Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 8, 104451 (2026)
https://doi.org/10.29140/tltl.2026.104451

GenAI-Supported Adaptation of Children's Storybooks into Vocabulary Materials in Indonesian Primary Education

SALMAN ALADE[a] (Corresponding author)
ENDANG NURHAYATI[b]
SRI HARTI WIDYASTUTI[c]
WIWIK PRATIWI[d]

[a]*Department of Language Education, Faculty of Language, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia*
salmanalade.2023@student.uny.ac.id

[b]*Department of Language Education, Faculty of Language, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia*
endang_nurhayati@uny.ac.id

[c]*Department of Language Education, Faculty of Language, Arts, and Culture, Universitas Negeri Yogyakarta, Indonesia*
sriharti@uny.ac.id

[d]*Department of Early Childhood Education, Faculty of Education Universitas Negeri Yogyakarta, Indonesia*
wiwikpratiwi.2025@student.uny.ac.id

Abstract

Generative Artificial Intelligence (GenAI) is increasingly used to support the adaptation of children's storybooks into English as a foreign language vocabulary material in Indonesian primary education, yet limited evidence explains how such use shapes vocabulary, cognitive, multimodal, and cultural aspects of material quality. This qualitative multiple-case study examined three areas: (1) vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance of children's storybooks and teachers' existing GenAI-generated vocabulary materials; (2) the prompt components, design problems, and teacher revision patterns emerging from comparisons of storybooks, standard GenAI outputs, curated GenAI outputs, and teacher revisions; and (3) teacher curation and parent mediation in GenAI-supported storybook-based vocabulary practices. Participants included primary teachers, parents, and supporting school informants, while the corpus comprised 32 children's storybooks and related GenAI artifacts. A comparative analysis matrix was used to examine vocabulary quality,

cognitive demand, multimodal coherence, and cultural relevance. Artifact documentation and walk-through protocols traced prompt-output-revision patterns, while interviews and focus group discussions examined teacher curation, parent mediation, and classroom/home enactment. Data were analyzed using directed qualitative content analysis and reflexive thematic analysis. Within this bounded qualitative dataset, standard GenAI outputs tended to reduce vocabulary specificity, shift tasks toward factual recall, weaken multimodal coherence, and introduce culturally generic references. Curated GenAI outputs showed stronger alignment with target vocabulary, inferencing demands, local contexts, and classroom usability. The findings suggest that GenAI is most useful as a drafting tool within teacher-led curation, culturally responsive prompting, human-in-the-loop review, and AI literacy support.

Keywords: GenAI-generated materials, children's storybooks, teacher curation, parent mediation, vocabulary quality

Introduction

Children's storybooks play an important role in primary education because they provide meaningful contexts for vocabulary learning, language development, and early meaning-making. In storybook-based learning, children encounter words through characters, actions, settings, emotions, and narrative events rather than through isolated word lists (Spencer et al., 2012). Previous studies have shown that storybooks can support language and literacy development when children are guided to connect words with story meaning, illustrations, and adult explanation (Melzi et al., 2023; van den Berg & Klapwijk, 2020). This makes storybooks particularly relevant for primary EFL classrooms, where vocabulary learning needs to be both linguistically accessible and contextually meaningful.

Digital adaptation has extended the pedagogical role of storybooks by allowing teachers to redesign stories with audio-visual scaffolding, repeated access, interactive tasks, and multimodal input. Research on digital storytelling and electronic storybooks suggests that digital features can support vocabulary learning when they reinforce story meaning rather than distract from it (Cordes et al., 2023; Lee, 2020). In Indonesian EFL contexts, digital storytelling has also been associated with vocabulary development, learner engagement, and cultural identity representation (Ferdiansyah, 2024; Kristiawan et al., 2022). However, these benefits depend on how teachers transform storybooks into instructional materials. When adaptation focuses mainly on quick worksheets or assessable word practice, the original story may lose some of its vocabulary, cognitive, multimodal, and cultural affordances.

Vocabulary learning from storybooks is not only a matter of introducing new words. Prior research indicates that effective storybook-based vocabulary instruction involves explicit explanation, repeated exposure, contextual use, and adult-guided discussion (Marulis & Neuman, 2010; Moore et al., 2014; Spencer et al., 2012). In L2 research, vocabulary has commonly been examined in terms of size, sophistication, and variation. In this study, however, vocabulary quality is assessed through features directly observable in the materials, particularly whether target words remain specific, contextually meaningful, developmentally appropriate, and connected to story meaning. This operational definition is appropriate because the study examines whether GenAI-supported adaptation preserves precise and context-rich vocabulary or replaces it with simpler generic alternatives.

Storybooks also provide opportunities for cognitive engagement. Children are encouraged to follow event sequences, infer motives, interpret emotions, and take characters' perspectives when teachers use guided questions and dialogic interaction (de Koning et al., 2020; Price et al., 2012; Xu et al., 2021). This literature suggests that cognitive demand in storybook-based vocabulary tasks should be

examined through the kinds of thinking required by the task, not merely through the presence of comprehension questions. In this study, cognitive challenge refers to the extent to which vocabulary tasks require inference, prediction, explanation, comparison, or perspective-taking rather than simple recall. This focus is important because GenAI-generated tasks may retain the topic of a story while reducing its reasoning demands.

Cultural relevance is also central to storybook-based EFL learning. Storybooks introduce values, social practices, identities, and everyday experiences through names, settings, illustrations, conflicts, and resolutions. In Indonesian primary EFL classrooms, children often make sense of English vocabulary through familiar objects, family routines, school practices, and local social relationships (Hastomo et al., 2025). Prior research on culturally meaningful digital storytelling and cultural bias in large language models (LLMs) suggests that learning materials should be checked for local recognizability, cultural appropriateness, and alignment with learners' lived experiences (Ferdiansyah, 2024; Tao et al., 2024). Accordingly, cultural relevance in this study refers to how far materials reflect children's familiar social practices, local values, names, settings, visual cues, and everyday experiences.

The increasing availability of GenAI has introduced new possibilities for adapting storybooks into EFL vocabulary materials. GenAI can help teachers generate vocabulary lists, glosses, examples, worksheets, image prompts, sequencing tasks, and comprehension questions more efficiently (Jauhiainen & Garagorry Guerra, 2024). In this study, GenAI-generated materials refer to vocabulary tasks, prompts, worksheets, or reading-support activities produced with GenAI tools from storybook content. This definition follows the study's focus on instructional artifacts rather than on GenAI as a general educational technology. However, research also warns that AI-generated materials may be inaccurate, overly simplified, culturally biased, or pedagogically weak without human review (Akgun & Greenhow, 2021; Jain, 2025; Tao et al., 2024). Therefore, teacher and parent curation becomes important. Curation in this study refers to human review, selection, modification, localization, and pedagogical alignment of AI outputs before they are used with children.

