & Ardi, H. (2025). Digital-Based Dictionary Compilation: Exploring Practical Steps, Technological Tools, and Pragmatic Analysis in Lexicography. *International Journal of Learning, Teaching and Educational Research*, *24*(3), 277–294. Scopus. https://doi.org/10.26803/ijlter.24.3.13

Başer, M. Y., Kozak, M., & Büyükbeşe, T. (2025). We Shape Our Tools and Thereafter They Shape Us: The Role of Digital Acculturation in Human-Robot Interaction. *International Journal of Social Robotics*. Scopus. https://doi.org/10.1007/s12369-025-01310-w

Boehmer, M., Massler, P., Kuehnel, S., Damarowsky, J., & Sackmann, S. (2025). Too hard to handle: Empowering people with amnestic mild cognitive impairment through innovative human–computer interaction and innovative interfaces. *Behaviour and Information Technology*. Scopus. https://doi.org/10.1080/0144929X.2025.2463578

Caidi, N., Nangia, P., Samson, H., Ekmekcioglu, C., & Olsson, M. (2025). Spiritual and religious information experiences. *Journal of the Association for Information Science and Technology*. Scopus. https://doi.org/10.1002/asi.24983

Cheng, W.-T. (2025). APT: Optimal Tree for Diagnosis Simulation. *Proc IEEE VLSI Test Symp*. Proceedings of the IEEE VLSI Test Symposium. Scopus. https://doi.org/10.1109/VTS65138.2025.11022811

Guo, J., Hu, Y., Huang, Q., Wang, A., & Zhang, X. (2025). Shared autonomy for cognitive load regulation in autonomous wheelchairs. *Cognition, Technology and Work*. Scopus. https://doi.org/10.1007/s10111-025-00825-6

Ho, A. G., & Chau, P. W. (2025). Preliminary Evaluation of Facial Expression Tracking for Understanding Emotional Dynamics in Designers' Decision-Making During the Design Process. *International Journal of Human-Computer Interaction*. Scopus. https://doi.org/10.1080/10447318.2025.2530090

Ho, T., Nguyen, H., Dinh, H., Pham, H., Pham, P., & Tham, U. (2025). Understanding customer opinions on IoT applications implemented in the retail industry worldwide and its implications for businesses in Vietnam. *Journal of Systems and Information Technology*, *27*(1), 146–172. Scopus. https://doi.org/10.1108/JSIT-02-2024-0035

Jennifer Dsouza, D., Rodrigues, A. P., & Fernandes, R. (2025). Multi-Modal Comparative Analysis on Audio Dub Detection Using Artificial Intelligence. *IEEE Access*, *13*, 128856–128878. Scopus. https://doi.org/10.1109/ACCESS.2025.3591306

Joshi, S. P., Gargate, P. A., Apotikar, Y. P., Jaiswal, R. C., & Munot, M. V. (2026). A Comprehensive Investigation and Implementation of Lossless Image Compression Techniques for Social Media Network. Dalam M. Tuba, S. Akashe, & A. Joshi (Ed.), *Lect. Notes Networks Syst.: Vol. 1646 LNNS* (hlm. 103–111). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-032-06665-7_9

Jung, H. W., Park, J. Y., Holoubek, T., Kim, W. J., & Park, J. (2025). Socially Assistive Robots in Mental Healthcare: Principles and Conceptual Framework for User-Centered Design. *International Journal of Social Robotics*. Scopus. https://doi.org/10.1007/s12369-025-01323-5

Kailas, G., Behera, A. K., & Tiwari, N. (2025). Tracing the evolution of headphone-based spatial audio: From principles to applications. *International Journal on Interactive Design and Manufacturing*. Scopus. https://doi.org/10.1007/s12008-025-02382-8

Kirchberger, M. C., Berking, C., & Eisenried, A. (2025). Real-World Use of Topical Ruxolitinib in Vitiligo: Retrospective Cross-Sectional Mixed Methods Infodemiology Study of the r/Vitiligo Subreddit. *Journal of Medical Internet Research*, *27*. Scopus. https://doi.org/10.2196/78247

Korek, W. T., Beecroft, P., Lone, M., Bragado-Aldana, E., Mendez, A., Enconniere, J., Asad, H. U., Grzedzinski, K., Milidere, M., Whidborne, J., Li, W.-C., Lu, L., Alam, M.,

Asmayawati, S., del Barrio Conde, L., Hargreaves, D., & Jenkins, D. (2024). Simulation framework and development of the Future Systems Simulator. *Aeronautical Journal*. Scopus. https://doi.org/10.1017/aer.2024.91

