

The effect of LLM-powered value-integrated story cards on English vocabulary acquisition: A generative AI-supported convergent mixed methods study in middle school students

Halil Şentürk * 

Science and Art Center (BİLSEM), Ordu/Ünye, Türkiye

Suggested citation: Şentürk, H. (2026). The effect of LLM-powered value-integrated story cards on English vocabulary acquisition: A generative AI-supported convergent mixed methods study in middle school students. *Journal of Educational Technology and Online Learning*, 9(3), 397-409.

Highlights

- LLM-generated value-integrated story cards yielded statistically significant and practically large vocabulary gains in a K-12 EFL context (Cohen's $d = 1.52$).
- Qualitative content analysis identified four themes: contextual recall, motivation-enjoyment, value awareness, and material usability.
- A transparent multi-role GenAI integration model is presented for material development, test item generation, and preliminary qualitative coding.
- Convergent mixed methods design enabled triangulation of cognitive, affective, and behavioral dimensions of instructional effectiveness.
- Findings support the use of prompt-engineered, expert-validated GenAI tools as scalable resources for value-integrated foreign language material design.

Abstract

This study investigates the effectiveness of large language model (LLM)-powered value-integrated story cards in English vocabulary instruction using a convergent mixed methods design with 30 sixth-grade middle school students over a four-week intervention. Story cards embedding universal values (empathy, responsibility, honesty, and cooperation) were developed via systematic prompt engineering with GPT-4, a generative AI (GenAI) tool, reviewed by domain experts, and applied within a quasi-experimental pre-test/post-test control group design. Generative AI was additionally employed in the preliminary coding phase of qualitative data analysis and in generating the item pool for the Vocabulary Achievement Test. Data were collected via a Vocabulary Achievement Test (KR-20 = .83), structured classroom observation forms, and semi-structured individual interviews. Quantitative findings revealed a statistically significant and practically large difference in favour of the experimental group [$t(28) = 4.15$, $p = .001$, Cohen's $d = 1.52$]. Qualitative content analysis generated four main themes: contextual recall, motivation-enjoyment, value awareness, and material usability. Findings offer pedagogical implications for GenAI-supported, holistic, and value-integrated material design in foreign language education.

Article Info: Research Article

Keywords: *foreign language teaching, values education, story cards, vocabulary acquisition, context-based instruction*

1. Introduction

Vocabulary knowledge is widely recognized as a fundamental building block of foreign language proficiency. Adequate performance across listening, reading, speaking, and writing requires students to possess a broad and functional vocabulary repertoire (Nation, 2013; Schmitt & Schmitt, 2020). Research indicates that English language learners need to recognize at least 3,000-4,000 word families to achieve a

* Corresponding author. Science and Art Center (BİLSEM), Ordu/Ünye, Türkiye.

e-mail address: halilsenturk411@gmail.com

Doi: <http://doi.org/10.31681/jetol.1903748>

Received 6 Mar 2026; Accepted 3 Jun 2026

ISSN: 2618-6586. This is an open Access article under the CC BY license.



communicatively viable level in everyday language use (Nation, 2013). Despite this, studies conducted in Turkish middle school settings consistently reveal that the majority of students fall considerably short of this threshold (Tok & Yigin, 2013).

Current vocabulary instruction practices largely rely on repetition-based pedagogy in which words are presented in list form, decontextualized from meaningful use. Research has demonstrated that this approach impedes vocabulary learning and hinders transfer to long-term memory (Laufer & Hulstijn, 2001; Webb & Nation, 2017). In contrast, vocabulary presentation supported by contextual cues has been shown to substantially improve both learning speed and retention rates (Schmitt, 2010; Webb, 2020), making the embedding of words in meaningful, affectively rich contexts a pedagogical imperative.

The Turkish Ministry of National Education (MoNE, 2018) curriculum prioritizes the development of universal values such as love, respect, responsibility, helpfulness, and justice. Recent research affirms that values education aligns organically with the nature of language learning and can be systematically integrated into foreign language classrooms (Byram, 1997; Ghosn, 2013). Value-embedded narrative texts have been shown to enhance students' empathy and enable character education and language learning to mutually reinforce one another within the same instructional process (Mages, 2006; Lickona, 2009). Nevertheless, the existing body of research on values integration in foreign language teaching has been dominated by textbook analyses and critical discourse studies (Puspitasari et al., 2021; Pratiwi et al., 2023) rather than empirical intervention research. Recent explorations at the intersection of GenAI and values-based EFL content signal the pedagogical potential of this integration, yet they remain largely descriptive in scope (Lo, 2023). Controlled experimental studies that empirically test the effect of value-integrated instructional materials on measurable language outcomes such as vocabulary acquisition are, to the best of the authors' knowledge, virtually absent from the literature (Ghosn, 2013; Puspitasari et al., 2021). This gap is particularly pressing given that national curricula — including Turkey's MoNE (2018) framework — explicitly mandate the integration of universal values into foreign language instruction, creating an urgent need for evidence-based material designs that bridge language pedagogy, values education, and emerging GenAI capabilities.