Existing literature has examined several related areas, but these areas remain insufficiently connected. Studies on storybook-based vocabulary learning have shown the value of explicit vocabulary instruction, repeated exposure, and adult mediation (Marulis & Neuman, 2010; Melzi et al., 2023; van den Berg & Klapwijk, 2020). Studies on digital storybooks and digital storytelling have shown that multimodal features can support engagement and vocabulary learning when they are pedagogically aligned (Cordes et al., 2023; Ferdiansyah, 2024; Lee, 2020). Research on children's inferencing and dialogic reading has demonstrated that higher-level comprehension depends on the quality of questioning and adult scaffolding (de Koning et al., 2020; Price et al., 2012; Xu et al., 2021). Meanwhile, GenAI studies have highlighted both the potential of AI-supported material generation and the risks of inaccuracy, oversimplification, bias, and weak contextual fit (Jain, 2025; Jauhiainen & Garagorry Guerra, 2024; Nguyen et al., 2024; Tao et al., 2024). Recent work on GenAI-assisted materials development also shows the importance of prompting, evaluation, and human refinement (Fu & Zhang, 2026; Ngo & Fians, 2026). However, less is known about how adaptation from an existing storybook changes its pedagogically relevant features, including vocabulary specificity, reasoning opportunities, text–visual alignment, and culturally recognizable references, or how teachers address such changes during curation.

This study addresses that gap by examining how GenAI-supported adaptation shapes vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance in storybook-based EFL vocabulary materials. These four dimensions provide complementary perspectives on material quality, with each capturing a specific aspect of storybook adaptation: vocabulary quality concerns whether selected target words remain specific, contextually meaningful, developmentally appropriate, and connected

to story meaning, cognitive demand concerns the level of reasoning required by the task, multimodal coherence concerns the alignment between text, visuals, mood, and sequence, and cultural relevance concerns the fit between generated materials and learners' familiar social and cultural contexts. Examining these dimensions within the same artifact corpus allows the study to identify different forms of preservation, weakening, and revision that may occur during GenAI-supported adaptation. It also examines the prompt components, design problems, and teacher revision patterns that emerged from artifact comparison and explains how teachers and parents curate or mediate GenAI-supported vocabulary practices. Rather than treating GenAI as an independent solution, the study positions GenAI as a material-production tool whose pedagogical quality depends on the interaction between storybook affordances, prompt design, teacher curation, and parent mediation. The study was guided by the following research questions:

1. How are vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance preserved, weakened, or revised when children's storybooks are transformed into standard and teacher-curated GenAI-supported EFL vocabulary materials?
2. What prompt components, design problems, and teacher revision patterns emerge from the comparison of source storybooks, standard GenAI outputs, curated GenAI outputs, and teacher revision records?
3. How do teachers and parents mediate, review, and support GenAI-supported storybook-based EFL vocabulary practices in classroom and home contexts?

Literature Review

Storybook-Based Vocabulary Learning

Storybooks offer rich and contextualized input because children encounter vocabulary through narrative events, character actions, emotions, settings, and illustrations. Previous research has shown that vocabulary learning is strengthened when children receive explicit explanation, repeated exposure, and opportunities to connect target words with meaningful contexts (Marulis & Neuman, 2010; Moore et al., 2014; Spencer et al., 2012). Dialogic reading also supports vocabulary development because adults use prompts, questions, and elaboration to guide children's attention to word meanings and story events (van den Berg & Klapwijk, 2020). These findings suggest that storybook-based vocabulary learning depends not only on the number of words introduced but also on the quality of word selection, contextualization, and adult mediation. For the present artifact analysis, vocabulary quality is examined through the specificity, developmental appropriateness, and contextualization of selected target words.

Digital and multimedia storybooks can extend these affordances when audio narration, visuals, animations, and interactive features support the vocabulary focus. Lee (2020) showed that e-book reading and recorded word explanation can support vocabulary learning, particularly when digital features provide meaningful support. Cordes et al. (2023) similarly found that digital storybook interventions can foster child language. However, multimedia features may not automatically improve learning when they distract from story meaning or reduce attention to vocabulary. For the present study, this distinction is important because GenAI-generated vocabulary materials may include glosses, prompts, images, and worksheets, but these features need to remain aligned with the original story.

Cognitive Demand in Storybook Reading

Storybook reading can support cognitive engagement when children are encouraged to infer, explain, sequence events, interpret motives, and consider characters' perspectives. de Koning et al. (2020)

found that inferencing questions embedded in children's books can help children make more inferences. Price et al. (2012) showed that the level of cognitively demanding talk depends on how adults scaffold discussion during read-alouds. Xu et al. (2021) also showed that both adults and conversational agents can support comprehension, although human interaction may produce richer language responses. Therefore, these studies show that cognitive demand is not inherent in the story alone; it is shaped by task design, questioning, and adult mediation.

This point is directly relevant to GenAI-supported material adaptation. A GenAI tool may generate questions quickly, but the questions may focus on factual recall unless the prompt explicitly requires inference, explanation, or perspective-taking. The key issue is whether GenAI-generated storybook tasks preserve the reasoning opportunities embedded in the original story or convert them into lower-demand vocabulary and comprehension exercises. This study examines that issue by comparing original storybook features, standard GenAI outputs, and teacher-curated revisions.

Multimodal Coherence and Cultural Relevance

Storybook meaning is produced through the relationship between words, images, layout, mood, sequence, and visual details. In digital or AI-supported materials, this relationship becomes especially important because teachers may use generated image prompts, visual worksheets, or redesigned story scenes. Multimodal coherence refers to whether textual and visual elements work to support vocabulary meaning and story comprehension. When images do not match the story's emotional tone, setting, or sequence, learners may receive conflicting cues. This concern is consistent with research on digital storybooks, which shows that multimodal features are useful when they reinforce rather than compete with story meaning (Cordes et al., 2023; Lee, 2020).

Cultural relevance is equally important because storybook-based EFL materials carry assumptions about everyday life, social practices, values, and identities. Tao et al. (2024) showed that large language models may reproduce dominant cultural assumptions when users do not provide sufficient contextual guidance. In Indonesian primary EFL settings, culturally meaningful examples may help children connect English vocabulary with familiar names, places, family routines, school practices, and community values (Ferdiansyah, 2024; Hastomo et al., 2025). Thus, cultural relevance in GenAI-supported storybook materials involves more than avoiding inappropriate content. It also requires checking whether examples, images, settings, moral situations, and interaction patterns remain locally recognizable and pedagogically useful.

GenAI-Generated Vocabulary Materials and Teacher Curation

GenAI can accelerate the production of learning materials by generating vocabulary lists, glosses, examples, worksheets, image prompts, and comprehension questions. Jauhiainen and Garagorry Guerra (2024) showed that ChatGPT can support dynamic personalization of school learning materials. Bédi et al. (2025) examined AI-generated pedagogical picture books and highlighted both the promise and limits of AI-generated educational content. These studies indicate that GenAI may be useful for drafting materials, but they do not remove the need to evaluate whether the outputs are accurate, age-appropriate, cognitively meaningful, culturally relevant, and connected to the source text.