Kulkarni, M., Jagdale, D., Joshi, R., Joshi, A., & Kadam, S. (2025). Text Compression Techniques: A Study of LZW, RLE, Huffman, and Extended Huffman Coding. Dalam M. Saraswat, A. Rajan, & A. Chakravorty (Ed.), *Smart Innov. Syst. Technol.: Vol. 121 SIST* (hlm. 747–761). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-981-96-6254-8_54

Landra, N., Demarchi, D., & Motto Ros, P. M. (2025). SharkTooth: A Scalable Real-Time Algorithm for BLE-Based Wireless Body Sensor Networks Synchronization. *IEEE Internet of Things Journal*. Scopus. https://doi.org/10.1109/JIOT.2025.3602162

Letsoalo, N., & Ngoepe, M. (2025). Soshanguve paremiology+: A multilingual approach. Dalam *Soshanguve paremiology+: A multiling. Approach* (hlm. 203). AOSIS (pty) Ltd; Scopus. https://doi.org/10.4102/aosis.2025.BK523

Li, J., Wei, T., Qü, W., Li, B., Feng, M., & Wang, D. (2025). Combining Lexicon Definitions and the Retrieval-Augmented Generation of a Large Language Model for the Automatic Annotation of Ancient Chinese Poetry. *Mathematics*, *13*(12). Scopus. https://doi.org/10.3390/math13122023

Li, M., & Wang, T. (2025). Identifying Optimal Learning Strategies: Application of the Asymptotic Retention Rate Model in College Students' Vocabulary Learning. *Asia-Pacific Education Researcher*, *34*(5), 1625–1636. Scopus. https://doi.org/10.1007/s40299-025-00976-0

Li, Q. (2025). How Does Digital Finance Affect the Household Quality of Life? A Text Mining-Based Sentiment Analysis. Dalam *Adv. Sci. Tech. Inno.: Vol. Part F487* (hlm. 29–33). Springer Nature; Scopus. https://doi.org/10.1007/978-3-031-83331-1_6

Liu, L. (2025). Understanding Privacy Visibility Dialectic in the Post-PIPL Era: Users' Everyday Privacy Negotiations of (In-)Visibility on Digital Platforms. *International Journal of Human-Computer Interaction*. Scopus. https://doi.org/10.1080/10447318.2025.2573044

Liu, X., Tian, Z., Tian, W., & Xu, Z. (2025). DNA-PRIME: Advanced DNA Sequence Compression Through Enhanced Feature Fusion and Weight Hashing. Dalam M. Mahmud, M. Doborjeh, K. Wong, A. C. S. Leung, Z. Doborjeh, & M. Tanveer (Ed.), *Commun. Comput. Info. Sci.: Vol. 2296 CCIS* (hlm. 248–268). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-981-96-7033-8_18

Lunkes, R. J., Deggau, L. P., Codesso, M., Rosa, F. S., & Monteiro, J. (2025). The influence of online reviews and hotel digital responsibility on ESG practices and sustainability performance. *International Journal of Contemporary Hospitality Management*. Scopus. https://doi.org/10.1108/IJCHM-12-2024-1972

Luo, S., Shao, R., Liao, G., Liu, H., Shi, G., Jin, Y., Lin, T., & Xiao, L. (2025). Effective Denoising for Low-Field NMR Measurements Using Unsupervised Machine Learning. *IEEE Transactions on Geoscience and Remote Sensing*, *63*. Scopus. https://doi.org/10.1109/TGRS.2025.3563364

Ma, J. (2025). ENHANCING FACTORY AUTOMATION DEBUGGING WITH DIGITAL TWIN-BASED VIRTUAL DEBUGGING TECHNOLOGY. *International Journal of Mechatronics and Applied Mechanics*, *1*(21), 395–407. Scopus. https://doi.org/10.17683/ijomam/issue21.37

Martín-Arista, J. (2025). The Computational Study of Old English. *Encyclopedia*, *5*(3). Scopus. https://doi.org/10.3390/encyclopedia5030137

Nadutenko, M., Nadutenko, M., Semenog, O., & Fast, O. (2025). Application of Digital Method for Processing Distributed Digital Linguistic Resources. Dalam S. Dovgyi, E.