This study aims to address this gap by investigating the effect of value-integrated story cards developed using LLM-based generative AI (GenAI) tools on sixth-grade students' English vocabulary acquisition through a convergent mixed methods design. Notably, GenAI was integrated not only in material development but also in test item pool generation and preliminary qualitative coding, making this study methodologically distinctive in the literature. The integration of generative AI into language teaching material development is considered an emerging domain with the potential to enhance both material quality and personalization capacity (Kasneji et al., 2023; Mollick & Mollick, 2023; Lo, 2023).

Recent systematic reviews indicate that while GenAI research in language education has grown exponentially since 2023, vocabulary acquisition has received comparatively less systematic attention and K-12 settings remain notably underrepresented in the literature (Li, 2025; Crompton, 2024). Nevertheless, emerging evidence suggests that LLM-powered tools hold considerable promise for vocabulary instruction: studies have demonstrated that generative AI tools can produce contextually rich, levelled vocabulary exercises and personalized word learning materials that surpass the decontextualized, list-based approaches prevalent in traditional instruction (Song & Song, 2023; Al Mahmud, 2023). Despite this potential, controlled experimental studies that empirically test the effect of LLM-generated materials on vocabulary acquisition outcomes in K-12 EFL contexts remain scarce, further underscoring the significance of the present study.

2. Theoretical Framework

The theoretical foundation of this study integrates five interrelated frameworks that together account for the cognitive, affective, social, and technological dimensions of value-integrated vocabulary instruction.

The first is Ausubel's (1963) theory of meaningful learning, which posits that new knowledge is most effectively acquired when it is linked to the learner's existing cognitive schemas. In the context of this study,

value-embedded stories provide the meaningful anchoring context through which target vocabulary is introduced, ensuring that words are not encountered in isolation but within a semantically and emotionally coherent narrative structure.

The second is Paivio's (1991) dual coding theory, which holds that simultaneous activation of verbal and visual representational systems deepens information processing and strengthens memory encoding. The story card format — featuring a target word and supporting visual on the front, and a narrative text on the back — was deliberately designed to activate both channels concurrently, maximizing encoding depth.

The third is self-determination theory (Deci & Ryan, 2000), which identifies intrinsic motivation as a primary driver of sustained learning. Value-laden narratives that resonate with students' personal experiences and moral sensibilities are hypothesized to support autonomy, relatedness, and competence — the three basic psychological needs identified by this theory — thereby sustaining engagement throughout the learning process.

The fourth is Byram's (1997) model of intercultural communicative competence, which frames language learning as inherently bound to values, attitudes, and cultural awareness. This framework justifies the systematic embedding of universal values into language instruction, positioning values education not as an add-on but as an organically integrated dimension of foreign language pedagogy.

The fifth is the emerging framework of GenAI integration into pedagogical material design and research processes (Kasneci et al., 2023; Mollick & Mollick, 2023; Lo, 2023). LLM-based systems such as GPT-4 are positioned in this study as tools that accelerate and personalize material development while requiring human expert oversight to ensure pedagogical appropriateness, cultural sensitivity, and alignment with curriculum goals.

Together, these five frameworks form a coherent theoretical architecture in which meaningful narrative context deepens cognitive encoding, multimodal material design activates dual representational channels, intrinsic motivation sustains engagement, values are treated as integral to language pedagogy, and GenAI tools serve as efficient yet expert-validated resources for material development.

3. Research Questions

This study seeks to answer the following research questions:

RQ1 (Quantitative): Do the post-test vocabulary achievement scores of the experimental group students taught with value-integrated story cards differ significantly from those of the control group?

RQ2 (Quantitative): How do value-integrated story cards affect students' active classroom participation levels compared to the control group?

RQ3 (Qualitative): How do students evaluate value-integrated story cards in terms of contextual recall, motivation, value awareness, and material usability?