The need for human review is emphasized in research on AI use in education. Akgun and Greenhow (2021) warned that AI in K–12 settings raises ethical issues related to privacy, bias, and responsible implementation. Jain (2025) cautioned that AI-generated educational content may be pedagogically weak if used without careful evaluation. Nguyen et al. (2024) also showed that human-AI collaboration depends on iterative evaluation and revision rather than simple acceptance of AI output. In this

study, teacher curation is conceptualized as the professional process of reviewing, revising, selecting, localizing, and pedagogically aligning GenAI outputs before classroom use. In storybook-based EFL vocabulary instruction, this includes checking whether target vocabulary is retained, whether questions require appropriate reasoning, whether examples remain connected to the story, whether images match the narrative, and whether cultural references fit children's lived experiences.

Parent Mediation in Home Vocabulary Practice

Storybook-based vocabulary learning often extends beyond classroom instruction into home reading routines. Parents may support children by translating English words, explaining task instructions, monitoring app use, and encouraging practice. Khan et al. (2024) showed that parent support remains relevant in AI-assisted vocabulary learning for young learners. Xiao et al. (2025) further demonstrated that AI-guided reading needs to be considered alongside parent-led dialogic reading because adult-child interaction remains important for meaningful comprehension. These studies suggest that AI-supported home practice may help parents participate in children's vocabulary learning, but it may also become shallow if children focus only on task completion, points, or screen-based activities.

In the present study, parent mediation is treated separately from teacher curation. Teachers mainly curate instructional materials for classroom use, while parents mediate children's engagement with vocabulary activities at home. This distinction is important because GenAI-supported vocabulary learning may involve different forms of adult mediation across school and home contexts.

Research on storybook-based vocabulary materials has approached material quality from several complementary perspectives. Research on vocabulary instruction informs the examination of vocabulary quality (Marulis & Neuman, 2010; Melzi et al., 2023), while studies of inferencing and dialogic reading provide a basis for examining cognitive demand (de Koning et al., 2020; Price et al., 2012). Digital storybook research supports attention to multimodal coherence (Cordes et al., 2023; Lee, 2020), and work on culturally meaningful storytelling and cultural alignment in AI-generated content informs the examination of cultural relevance (Ferdiansyah, 2024; Tao et al., 2024). These four dimensions provide complementary analytical lenses for examining different aspects of material quality during GenAI-supported storybook adaptation.

Research Gap

The reviewed literature shows three main patterns. First, storybook-based vocabulary research has established the importance of contextualized input, explicit vocabulary instruction, repeated exposure, and adult-guided discussion (Marulis & Neuman, 2010; Melzi et al., 2023; van den Berg & Klapwijk, 2020). Second, research on digital storybooks and dialogic reading shows that multimodal support and cognitive questioning can strengthen vocabulary and comprehension when they are pedagogically aligned (Cordes et al., 2023; de Koning et al., 2020; Lee, 2020; Price et al., 2012; Xu et al., 2021). Third, research on GenAI in education shows that AI can support material production but may also generate inaccurate, generic, oversimplified, or culturally biased outputs without human review (Akgun & Greenhow, 2021; Jain, 2025; Jauhiainen & Garagorry Guerra, 2024; Nguyen et al., 2024; Tao et al., 2024). The unresolved issue is how GenAI-supported adaptation changes the pedagogically relevant features of source storybooks and how teacher curation responds to these changes in actual material-development practices. Evidence is particularly limited on source-to-output changes in vocabulary quality, reasoning demand, multimodal coherence, and cultural relevance in primary EFL storybook materials. This study addresses this gap by tracing these changes across source storybooks, standard GenAI outputs, teacher-curated materials, and revision records, while examining how the resulting materials are mediated in classroom and home use.

Methods

Research Design

This study employed a qualitative multiple-case design in which each case was defined as a school-based storybook-supported EFL vocabulary practice involving the participating teachers within that school. The focal practice included classroom preparation, GenAI-supported material adaptation, and related home vocabulary support. Parent/home literacy practices were treated as nested data sources within each school case to capture how GenAI-mediated vocabulary support extended from classroom planning and instruction into home routines. Supporting informants were included to contextualize each case at the institutional level (VanWynsberghe & Khan, 2007). Consistent with a multiple-case logic, cases were selected for replication and theoretical contrasts, enabling within-case interpretation followed by cross-case synthesis (Ridder, 2017).

This study was designed as a document-centered, interview-supported qualitative multiple-case study. Storybooks and naturally occurring GenAI instructional artifacts constituted the primary analytic evidence, while interviews/focus group discussions (FGDs) and supporting-informant interviews were used to explain enactment, contextualize institutional conditions, and strengthen cross-source interpretation rather than to replace artifact-based analysis (Bowen, 2009). The analysis used directed qualitative content analysis for the documents and artifacts. Vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance were derived deductively from prior literature, while related subpatterns, including vocabulary simplification, cognitive flattening, multimodal mismatch, and cultural misalignment, were identified and refined through iterative comparison of storybooks, GenAI outputs, curated materials, and teacher revisions, consistent with directed qualitative content analysis (Hsieh & Shannon, 2005).

Naturally occurring instructional artifacts refer to storybook-based vocabulary materials, prompts, AI outputs, worksheets, visual prompts, and teacher revisions that were independently created by teachers before or during the study as part of routine lesson preparation. Artifact walkthroughs were used only

Table 1. *Overview of School-Based Cases*

Case	School type	Grade band	Teachers	Parents/ guardians	Storybooks	GenAI artifacts	Main form of GenAI use
Case 1	Public school	Grades 1–3	3	3	8	24	Vocabulary lists, simple worksheets, bilingual glosses
Case 2	Public school	Grades 4–6	4	2	8	28	Inferencing questions, sequencing tasks, worksheet adaptation
Case 3	Private school	Grades 1–3	2	3	8	22	Visual prompts, simplified story tasks, home vocabulary support
Case 4	Private school	Grades 4–6	3	2	8	26	Higher-order thinking skills (HOTS)-oriented prompts, teacher-edited outputs, local-context adaptation
Total			12	10	32	100	

to clarify the origin, purpose, revision process, and classroom or home use of the materials. Alignment or divergence refers to whether GenAI-mediated materials preserved, weakened, or revised the vocabulary, cognitive, multimodal, and cultural features of source storybooks. An overview of these school-based cases is presented in Table 1.

Participants

Participants were recruited from four school-based cases in the Special Region of Yogyakarta, consisting of two public and two private schools. Using purposive, criterion-based sampling to obtain information-rich participants aligned with GenAI-supported storybook vocabulary practices, the study included N = 26 participants: 12 primary teachers, 10 parents/guardians, and 4 supporting informants, consisting of 2 principals and 2 curriculum coordinators. The four schools were selected because they represented variation in school type, grade band, and existing storybook-based GenAI vocabulary practices. Teachers were eligible if they taught English or literacy-related lessons, used children's storybooks for vocabulary activities, and had produced at least one GenAI-assisted instructional artifact during routine preparation. Their GenAI experience varied from frequent use to recent adoption of tools such as ChatGPT, Gemini, Microsoft Copilot, or AI-assisted design and translation tools.