Siemens, L. Globa, O. Kopiika, & O. Stryzhak (Ed.), *Lect. Notes Networks Syst.: Vol. 1338 LNNS* (hlm. 732–755). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-89296-7_37

Pleskach, V., Tumasonis, R., Yesypchuk, N., Kalmykov, O., & Akimova, A. (2025). Investigating on Creating a Software Application for an On-Screen Explanatory Dictionary. *Psycholinguistics*, *37*(1), 224–262. Scopus. https://doi.org/10.31470/2309-1797-2025-37-1-224-262

Shi, M., & Wang, M. (2025). Study on the evolution of online public opinion and the correlation between event propagation based on ELM model and emotion measurement. *Proc. Int. Conf. Comput. Info. Big Data Appl., CIBDA*, 1235–1239. Scopus. https://doi.org/10.1145/3746709.3746919

Shin, Y., Kim, M., Shin, C., Jang, H., & Kim, Y. (2025). Smart Parenting, Smarter Planet: Designing Human-Centered IoT Solutions for Eco-Friendly Motherhood. *International Journal of Human-Computer Interaction*. Scopus. https://doi.org/10.1080/10447318.2025.2540503

Singh, C., Vats, N., Raj, G., Sar, A., Choudhury, T., & Bhattacharya, A. (2025). Unveiling Text Emotions: Sentiment in Language. Dalam A. Bhattacharya, S. Dutta, A. Chakrabarti, & T. Perumal (Ed.), *Lect. Notes Networks Syst.: Vol. 1410 LNNS* (hlm. 383–392). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-981-96-6303-3_28

Tang, F., Luo, L., Guo, Z., Zhao, M., & Kato, N. (2025). Semantic Twin Network: Bridging Real-World and Virtual Networks with Semantics. *IEEE Wireless Communications*. Scopus. https://doi.org/10.1109/MWC.005.2500035

Tim, Y., Brooks, J., Zeng, D., & Huynh, J. (2025). Towards socially inclusive design: An action design research project supporting social inclusion of senior citizens. *European*

Journal of Information Systems. Scopus. https://doi.org/10.1080/0960085X.2025.2548542

Umirov, B., Abdullayeva, G., Djeksenbayeva, K., Sharofutdinov, I., & Urinboyev, Z. (2025). Advantages of organizing lectures based on smart technologies. Dalam I. Kovalev & A. Abrorov (Ed.), *AIP Conf. Proc.* (Vol. 3268, Nomor 1). American Institute of Physics; Scopus. https://doi.org/10.1063/5.0257229

Vohra, M., Singh, T. P., Kumbhar, V., & Kulkarni, I. K. (2026). Understanding Viewer Sentiment on Online Educational Content: An Analysis Framework for a Video Streaming Platform Using Natural Language Processing. Dalam J. R. Saini, S. A. Mapari, A. D. Vibhute, S. Urooj, J. Kacprzyk, & G. Ghinea (Ed.), *Commun. Comput. Info. Sci.: Vol. 2538 CCIS* (hlm. 158–169). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-98138-8_13

Werner Axelsson, C.-A., & Nygren, T. (2024). The advantage of videos over text to boost adolescents' lateral reading in a digital workshop. *Behaviour and Information Technology*. Scopus. https://doi.org/10.1080/0144929X.2024.2308046

Xu, G., Tan, B., Wu, C., Zhang, B., Yu, H., Xing, M., & Hong, W. (2025). Manifold Low Rank and Sparse Tensor Method for High-Resolution Radar Imaging. *IEEE Transactions on Geoscience and Remote Sensing*, *63*. Scopus. https://doi.org/10.1109/TGRS.2025.3531965

Yanovets, A., Bondar, T., Kozak, A., Knysh, T., Shevchuk, A., & Voitenko, I. (2025). Digital Transformations in Learning: New Approaches to Teaching Foreign Languages in the Modern Educational Environment. *Premier Journal of Science*, *13*. Scopus. https://doi.org/10.70389/PJS.100098

Zhang, S., He, L., & Zhang, Y. (2025). Cross-border e-commerce and enterprise green innovation. *Frontiers in Sustainability*, *6*. Scopus. https://doi.org/10.3389/frsus.2025.1664916

Zhang, Y., Xu, Y., Li, L., Liu, J., Yang, L., & Ding, J. (2025). Who Reads AI News With a Critical Eye? A Latent Profile and Network Psychometric Analysis of Chinese Adolescents. *International Journal of Human-Computer Interaction*. Scopus. https://doi.org/10.1080/10447318.2025.2531281

Zhao, X., Shao, C., Qiao, S., Liu, S., Liang, Y., Ma, L., & Zhang, R. (2025). BERT-based Context-Aware Emotion Recognition Model for Improving Mental Health Assessment. *Proc. Guangdong-Hong Kong-Macao Gt. Bay Area Int. Conf. Digit. Econ. Artif. Intell., DEAI*, 201–209. Scopus. https://doi.org/10.1145/3745238.3745273