4. Method

4.1. Research Design

The study employed a convergent mixed methods design (Creswell & Plano Clark, 2018), in which quantitative and qualitative data were collected simultaneously and integrated at the interpretation stage. This design was selected because neither quantitative nor qualitative data alone would be sufficient to capture the full scope of the research problem: while experimental data can establish whether a statistically significant vocabulary gain occurred, qualitative data are necessary to explain the cognitive and affective mechanisms through which story cards may have produced that effect. The convergent design enabled triangulation across data sources, thereby strengthening both the internal validity of the quantitative findings and the credibility of the qualitative interpretations.

On the quantitative dimension, a quasi-experimental pre-test/post-test control group design (Campbell & Stanley, 1963) was adopted. Random assignment of students to experimental and control groups was

employed to minimize selection bias, and pre-test equivalence was statistically verified prior to the intervention [$t(28) = 0.12, p = .906$], confirming group comparability. This design was preferred over a true experimental design given the naturalistic school setting, in which full randomization across classrooms was not feasible. On the qualitative dimension, semi-structured interviews and structured observations were employed to obtain an in-depth examination of student perspectives and classroom behaviors, providing interpretive depth that complements the experimental outcomes.

4.2. Participants and Ethics

The study was conducted with 30 sixth-grade students (14 female, 16 male; M age = 11.7, $SD = 0.4$) enrolled in a state middle school in the Black Sea region of Turkey. Purposive sampling was employed in school selection, as the researcher had obtained prior institutional access permission and could therefore ensure full administrative cooperation throughout the intervention — a practical prerequisite for quasi-experimental research in naturalistic school settings (Fraenkel et al., 2019). Within the selected school, students were randomly assigned to two equal groups: an experimental group ($n = 15$) and a control group ($n = 15$).

Although a sample of 30 participants is modest, it is consistent with the scope of quasi-experimental intervention studies conducted in single-school EFL contexts at the middle school level, where access constraints and ethical considerations related to minors typically limit sample size (Creswell, 2014). The pre-test equivalence of the two groups was statistically confirmed [$t(28) = 0.12, p = .906$], indicating that any post-test differences can be attributed to the intervention rather than to pre-existing differences in vocabulary knowledge. For the qualitative dimension, all 15 experimental group students participated in semi-structured interviews, ensuring that the full range of participant perspectives was represented in the qualitative data.

Written informed consent was obtained from all parents prior to data collection, institutional permission was secured from school administration, and participant anonymity was ensured throughout through the use of participant codes.

4.3. Instructional Material Design

The value-integrated story cards were an original instructional material developed by the researcher with support from a GPT-4-based generative AI tool. The design process followed five phases: (1) Needs analysis: the target vocabulary and values framework in the MoNE (2018) sixth-grade English curriculum were reviewed. (2) GenAI-supported content generation: via structured system prompts, GPT-4 was instructed to generate short English stories at the CEFR A2 level centered on a universal value (empathy, responsibility, honesty, cooperation, patience) for each target word, using a template incorporating context specification, constraint definition, and sample output demonstration. GenAI outputs were treated as raw material and revised by the researcher for pedagogical appropriateness, language level, and value consistency. Each card featured the target word and a supporting visual on the front, and the GenAI-supported story text on the back. A total of 20 cards were produced. (3) Expert review: two English language education specialists and one values education specialist independently evaluated the cards for content validity, language level, and value consistency (agreement rate = 91%), serving as a critical quality-assurance step to control for potential bias and cultural inappropriateness in GenAI outputs. (4) Pilot and revision: word selection, contextual fit, and naturalness issues in four cards were resolved. (5) GenAI-supported item pool generation: GPT-4 was prompted to generate CEFR A2-level multiple-choice items for the Vocabulary Achievement Test; outputs were reviewed by a measurement specialist to form the item pool, from which final test items were selected by the researcher.

4.4. Data Collection Instruments

4.5. Vocabulary Achievement Test (VAT)

The researcher-developed VAT consists of 20 multiple-choice items measuring both direct word knowledge (50%) and contextual use ability (50%). Content validity was established through a panel of two subject matter experts and one measurement specialist (average content validity ratio = .87; Lawshe, 1975). Internal

consistency reliability was $KR-20 = .83$ with a pilot group of 30 students. The test is scored out of 100 points.

4.6. *Structured Observation Form*

The observation form covers four behavioral categories: question asking/answering, story enactment/role play, contextual word use, and peer interaction. Each category was recorded on a frequency basis. Observations were conducted by the researcher and an independent observer; inter-rater reliability was Cohen's kappa = .84.

4.7. *Semi-Structured Interview Form*

The interview form consists of five open-ended questions covering material comprehensibility, enjoyment level, contribution to value awareness, contextual recall, and material usability. Item appropriateness was reviewed by two qualitative research experts; following pilot interviews, two items were reworded. Interviews were conducted individually, audio-recorded, and transcribed by the researcher (average duration: 12-15 minutes).