Within each case, teacher-school literacy practice served as the focal source for within- and cross-case comparison, while parent/guardian home-based literacy practices were treated as nested sources to capture how GenAI-supported storybook vocabulary mediation extended beyond the classroom. Parents were recruited through the participating schools and were linked to the corresponding school cases rather than treated as a separate sample. They were included if they supported children's reading or vocabulary practice at home and had experience using AI-supported vocabulary, translation, or learning-support tools. In addition, the document corpus comprised 32 children's storybooks demonstrably used in classroom and home reading routines during the study period, forming the core dataset for document analysis. Principals and curriculum coordinators provided contextual information on literacy routines, digital infrastructure, and technology-use expectations, but their interviews were used for corroboration and contextualization only, not as the primary basis for thematic findings.

Table 2. *Participant Demographic Profile*

Variable	Category	n	%
Role	Teachers	12	46.2
	Parents/guardians	10	38.5
	Supporting informants	4	15.4
Grade band taught (Teachers, n = 12)	Grades 1–3	6	50.0
	Grades 4–6	6	50.0
School type (Teachers, n = 12)	Public	7	58.3
	Private	5	41.7
Supporting informant role (n = 4)	Principal	2	50.0
	Curriculum coordinator	2	50.0
Child's grade band (Parents, n = 10)	Grades 1–3	6	60.0
	Grades 4–6	4	40.0

Note. Percentages for "Teachers/Parents/Supporting informants" are based on N = 26; percentages for subgroup variables are calculated within each subgroup. The descriptive statistics for these continuous variables are presented in Table 3.

Table 3. *Participant Characteristics (Continuous Variables)*

Variable	Group	Mean (SD)	Range
Age (years)	Teachers (n = 12)	35.8 (7.2)	24–50
	Parents (n = 10)	37.9 (6.1)	28–49
	Supporting informants (n = 4)	44.8 (6.0)	36–53
Teaching experience (years)	Teachers (n = 12)	9.2 (5.8)	2–22
GenAI use frequency (times/week)	Teachers (n = 12)	3.2 (1.5)	1–6
	Parents (n = 10)	4.3 (2.1)	1–10

Supporting-informant interviews contextualized case boundaries, school literacy routines, digital infrastructure, and variation across school types and grade bands. A detailed demographic profile of the participants is presented in Table 2.

Instrument and Data Collection

Consistent with the document-centered, interview-supported design, primary evidence for the study came from documents and naturally occurring instructional artifacts, in line with document analysis as a qualitative method. The core dataset comprised: (1) 32 storybooks used in school/home routines and (2) teachers' existing GenAI traces, including copied prompts, AI outputs, vocabulary lists, worksheets, sequencing tasks, inferencing tasks, image prompts, visual materials, and teacher edits where available. These artifacts captured different forms of GenAI-supported material development, including vocabulary outputs, reasoning-oriented tasks, visual prompts, worksheets, and teacher-curated revisions. Teacher-curated versions of these artifacts were retained to document subsequent revisions.

In addition, semi-structured interviews and FGDs with teachers and parents/guardians were conducted to explain enacted routines and how GenAI shaped planning, task design, safeguards, and oversight in practice. Teacher interviews examined how GenAI outputs were generated, checked, revised, localized, and aligned with storybook and vocabulary goals. Example questions included: "How do you usually use GenAI when adapting a storybook into vocabulary materials?", "What aspects of an AI-generated material do you check before using it?", and "Can you describe a GenAI output that you revised and explain what you changed?" Parent interviews examined home use of GenAI-supported activities, including translation, explanation, monitoring, and vocabulary support. Example questions included: "How do you use AI-supported tools when helping your child with vocabulary activities at home?", "What kind of help do you usually provide during these activities?", and "What do you check or monitor when your child uses an AI-supported vocabulary activity?" FGDs explored shared experiences with GenAI-supported materials, including perceived benefits, problems, revision practices, adult review, and safeguards in classroom and home use. Document and artifact analysis addressed textual/material evidence and alignment with source storybooks, whereas interview and FGD data were analyzed thematically to examine teacher curation, parent mediation, and enacted GenAI-supported practices.

A separate artifact was counted when it represented a distinct prompt, AI output, revised output, worksheet/task, image prompt/visual material, or teacher revision note. Not every storybook produced the same number or type of GenAI artifact because teachers adapted storybooks according to classroom needs. Therefore, the unit of artifact analysis was the discrete instructional trace. Copied prompts and written revision notes were counted only when explicitly documented. The eight revision notes reported in Table 4 were standalone written records. Teacher revisions were also identified from observable

Table 4. *GenAI Artifact Corpus by Case*

Case	Storybooks	Prompts	Standard outputs	Curated/ revised outputs	Worksheets/ tasks	Image prompts/visual materials	Teacher revision notes	Total GenAI artifacts
Case 1	8	6	5	4	5	2	2	24
Case 2	8	7	6	6	5	2	2	28
Case 3	8	5	5	4	4	2	2	22
Case 4	8	7	5	6	4	2	2	26
Total	32	25	21	20	18	8	8	100

Note. Teacher revision notes were standalone records. Additional revision evidence was identified in curated outputs and worksheets and clarified through walkthroughs and teacher explanations.

changes in curated outputs and worksheets. When revision decisions were embedded in these artifacts, walkthroughs and teacher explanations were used to clarify the documented changes.

For analytical purposes, the artifacts were classified into original storybooks, standard GenAI materials, and curated GenAI materials. Standard GenAI materials were outputs generated from brief or minimally specified prompts that did not clearly state story context, target vocabulary, learner grade level, cognitive-skill focus, cultural context, or output format. Curated GenAI materials were outputs generated through structured prompts and/or teacher revisions that specified target vocabulary, grade band, inferencing or sequencing focus, local examples, bilingual glosses, or visual-scene alignment. Standard GenAI materials reflected teachers' authentic quick-prompting practices during routine preparation. Curated GenAI materials represented teacher-improved outputs developed through more specific prompting, direct editing, or post-generation revision. Ambiguous cases were checked against prompt logs, output versions, teacher revision notes, and artifact walkthrough explanations. Prompts and outputs appeared in Indonesian, English, or mixed Indonesian-English, and the tools reported by participants included ChatGPT, Gemini, Microsoft Copilot, and AI-assisted design or translation tools. Tool differences were recorded as metadata but were not used as the main basis for comparison.

Additional data were collected through teacher interviews, parent/guardian interviews, FGDs, artifact walkthroughs, supporting-informant interviews, and limited routine observations. Document and artifact analysis focused on what was present in the materials, whereas interviews and FGDs explained teacher curation, parent mediation, classroom/home enactment, safeguards, and oversight practices. Artifact walkthroughs clarified the origin, purpose, prompt intention, revision process, and use of selected materials, while supporting-informant interviews provided school-level context on literacy routines, digital infrastructure, and technology-use expectations. Thus, textual artifacts and human practices were analyzed as related but distinct evidence sources. Table 5 clarifies how each data source informed the research questions, with RQ1 interview excerpts used to contextualize artifact-based patterns.

Data Analysis

All interviews and FGDs were audio-recorded with consent, transcribed verbatim, and linked to the document corpus through metadata on case ID, school type, grade band, tool, language, artifact type, task type, and revision status. The analysis separated textual evidence from human-practice evidence: document and artifact analysis identified what appeared in storybooks and GenAI materials, while interview and FGD analysis explained how teachers and parents produced, revised, selected, and

Table 5. *Data Sources and Analytic Contributions*

Research question	Primary data source	Supporting data source	Role in analysis
RQ1	32 storybooks and GenAI artifacts, including standard and curated outputs	Artifact walkthroughs and selected teacher explanations	Identified material-level alignment or divergence in vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance. Interview data were used only to contextualize artifact-based patterns.
RQ2	Prompts, standard outputs, curated outputs, worksheets/tasks, image prompts, and standalone revision notes	Artifact walkthroughs and teacher interviews	Traced how recurring artifact problems and documented revisions across outputs, worksheets, and revision notes informed prompt components and design specifications.
RQ3	Teacher interviews, parent interviews, FGDs, and limited observations	Selected artifacts used as contextual examples	Explained how teachers and parents mediated, reviewed, and supported GenAI-supported vocabulary practices in classroom and home contexts.
Contextual interpretation	Supporting-informant interviews	School-level documents or contextual notes where available	Clarified school-level literacy routines, technology access, and contextual variation. These data were not used as primary thematic evidence.

used those materials. Original storybooks were examined through document analysis, GenAI traces through qualitative content analysis, and interview/FGD data through reflexive thematic analysis.

Artifact analysis followed a directed qualitative content analysis procedure. Storybooks, prompts, GenAI outputs, revised materials, worksheets, visual prompts, revision notes, walkthroughs, interviews, FGDs, and observation notes were organized by case. The four material-quality dimensions served as deductive categories derived from the literature. Recurring patterns identified during comparison were then refined and formalized as vocabulary simplification, cognitive flattening, multimodal mismatch, and cultural misalignment. Prompt/design specifications were derived by tracing artifact problems, teacher revisions, and resulting review requirements. Interview and FGD data were analyzed thematically to interpret teacher curation and parent mediation.

For RQ1, the 32 storybooks and related GenAI artifacts were compared across vocabulary, cognitive, multimodal, and cultural features. Following iterative comparison, the recurring patterns were operationalized as diagnostic labels for consistent cross-case analysis: vocabulary simplification referred to replacing precise or contextually meaningful words with simpler generic terms; cognitive flattening referred to converting inference, sequencing, perspective-taking, or moral reasoning into recall-based tasks; multimodal mismatch referred to weak alignment between visuals, layout, mood, sequence, and vocabulary focus; and cultural misalignment referred to replacing or weakening locally meaningful names, settings, objects, values, social practices, or visual cues with generic, unfamiliar, or culturally less appropriate references that reduced relevance for Indonesian primary learners.

For RQ2, teachers' prompts, AI outputs, revised outputs, worksheets, image prompts, and revision notes were analyzed to identify recurring prompt components, design problems, and teacher revisions. Teacher revisions were identified from explicit revision notes and observable changes between standard outputs and curated/revised artifacts or worksheets. Walkthroughs and teacher explanations were

Table 6. *Operational Framework for Artifact Analysis*

Material-quality dimension	Literature basis	Operational focus in this study	Diagnostic label used in analysis
Vocabulary quality	Storybook-based vocabulary learning and explicit vocabulary instruction (Marulis & Neuman, 2010; Melzi et al., 2023; Moore et al., 2014; Spencer et al., 2012)	Whether selected target words remained specific, contextually meaningful, developmentally appropriate, and connected to story meaning	Vocabulary simplification: specific or contextually meaningful target words were replaced by generic/ high-frequency alternatives or removed from the task.
Cognitive demand	Inferencing, dialogic reading, adult scaffolding, and perspective-taking in storybook reading (de Koning et al., 2020; Price et al., 2012; Xu et al., 2021)	Whether tasks required inference, sequencing, explanation, motive interpretation, consequence reasoning, or perspective-taking	Cognitive flattening: inferential, sequential, moral, or perspective-taking demands were converted into factual recall or simple plot-order questions.
Multimodal coherence	Digital storybook and multimodal learning research showing that visual and interactive features should reinforce story meaning (Cordes et al., 2023; Lee, 2020)	Whether text, image, mood, layout, sequence, and vocabulary focus were mutually reinforcing	Multimodal mismatch: visuals, image prompts, layout, mood, or sequence did not align with the story event or vocabulary focus.
Cultural relevance	Culturally meaningful digital storytelling and cultural alignment/bias in AI-generated content (Ferdiansyah, 2024; Hastomo et al., 2025; Tao et al., 2024)	Whether names, settings, objects, values, social practices, and visual cues were recognizable and appropriate for Indonesian primary learners	Cultural misalignment: locally meaningful names, settings, objects, values, social practices, or visual cues were replaced, weakened, or omitted in favor of generic, unfamiliar, or culturally less appropriate references.

used to clarify revision decisions when the changes were embedded in the artifacts. The comparison proceeded in two stages. First, within each case, source storybook features were compared with standard GenAI outputs, curated outputs, and documented teacher revisions. Second, the resulting patterns were compared across the four cases to identify recurring similarities and case-level differences. Revision patterns were then established through repeated evidence within and across cases. Prompt/design specifications were derived from patterns that recurred across cases or were supported by repeated artifact and walkthrough evidence. For RQ3, teacher and parent enactment was analyzed through a hybrid deductive-inductive thematic procedure. This procedure applied to the interview and FGD data, separately from the directed content analysis used for RQ1 artifacts.

The first author conducted the primary coding and maintained analytic memos, while three co-researchers reviewed selected excerpts, artifact classifications, and theme summaries. Because the study used reflexive thematic analysis, coding agreement was treated as an interpretive discussion process rather than a statistical reliability target. Disagreements refined code definitions and challenged assumptions. Disagreements in artifact analysis concerned the application of Table 6 indicators, while disagreements in interview/FGD analysis concerned interpretations of teacher curation and parent mediation. Trustworthiness was strengthened through reflexive journaling, an audit trail, within-case memos, cross-case synthesis, selective member checking, and a case-by-source matrix. The analysis focused on interpretive comparisons of materials and adult enactment.

Trustworthiness and Ethical Considerations

Methodological rigor was supported through cross-source comparison, analytic memoing, a reflexive journal, an audit trail, and selective member checking with four teachers. Interpretations were developed by comparing storybooks, GenAI artifacts, interviews/FGDs, artifact walkthroughs, and limited routine observations, with attention to both convergence and divergence across cases. Researchers' reflexivity was addressed through positionality memoing. The research team had professional experience in Indonesian EFL literacy, primary education, and educational technology, as well as linguistic and cultural familiarity with Indonesian school contexts. These backgrounds supported contextual interpretation but could also shape assumptions about GenAI, local relevance, and cultural misalignment. To reduce this risk, the researchers documented assumptions in reflexive memos, discussed interpretations in analytic meetings, and checked selected findings with teachers.