4.8. *Procedure*

The intervention followed a four-phase process. In Week 1, the VAT was administered as a pre-test to both groups. During Weeks 2 and 3 (eight class hours total), experimental group lessons were structured around story cards, incorporating in-context word discovery, role play, group discussion, and enactment activities. The control group used the course book, word list exercises, and direct instruction with the same target words. In Week 4, the post-test was administered and individual interviews with experimental group students were conducted. Both groups received equal weekly class hours (4 hours) and studied an identical vocabulary set.

4.9. *Data Analysis*

Quantitative data were analyzed using SPSS 28. An independent-samples t-test was selected to compare post-test scores between the experimental and control groups, as this test is appropriate for comparing two independent groups on a continuous dependent variable when group variances are homogeneous (Fraenkel et al., 2019). Prior to conducting inferential analyses, the statistical assumptions underlying the independent-samples t-test were systematically examined. Normality of score distributions was assessed using the Shapiro-Wilk test, which is recommended for samples of $n < 30$ (Field, 2018). Results indicated that post-test scores were normally distributed for both the experimental group ($W = .951$, $p = .534$) and the control group ($W = .944$, $p = .449$). Homogeneity of variance was examined using Levene's test, which yielded a non-significant result ($F = 1.87$, $p = .182$), confirming that the equal variances assumption was met. These results justify the use of the independent-samples t-test for between-group comparison. A paired-samples t-test was used to assess within-group pre-test/post-test change, enabling evaluation of learning gains within each group independently. The significance threshold was set at $\alpha = .05$. Cohen's d effect size was calculated to assess practical significance beyond statistical significance ($d < .20$ = negligible, $.20-.49$ = small, $.50-.79$ = medium, $\geq .80$ = large; Cohen, 1988), as effect size provides a more informative and replicable indicator of intervention impact than p -values alone. To address the risk of effect size inflation associated with small samples, a bootstrap confidence interval (1,000 resamples) was calculated for Cohen's d . The bootstrapped 95% CI for d was $[0.84, 2.19]$, confirming that the observed effect size falls within a plausible range and is unlikely to be an artefact of sample size (Field, 2018). Observation data were analyzed through descriptive frequency analysis to document behavioral patterns across the intervention period.

Qualitative data were subjected to content analysis following Miles, Huberman, and Saldana's (2014) three-stage model: data condensation, data display, and drawing and verifying conclusions. Content analysis was selected over thematic analysis or grounded theory because the study's research questions were directed at specific, pre-identified dimensions of student experience (contextual recall, motivation, value awareness, and material usability) rather than at the generation of new theory from raw data. Trustworthiness was established through multiple strategies: expert code comparison between two independent coders

(agreement = 87%), member checking through participant verification of interview transcripts, and data triangulation across quantitative, qualitative, and observational data sources (Lincoln & Guba, 1985).

A GenAI-assisted preliminary coding procedure was additionally applied: interview transcripts were submitted to GPT-4 with a structured system prompt to generate suggestions for potential themes and codes. Critically, these outputs were produced before human coders began independent coding, and coders did not access GenAI suggestions until after completing their own analysis. Final coding decisions were made exclusively by human coders; GenAI outputs served only as a post-hoc comparison point. This procedure was designed to ensure that GenAI assistance enhanced rather than replaced human analytical judgment, and its role is explicitly reported here in accordance with emerging transparency standards for GenAI use in qualitative research (Lincoln & Guba, 1985; Lo, 2023).

5. Results

Quantitative and qualitative findings are presented separately in line with the research questions, followed by convergent interpretation.

5.1. Vocabulary Achievement: Pre-Test / Post-Test Comparison (RQ1)

Table 1. Vocabulary Achievement Test Descriptive Statistics and Between-Group Comparison

Group	Pre-Test M	Pre-Test SD	Post-Test M	Post-Test SD	Gain M	Gain %
Experimental	50.13	6.31	85.40	5.14	+35.27	70.3%
Control	49.80	6.09	60.20	7.42	+10.40	20.9%
Post-test: $t(28)$ = 4.15, p = .001, Cohen's d = 1.52 (large)						

Note. M = Mean; SD = Standard deviation; $Gain$ = post-test minus pre-test mean. Cohen's $d \geq .80$ = large effect (Cohen, 1988).