Ethical approval was obtained from an institutional review board at a public university in Indonesia. All participants gave written informed consent, participated voluntarily, and could withdraw at any time. Pseudonyms were used, identifiable data were removed, and files were stored securely. Additional safeguards were applied because the study involved primary education and GenAI artifacts. The researchers did not enter children's names, student work, school-identifying details, photographs, or identifiable classroom data into GenAI systems. Teachers and parents were also instructed not to input identifiable child data. Prompts, outputs, and revision histories were collected only with consent, screened before analysis, and anonymized when they contained names, school references, or other identifiable details.

Results

The findings are aligned with the three research questions and organized by evidence type. RQ1 compares vocabulary, cognitive, multimodal, and cultural features in original storybooks and GenAI-generated EFL vocabulary materials, drawing mainly on storybooks and GenAI artifacts. RQ2 traces recurring artifact problems into teacher revisions and prompt/design specifications. RQ3 examines teacher and parent enactment of GenAI-supported vocabulary practices through interviews, FGDs, and observations.

Findings on RQ1: Vocabulary, Cognitive, Multimodal, and Cultural Features

To answer RQ1, the analysis compared 32 original storybooks and 100 GenAI instructional artifacts across four dimensions: vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance. The four dimensions were used as literature-based material-quality categories, while vocabulary simplification, cognitive flattening, multimodal mismatch, and cultural misalignment described specific alignment or divergence patterns in the artifacts. Table 7 summarizes the main cross-case patterns.

Table 8 shows paired examples from anonymized artifacts to support the analysis. These examples illustrate the types of source storybooks, standard GenAI outputs, visual prompts, worksheets, and teacher-curated revisions included in the artifact corpus. The excerpts were translated and lightly edited for readability. Case and artifact IDs are included for transparency.

The vocabulary comparison showed that standard GenAI outputs often preserved the general topic of the story but reduced the specificity of vocabulary. In all four cases, precise or culturally meaningful words in the source storybooks were sometimes replaced with simpler, high-frequency alternatives. This pattern indicates material-level vocabulary simplification in the standard GenAI outputs. During

Table 7. *Summary of RQ1 Artifact Analysis*

Feature of analysis	Original storybooks	Standard GenAI materials	Curated GenAI materials	Cross-case pattern
Vocabulary quality	Used more specific and context-rich vocabulary such as *rimba, gundah, jejak*, and *berbisik*.	Often replaced precise vocabulary with generic alternatives such as *hutan, sedih*, or *jalan*.	More likely to retain target vocabulary when teachers explicitly listed words to preserve.	Appeared in all four cases.
Cognitive demand	Often required inference, sequencing, motive interpretation, or perspective-taking.	Frequently shifted toward factual recall or plot-order questions.	Included stronger reasoning demands when teachers added prompts for reasons, motives, consequences, and character feelings.	Appeared in all four cases, especially Grades 4–6.
Multimodal coherence	Text and images jointly conveyed setting, emotion, and plot cues.	Outputs were often text-heavy; visual prompts sometimes mismatched story mood or sequence.	Improved when teachers revised image prompts and added scene-specific visual cues.	Appeared across all four cases, but was limited in Case 1 and most evident in Cases 3 and 4.
Cultural relevance	Included local settings, values, and familiar practices such as pasar, gotong royong, and school routines.	Sometimes shifted toward generic or unfamiliar settings such as supermarkets or non-local names.	Improved when teachers specified Indonesian names, places, routines, and social practices.	Appeared in all four cases.

an artifact walkthrough, one teacher explained that after asking ChatGPT to simplify a story, "*the emotion word disappeared*," which meant that students no longer encountered "*a more precise feeling term*" (T01). This comment helps clarify the artifact-based pattern identified in RQ1.

A similar pattern appeared in cognitive demand. Original storybooks often invited learners to infer motives, interpret emotions, connect events, and consider character perspectives. However, standard GenAI tasks frequently converted these demands into factual questions about characters, objects, or event order. The standard GenAI materials showed a shift toward recall-oriented task design. As T02 noted, "*The AI can generate factual questions quickly*," but she usually revised them so that students would think about "*reasons, motives, and consequences*." This explanation clarifies why curated materials included more inferencing and perspective-taking prompts.

Multimodal coherence also varied across the artifacts. In the original storybooks, illustrations often helped children interpret mood, setting, and narrative sequence. Standard GenAI materials were more text-heavy, and visual prompts sometimes failed to match the story atmosphere. One teacher stated that "*sometimes the text says one thing, but the image shows a different atmosphere*" (T03). This pattern was most visible in Cases 3 and 4, where visual prompts and AI-assisted design tools were used more frequently.

Cultural relevance showed a recurring movement from local specificity to more generic examples. The storybooks included familiar Indonesian settings and social practices, while some standard GenAI outputs replaced them with less locally recognizable references. For example, *pasar* became

Table 8. *Paired Examples from Storybooks, Standard GenAI Outputs, and Teacher-Curated Revisions*

Dimension	Original storybook feature	Standard GenAI output	Teacher-curated revision	Interpretation
Vocabulary (Case 1; SB-C1-03/ SO-C1-04/ CR-C1-04)	The story used *rimba, gundah*, and *jejak* to describe a child's anxious walk through a forest.	"The child walks in the forest and feels sad."	"The child follows the *jejak* in the *rimba* and feels *gundah* because he cannot find his way home."	The standard output simplified specific vocabulary, while teacher revision restored target words and emotional nuance.
Cognitive (Case 2; SB-C2-05/ SO-C2-06/ CR-C2-06)	The story required children to infer why a character stayed silent after losing a friend's toy.	"What toy did the child lose?"	"Why did the child stay silent, and what might the friend feel after knowing the truth?"	The standard output shifted the task toward factual recall, while the revision restored inferencing and perspective-taking.
Multimodal (Case 3; SB-C3-02/ VP-C3-02/ CR-C3-03)	The illustration showed a dark path, a worried facial expression, and distant village lights.	Image prompt: "A happy child walking in a bright forest."	Image prompt: "A worried child walking on a dim forest path, looking toward distant village lights."	The standard image prompt mismatched the story mood, while the revision aligned visual cues with the narrative.
Cultural (Case 4; SB-C4-06/ WS-C4-04/ CR-C4-05)	The story referred to children helping at a pasar and joining *gotong royong* after school.	Worksheet example used "supermarket shopping" and "cleaning the garage."	Worksheet revised to "buying vegetables at the pasar" and "cleaning the neighborhood together."	The standard output shifted local references into generic contexts, while the revision localized the task.

Note. SB = original storybook record; SO = standard GenAI output; CR = teacher-curated revision; VP = visual prompt; WS = worksheet/task artifact. The examples are translated and lightly edited excerpts from anonymized artifact records to improve readability while preserving the original meaning and analytic contrast. They are not composite examples.

"supermarket," and *gotong royong* was changed into a more generic cleaning activity. These examples indicate cultural misalignment in some generated materials. Cultural misalignment appeared when locally meaningful references were replaced, weakened, or omitted in specific artifacts.

Findings on RQ2: Prompt Components, Design Problems, and Teacher Revision Patterns

RQ2 examined the prompt components, design problems, and teacher revision patterns that emerged from comparisons of source storybooks, standard GenAI outputs, curated GenAI outputs, and teacher revision records. Within-case comparisons identified recurring problems and corresponding teacher revisions, while cross-case comparisons examined how these patterns converged or varied across the four cases. Vocabulary simplification and cultural misalignment appeared across all cases; recall-oriented questions were more evident in Cases 2 and 4, while multimodal mismatch was strongest in Cases 3 and 4. These recurring problems and revision patterns were subsequently synthesized into the prompt and design specifications summarized in Table 9.

The specifications in Table 9 were derived from repeated artifact-level problems. Vocabulary simplification and cultural misalignment occurred across all four cases, supporting explicit target-vocabulary and cultural-localization specifications. Recall-oriented questions were more prominent

Table 9. *Derived Prompt and Design Specifications for GenAI-Supported Storybook Vocabulary Activities*

Observed issue	Within-case comparison	Cross-case pattern	Teacher response/ revision	Derived specification
Vocabulary simplification	Standard GenAI outputs replaced specific storybook words with generic alternatives; curated outputs restored target words.	Evident in all four cases.	Teachers explicitly retained or restored target vocabulary.	Specify target vocabulary to retain.
Recall-oriented questions	Standard outputs often reduced inference, motive, or perspective-taking tasks to factual questions; curated outputs restored reasoning demands.	Evident across all cases and strongest in Cases 2 and 4.	Teachers added questions about reasons, motives, consequences, and character feelings.	Specify the intended cognitive-skill focus.
Multimodal mismatch	Some generated visual prompts did not match story mood, sequence, or narrative cues; teachers revised scene descriptions.	Limited in Case 1, evident in Case 2, and strongly evident in Cases 3 and 4.	Teachers revised image prompts and added scene-specific visual cues.	Specify visual alignment with story mood, sequence, and scene.
Cultural misalignment	Standard outputs sometimes replaced local names, settings, or practices with generic alternatives; curated outputs restored local references.	Evident in all four cases and more frequently corrected in curated outputs.	Teachers localized names, places, routines, and social practices.	Specify culturally recognizable Indonesian contexts.

in Cases 2 and 4, supporting explicit cognitive-skill prompts, while multimodal mismatch was strongest in Cases 3 and 4, where visual prompts and AI-assisted design were used more frequently. These cross-case patterns, together with teacher revisions, informed the resulting prompt and review specifications.

Findings on RQ3: Teacher and Parent Mediation, Review, and Support of GenAI-Supported Practices

RQ3 examined how teachers and parents mediated, reviewed, and supported GenAI-supported storybook-based EFL vocabulary practices in classroom and home contexts. Teacher and parent practices are presented separately because they represent distinct classroom and home literacy practices. Table 10 summarizes the main patterns.

Teachers enacted GenAI-supported vocabulary practice through prompting, checking, revising, localizing, and aligning materials with storybook goals. They used GenAI to draft vocabulary lists, worksheets, glosses, sequencing tasks, and inference questions. However, they rarely accepted the first output without revision. One teacher explained, "*I do not write everything anymore. I edit the AI draft and fix the culture so it fits my students*" (T01). This pattern highlights teachers' pedagogical judgement in adapting GenAI-supported materials.

Parents enacted GenAI differently. Their practices focused on translating English words, explaining task instructions, monitoring app use, and supporting home vocabulary practice. For some parents, AI-supported translation and explanation helped them remain involved in English vocabulary activities

Table 10. *Teacher and Parent Enactment of GenAI-Supported Practices*

Actor	Main GenAI-supported practices	Evidence from data	Interpretation
Teachers	Prompting, checking, revising, localizing, and aligning materials with storybook goals.	Teachers revised AI outputs by restoring target vocabulary, adding inference questions, correcting cultural references, and adjusting worksheet format.	Teachers worked as pedagogical curators of GenAI-generated materials.
Parents	Translating, explaining, and supporting home vocabulary practice.	Parents used AI tools to translate English words, explain instructions, and help children complete vocabulary tasks.	GenAI functioned as a bridge for parent mediation.
Parents	Monitoring app use and game-like vocabulary practice.	Some children focused on points, quizzes, or task completion but could not explain story meaning.	Some parent accounts indicated task-completion-oriented engagement without accompanying evidence of story explanation.
Teachers and parents	Reviewing AI-supported materials or activities before use.	Adults checked whether outputs were understandable, age-appropriate, culturally familiar, and connected to the story.	Adult mediation involved reviewing GenAI-supported materials for comprehensibility, age appropriateness, cultural familiarity, and connection to the story.

despite limited confidence in English. PR01 reported that her child could stay with the vocabulary game "for a long time" but still could not explain "the story meaning." This account illustrates task-completion-oriented engagement without corresponding evidence of story comprehension.

Across cases, GenAI did not replace adult mediation. Teachers and parents reported checking whether generated tasks remained connected to story meaning, target vocabulary, cultural context, and intended task demands. As PR02 stated, "*AI helps when adults guide it carefully; without that, the activity can become fast but shallow*." This account reinforces the perceived importance of adult guidance in GenAI-supported vocabulary activities.

The labels in Table 11 were assigned interpretively by comparing artifacts, interviews, FGDs, walkthroughs, and limited observations. "Strongly evident" refers to patterns that appeared repeatedly across artifacts and were supported by participant accounts. "Evident" refers to clearly present but less frequent patterns. "Moderate" refers to patterns found in some sources but not dominant. "Limited" refers to patterns appearing only occasionally or in a small portion of the case evidence.

The cross-case matrix shows that vocabulary simplification, recall-oriented questions, and cultural misalignment appeared across all four school cases. Multimodal mismatch appeared across all four cases but varied in prominence. It was limited in Case 1, evident in Case 2, and most prominent in Cases 3 and 4, where visual prompts and AI-assisted design materials were used more frequently. Teacher curation was strongest in Cases 2 and 4, where teachers used GenAI more frequently for inferencing questions, HOTS-oriented prompts, and local-context adaptation. Parent mediation was more visible in Cases 1 and 3, especially in lower-grade routines where parents supported translation, explanation, and home vocabulary practice. These patterns show both replication across cases and case-level variation by grade band, school type, and form of GenAI use, which strengthens the multiple-case logic of the results.

Table 11. *Cross-Case Pattern Matrix*

Pattern	Case 1 Public Grades 1–3	Case 2 Public Grades 4–6	Case 3 Private Grades 1–3	Case 4 Private Grades 4–6	Overall interpretation
Vocabulary simplification in standard GenAI outputs	Evident	Evident	Evident	Evident	Appeared across all cases.
Recall-oriented GenAI questions	Evident	Strongly evident	Evident	Strongly evident	More visible in upper-grade cases where inferencing tasks were expected.
Multimodal mismatch	Limited	Evident	Strongly evident	Strongly evident	More visible in cases using visual prompts or AI-assisted design tools.
Cultural misalignment	Evident	Evident	Evident	Evident	Appeared across all cases, but was more frequently revised in curated outputs.
Teacher curation	Evident	Strongly evident	Evident	Strongly evident	Strongest in cases with more experienced GenAI users.
Parent mediation	Evident	Moderate	Evident	Moderate	More visible in lower-grade cases where home support was more active.