As shown in Table 1, the groups were statistically equivalent on the pre-test [$t(28) = 0.12$, $p = .906$]. Post-test comparison revealed that the experimental group performed significantly higher than the control group [$t(28) = 4.15$, $p = .001$, Cohen's $d = 1.52$]. The d value of 1.52 indicates large practical significance. The 70.3% improvement rate in the experimental group is consistent with the theoretical model positing that words learned in meaningful contexts form stronger associative networks in long-term memory (Hulstijn, 2015).

5.2. Classroom Participation Levels (RQ2)

Table 2. Weekly Mean Participation Frequency per Structured Observation Form

Participation Behavior	Experimental (M)	Control (M)	Ratio
Question asking / answering	7.53	3.07	x2.45
Story enactment / role play	4.80	0.20	x24.0
Contextual word use	5.33	1.40	x3.81
Peer interaction (discussion)	3.87	0.93	x4.16

Note. M = Mean observed behaviors per lesson over 4-week intervention. Ratio = experimental M / control M .

Table 2 shows that experimental group students exhibited markedly higher active participation across all categories. The most striking difference was in story enactment and role play (ratio = x24.0), suggesting that value-based stories provide a powerful foundation for dramatization and embodied learning. The increase in peer interaction (x4.16) implies that the social dimension of vocabulary instruction can be supported through material design, consistent with self-determination theory (Deci & Ryan, 2000).

5.3. Student Perceptions (RQ3)

Table 3. Qualitative Content Analysis: Themes, Sub-codes, and Sample Quotes

Theme	Sub-codes	Sample Quote
Contextual Recall	Plot-based association, character linkage, emotional anchor	"I remember the word with the scene in the story, not alone." (S7)
Motivation and Enjoyment	Curiosity arousal, desire to learn, willingness to participate	"I was curious about the end of the story so I learned the words faster." (S12)
Value Awareness	Moral context, depth of meaning, real-life connection	"I am not only learning English; it also makes me think about being a good person." (S3)
Material Usability	Portability, visual support, ease of review	"When I see the picture, the word comes to mind." (S9)

Note. *S* = Student code. Table reflects content analysis findings from semi-structured interviews with experimental group students.

Under the Contextual Recall theme, students reported associating target words with mental images through story plot, character actions, and emotional moments, indicating verbal-visual integration as anticipated by dual coding theory (Paivio, 1991).

"When I think of kindness, I always picture the child holding the old woman's hand. There is no way I can forget the word." (S7)

Under the Motivation and Enjoyment theme, the majority of students reported that stories stimulated curiosity and made vocabulary learning enjoyable, consistent with self-determination theory (Deci & Ryan, 2000).

"I am curious about the end of the story so my attention does not wander. The words stay in my mind on their own." (S12)

Under the Value Awareness theme, students stated that the moral messages in the stories rendered vocabulary learning personal and meaningful (Lickona, 2009; Ghosn, 2013).

"I am not only learning English; when the story ends I think about myself too. That is why the word sticks better." (S3)

Under the Material Usability theme, students consistently highlighted three interconnected qualities of the story cards: portability, visual support, and ease of review. Students reported that the physical, card-based format allowed them to study independently outside the classroom and revisit target words at their own pace — a feature several participants described as making review feel considerably less burdensome than traditional word lists. Crucially, the visual element on the front of each card appeared to function as a retrieval cue that simultaneously activated the associated story context, suggesting that the material design effectively supported dual-channel encoding consistent with dual coding theory (Paivio, 1991).

"When I see the picture, the word comes to mind — and then the whole story comes too. It is much easier than trying to remember from a list." (S9).

5.4. Convergent Interpretation

Quantitative findings (statistically significant vocabulary gains and participation ratios) and qualitative findings (themes of contextual recall, motivation, value awareness, and material usability) together form a mutually reinforcing convergent pattern across three analytical dimensions. On the cognitive dimension, the statistically significant and practically large vocabulary gains observed in the experimental group (post-test: $M = 85.40$, $SD = 5.14$; Cohen's $d = 1.52$) are directly corroborated by the Contextual Recall theme, in which students described forming episodic memory traces linking target words to specific story moments, characters, and moral resolutions. This convergence suggests that the story card format did not merely

expose students to new words but created the conditions for deep, associative encoding consistent with meaningful learning theory (Ausubel, 1963) and dual coding theory (Paivio, 1991). On the affective and motivational dimension, the pronounced participation ratios observed across all four behavioral categories — most strikingly in story enactment (x24.0) — align directly with the Motivation and Enjoyment theme, in which students reported that narrative curiosity sustained their attention and reduced the cognitive burden of vocabulary learning. The Value Awareness theme further deepens this picture: students reported that the moral dimension of the stories rendered vocabulary personally meaningful, suggesting that values integration functioned not only as a character education strategy but as an affect-driven encoding mechanism. On the material design dimension, the Material Usability theme — highlighting portability, visual support, and ease of review — converges with the participation data to suggest that the physical and visual properties of the cards actively supported both in-class engagement and independent out-of-class review. Taken together, these convergent patterns indicate that the instructional effectiveness of value-integrated story cards operates simultaneously at cognitive, affective, and behavioral levels, and that neither quantitative nor qualitative data alone would have been sufficient to capture this multi-dimensional effect — thereby validating the convergent mixed methods design adopted in the present study.

6. Discussion

The findings demonstrate that value-integrated story cards exert a statistically significant and practically large effect on vocabulary acquisition. This result suggests that embedding target words within affectively rich, value-laden narratives creates a richer associative network in memory than isolated word presentation — a pattern consistent with the contextual learning framework, in which meaningful contexts facilitate associative retrieval (Schmitt & Schmitt, 2020; Nation, 2013). The 70.3% improvement rate observed in the experimental group indicates that the simultaneous activation of emotional, visual, and linguistic channels may have substantially deepened encoding. This interpretation aligns with dual coding theory (Paivio, 1991), as each card functioned as a visually-verbally-affectively integrated stimulus. Recent neuroimaging evidence lends further weight to this explanation: Zhou and Wu (2024), employing EEG-based measurement, demonstrated that internally generated contextual cues significantly activate memory and language processing areas during vocabulary learning. Similarly, Li et al. (2025) found that story-based contexts produced stronger word learning and retention outcomes in EFL learners compared to decontextualized conditions.

The pronounced difference in participation data across all four behavioral categories is particularly noteworthy. The x24.0 ratio in story enactment suggests that value-based narratives provided students with a genuine reason to use language — not merely to complete a task, but to embody a character and negotiate meaning around moral dilemmas. In the author's view, this social-emotional dimension of the instructional design may be as pedagogically significant as the cognitive one, since it transformed vocabulary learning from a passive reception activity into an active, identity-engaged process. This is consistent with research reporting that classroom interaction activates motivational processes and enhances cognitive processing capacity (Dornye & Ushioda, 2021; Wentzel, 2014), and is further supported by a recent systematic review identifying narrative-based approaches as consistently associated with heightened engagement and task participation among young EFL learners (Lena & Nikolov, 2025).

Qualitative findings revealed that students spontaneously formed emotional anchors between target words and story moments — recalling words through character actions, plot scenes, and moral resolutions rather than through definitions. This is a finding the author considers central to the study's contribution: it suggests that value integration does not merely add an ethical dimension to language learning, but actively restructures the cognitive encoding process by creating episodic memory traces. This converges with the affective filter hypothesis (Krashen, 1982), supporting the view that low anxiety and high motivation facilitate acquisition. Sembiring and Simajuntak (2023) similarly demonstrated that storytelling-based vocabulary instruction significantly outperformed traditional approaches, with learners reporting greater engagement and motivation — a pattern directly mirrored in the qualitative themes of the present study.

The findings extend those reported by Ghosn (2013) and Cameron (2001) regarding the function of children's literature in language learning, empirically demonstrating that value-embedded stories constitute cognitively and affectively powerful learning environments. Crucially, the present study advances this line of research by showing that GenAI can serve as a viable and efficient tool for generating such materials at scale, without sacrificing pedagogical quality when expert review is integrated into the process. A recent study published in *Computers and Education: Artificial Intelligence* (2025) found that AI-generated storytelling produced significant improvements in vocabulary retention and learner engagement among EFL learners, corroborating this conclusion. The integration of GenAI into material development, test item generation, and qualitative preliminary coding thus represents not only a methodological contribution but also a replicable model for researchers and practitioners seeking to combine technological efficiency with pedagogical intentionality (Kasneci et al., 2023; Mollick & Mollick, 2023; Lo, 2023). Complementing this, research conducted in EFL classroom contexts has highlighted that AI tools hold significant instructional promise while simultaneously raising ethical and pedagogical challenges that require structured institutional support — a finding that underscores the importance of the expert-validated, human-in-the-loop approach adopted in the present study (Tumen Akyildiz, 2025).

7. Limitations

This study has five core limitations. First, the study was conducted in a single state middle school with a sample of 30 students, restricting generalizability (Creswell, 2014). Second, the four-week intervention is insufficient for evaluating long-term vocabulary retention. Third, researcher-conducted observations do not entirely rule out observer effect despite measured inter-rater reliability (Miles et al., 2014). Fourth, material content was limited to five value themes, leaving unexplored the contributions of other value categories. Fifth, the potential biases and cultural assumptions embedded in GPT-4 outputs could not be systematically evaluated despite expert review, indicating that GenAI-assisted material development processes require additional validation for reliability and cultural sensitivity.

8. Conclusions and Implications

This study makes three original contributions to the field of educational technology and online learning. First, it provides the first empirical evidence — to the authors' knowledge — that LLM-powered, value-integrated instructional materials (story cards generated via GPT-4 prompt engineering) produce statistically significant and practically large gains in English vocabulary acquisition in a K-12 EFL context (Cohen's $d = 1.52$), demonstrating that generative AI can serve as a reliable tool for evidence-based material development beyond higher education settings. Second, it introduces and empirically tests a novel instructional material format — value-integrated story cards — that simultaneously targets vocabulary acquisition, character education, and learner engagement, offering foreign language educators a replicable, curriculum-aligned design model. Third, it presents a transparent, multi-role GenAI integration model in which LLM-based tools are employed across three distinct phases of an educational research study: instructional material development, test item pool generation, and preliminary qualitative coding — each with explicit expert validation procedures, thereby advancing methodological standards for GenAI use in educational research.

8.1. Implications for Practice

The findings carry several implications for educational technology and online learning theory and practice. First, this study demonstrates that LLM-based generative AI tools can function as effective, scalable resources for instructional material development in K-12 foreign language education — a context underrepresented in the ET/OL literature (Li, 2025; Crompton, 2024). The systematic prompt engineering process employed in this study offers practitioners a replicable model for leveraging GenAI to produce pedagogically appropriate, curriculum-aligned materials without sacrificing quality, provided that expert validation is embedded in the workflow. Foreign language teachers are therefore encouraged to explore GenAI-assisted material design as a complement to — rather than a replacement for — professional pedagogical judgment. Second, the story card format represents a low-technology, high-engagement

instructional design that can be readily adapted to both face-to-face and blended learning environments. The portability and ease of review highlighted by students under the Material Usability theme suggest that physical, card-based materials may offer affordances that screen-based alternatives do not fully replicate — a consideration relevant to EdTech designers balancing digital and analogue learning tools. Third, the integration of GenAI into qualitative research processes (preliminary coding) demonstrated in this study offers a methodologically transparent model for educational researchers seeking to incorporate AI assistance into data analysis while preserving human analytical primacy. Development of institutional guidelines and GenAI literacy programs for teachers and researchers is strongly recommended to ensure responsible and effective use.

8.2. Directions for Future Research

Several directions for future research emerge from this study. First, studies with larger samples spanning different school types, socioeconomic levels, grade levels, and geographical contexts are needed to establish the generalizability of these findings and to examine whether the effect of value-integrated, LLM-generated materials varies across learner populations.

Second, longitudinal designs are needed to evaluate the durability of vocabulary retention and to assess whether the value awareness outcomes observed in qualitative data translate into sustained behavioral or attitudinal change over time.

Third, comparative studies examining different GenAI tools (e.g., GPT-4 versus open-source LLMs), different prompt engineering frameworks, and different levels of expert intervention in the material review process would contribute to the development of evidence-based standards for GenAI-assisted material design in ET/OL contexts.

Fourth, research investigating the online and blended learning applications of value-integrated story cards — including digital card formats, gamified platforms, and AI-powered adaptive delivery — would extend the findings of this study to the online learning domain and address a significant gap in the current ET/OL literature.

Fifth, studies examining the methodological role of GenAI in educational research — including GenAI-human coder consistency across different qualitative traditions, prompt design effects on coding outcomes, and transparency reporting standards — will make valuable contributions to research methodology in the ET/OL field.

Ethics Statement

Ethical approval was obtained from Atatürk University Social and Humanities Sciences Ethics Committee, Education Sciences Unit Ethics Board (Meeting No: 02, Decision No: 11, Date: 22.01.2026).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

Acknowledgements

The author thanks the administration and students of BİLSEM Ordu/Ünye for their cooperation and participation.

References

- Al Mahmud, M. (2023). Generative AI tools and their impact on EFL vocabulary and grammar learning. *TESOL Journal*, 14(4), e734.
- Arnold, J. (2011). Attention to affect in language learning. *Anglistik: International Journal of English Studies*, 22(1), 11-22.
- Ausubel, D. P. (1963). *The psychology of meaningful verbal learning*. Grune & Stratton.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). Guilford Press.
- Brewster, J., Ellis, G., & Girard, D. (2002). *The primary English teacher's guide* (New ed.). Penguin English.
- Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. Multilingual Matters.
- Cameron, L. (2001). *Teaching languages to young learners*. Cambridge University Press.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Crompton, H. (2024). Artificial intelligence in English language teaching and learning: A systematic review. *Language Teaching Research*. <https://doi.org/10.1177/13621688241227693>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE.
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01
- Dornye, Z., & Ushioda, E. (2021). *Teaching and researching motivation* (3rd ed.). Routledge.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed.). McGraw-Hill.
- Ghosn, I. K. (2013). Four good reasons to use literature in primary school ELT. *ELT Journal*, 56(2), 172-179. <https://doi.org/10.1093/elt/56.2.172>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26. <https://doi.org/10.3102/0013189X033007014>
- Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

- Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1-26. <https://doi.org/10.1093/applin/22.1.1>
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575.
- Lena, M. S., & Nikolov, M. (2025). A systematic review of empirical studies on young EFL learners' motivation and task engagement. *SAGE Open*, 15(2). <https://doi.org/10.1177/21582440251386566>
- Li, B. (2025). A systematic review of empirical generative AI research in language learning and teaching from 2023 to 2024. *Computers & Education: Artificial Intelligence*, 8, 100352.
- Li, Z., Wu, S., & Liu, D. (2025). The effects of context, story-type, and language proficiency on EFL word learning and retention from reading. *International Review of Applied Linguistics in Language Teaching*, 63(2), 1507-1531. <https://doi.org/10.1515/iral-2023-0155>
- Lickona, T. (2009). *Educating for character: How our schools can teach respect and responsibility*. Bantam Books.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Mages, W. K. (2006). Drama and imagination: A cognitive theory of drama's effect on narrative understanding and story internalization. *Research in Drama Education*, 11(3), 329-340.
- Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. Jossey-Bass.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE.
- Ministry of National Education (MoNE). (2018). *English language curriculum (Primary and middle school grades 2-8)*. Board of Education and Discipline.
- Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. SSRN. <https://doi.org/10.2139/ssrn.4475995>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
- Nation, I. S. P. (2022). *Making and using word lists for language learning and testing* (2nd ed.). John Benjamins.
- Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle Cengage Learning.
- OpenAI. (2023). GPT-4 technical report. *arXiv*. <https://doi.org/10.48550/arXiv.2303.08774>
- Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3), 255-287. <https://doi.org/10.1037/h0084295>
- Pratiwi, Y., Andajani, K., Suyitno, I., Ismail, A., & Prastio, B. (2023). Representing and implementing moral values in foreign students in Indonesia textbooks for learners other than Indonesians. *International Journal of English Language and Literature Studies*, 12(1), 1-12.
- Puspitasari, D., Widodo, H. P., Widyaningrum, L., Allamnakhrah, A., & Lestariyana, R. P. D. (2021). How do primary school English textbooks teach moral values? A critical discourse analysis. *System*, 96, 102397. <https://doi.org/10.1016/j.system.2020.102397>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
- Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching* (2nd ed.). Cambridge University Press.
- Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1116>

- Sembiring, E., & Simajuntak, D. C. (2023). Digital storytelling to enhance vocabulary knowledge. *Journal of English Language Teaching and Applied Linguistics*, 5(1), 12-21.
- Song, C., & Song, Y. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *System*, 118, 103149. <https://doi.org/10.1016/j.system.2023.103149>
- Tok, H., & Yigin, M. (2013). Teachers' views on vocabulary instruction in teaching English as a foreign language. *Turkish Studies*, 8(9), 2151-2164.
- Tsou, W. (2008). Improving speaking skills through instruction in oral classroom participation. *Foreign Language Annals*, 38(1), 46-55.
- Tumen Akyildiz, S. (2025). Exploring pedagogical and ethical dimensions of AI use in EFL classrooms: teachers' views from private high schools. *Journal of Educational Technology and Online Learning*, 8(4), 409-428. <https://doi.org/10.31681/jetol.1744526>
- Webb, S. (2020). *The Routledge handbook of vocabulary studies*. Routledge.
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- Wentzel, K. R. (2014). Socialization in school settings. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of socialization: Theory and research* (2nd ed., pp. 251-275). Guilford Press.
- Wright, A. (1995). *Storytelling with children*. Oxford University Press.
- Yildirim, A., & Simsek, H. (2022). *Qualitative research methods in social sciences* (12th ed.). Seckin Publishing.
- Zhou, X., & Wu, S. (2024). The impact of internal-generated contextual clues on EFL vocabulary learning: Insights from EEG. *Frontiers in Psychology*, 15, 1332098. <https://doi.org/10.3389/fpsyg.2024.1332098>