Discussion

This study shows that GenAI can accelerate the adaptation of children's storybooks into EFL vocabulary materials, while the findings also demonstrate the importance of human mediation across classroom and home contexts. Across the four cases, standard GenAI outputs often retained the general topic of the source storybooks, yet they simplified specific vocabulary, shifted inferential tasks toward factual recall, weakened image–text alignment, and replaced local references with more generic contexts. Curated materials showed closer alignment with the material-quality criteria when teachers retained target vocabulary, requested reasoning-oriented questions, revised visual prompts, and localized examples. The findings are interpreted primarily at the level of material characteristics and adult mediation practices. Teacher curation played an important role in strengthening the alignment of generated materials with vocabulary, cognitive, multimodal, and cultural expectations, while parent mediation supported the explanation, monitoring, and use of AI-supported vocabulary activities at home.

The artifact comparison indicates that efficiency in GenAI-assisted material production does not automatically preserve vocabulary quality. Standard outputs sometimes changed context-rich words such as *rimba, gundah, jejak*, and *berbisik* into more generic alternatives, reducing the vocabulary specificity and semantic nuance of the tasks. This finding is consistent with studies showing that storybook-based vocabulary learning depends on repeated exposure to meaningful words in narrative contexts, adult-guided explanation, and opportunities to use words in relation to characters, actions, emotions, and events (Melzi et al., 2023; van den Berg & Klapwijk, 2020). It also supports Lee's (2020) view that digital storybook features are useful when they strengthen meaningful language input. Curated outputs

retained target vocabulary more consistently when teachers specified key words and revised the materials to maintain connections with story meaning. Vocabulary simplification was evident in comparisons of source storybooks, standard GenAI outputs, and teacher-curated revisions.

A similar issue appeared in the cognitive quality of the tasks. Original storybooks created opportunities for inference, sequencing, motive interpretation, and perspective-taking, while standard GenAI questions often focused on characters, objects, or plot order. This pattern reflects a material-level shift toward recall-oriented task design. The finding aligns with Price et al. (2012) and Xu et al. (2021), who show that cognitive demand in storybook reading depends on scaffolding, adult talk, and discussion prompts. GenAI could generate reasoning-oriented questions, but these emerged more reliably when teachers prompted for reasons, consequences, character feelings, and perspective-taking. Therefore, the level of cognitive demand represented in the tasks varied with teacher prompting and revision. Brief or minimally specified prompts more often produced recall-oriented tasks, while curated outputs more frequently included prompts requiring inferencing and perspective-taking.

The multimodal and cultural findings further explain why teacher review remained essential. In several artifacts, visual prompts did not match the story mood, sequence, or emotional cues, while local references such as pasar and gotong royong were sometimes replaced with generic settings. These patterns indicate multimodal mismatch and cultural misalignment at the material level. Tao et al. (2024) argue that LLMs may reproduce dominant cultural assumptions when users do not provide sufficient contextual guidance. The present study extends this concern to Indonesian primary EFL vocabulary materials by showing that teachers had to restore local names, familiar places, school routines, and social practices during curation. Cultural relevance therefore emerged as an important material-quality consideration, particularly when vocabulary activities retained familiar social and cultural references. This conclusion reflects the case-based interpretation of the analyzed artifacts. Across the four dimensions, the findings showed different forms of change in vocabulary specificity, reasoning demand, narrative–visual alignment, and cultural references, with vocabulary simplification and cultural misalignment appearing across all four cases and multimodal mismatch being more prominent in cases using visual prompts and AI-assisted design.

The prompt/design specifications derived from the findings represent practice-grounded design implications. They emerged from recurring artifact problems: target vocabulary needed to be specified because precise words were sometimes replaced, cognitive-skill focus was needed because questions became recall-oriented, cultural localization was needed because examples became generic, and multimodal support was needed because visual prompts sometimes failed to match the story. This finding aligns with Walter's (2024) argument that AI literacy and prompt engineering require critical thinking, and with Nguyen et al. (2024), who emphasize that AI-supported production depends on human evaluation and revision. The results show that these specifications were derived by tracing observed artifact problems, teacher revisions, and related prompt or post-generation review requirements. Prompt design served to address material-quality concerns involving vocabulary retention, reasoning demand, visual alignment, and cultural fit.

Teacher and parent practices also confirm the importance of human mediation. Teachers used GenAI mainly for drafting, checking, revising, localizing, and aligning materials with storybook goals. Their role shifted toward pedagogical curation, where professional judgment was needed to decide whether outputs retained vocabulary, supported reasoning, matched visual meaning, and fit local contexts. Parents used AI tools differently, mainly for translation, explanation, monitoring, and home vocabulary support. Some parents with limited English confidence reported using GenAI to remain involved in children's EFL activities, supporting the finding of Khan et al. (2024) that parent support remains important in AI-assisted vocabulary learning. Some parent accounts described

task-completion-oriented engagement, in which children focused on points, quizzes, or finishing activities but did not explain the associated story meaning. These accounts indicate that sustained engagement with an AI-supported activity does not necessarily correspond to engagement with story meaning. This concern is consistent with Xu et al. (2021) and Xiao et al. (2025), who show that AI-mediated reading still requires adult mediation for meaningful interaction and richer language production.

The findings have implications for teacher development, parent guidance, and future research in comparable primary EFL contexts. Teacher training should help teachers evaluate GenAI outputs for vocabulary retention, reasoning demand, image–text alignment, cultural relevance, age appropriateness, and classroom usability. Parent guidance may also help support shared reading, explanation, and vocabulary use alongside AI-supported home activities. For similar Indonesian primary EFL contexts, GenAI can serve as a drafting and scaffolding resource, while teachers and parents support material quality through review, revision, and contextual guidance. Future research should examine curated GenAI materials through longitudinal, mixed-methods, or design-based studies across more diverse school contexts.

Conclusion

This qualitative multiple-case study suggests that GenAI can support the development of storybook-based vocabulary materials in Indonesian primary education, with the quality of generated materials shaped by sustained human mediation and review. Across four school-based cases in Yogyakarta, involving 12 teachers, 10 parents/guardians, 4 supporting informants, and a document corpus of 32 children's storybooks, the analysis indicated that uncurated or minimally prompted GenAI outputs tended to require teacher curation to preserve vocabulary specificity, cognitive demand, multimodal coherence, and cultural fit, while parent mediation supported explanation and monitoring in home use. In these cases, standard GenAI outputs sometimes simplified precise vocabulary, shifted tasks toward factual recall, weakened visual-text alignment, or introduced culturally generic references when they were not reviewed or revised. By contrast, stronger materials emerged when teachers refined target vocabulary, inferential questions, visual alignment, cultural references, and instructional fit, while parents supported home use through explanation and monitoring. Conceptually, the study positions GenAI as a mediational tool shaped by storybook affordances, prompt design, teacher curation, and parent mediation. Vocabulary quality, cognitive demand, multimodal coherence, and cultural relevance were used as complementary dimensions for evaluating material quality across source storybooks, standard GenAI outputs, and teacher-curated materials. The study contributes two main implications. First, GenAI was most pedagogically useful as a drafting and scaffolding resource within teacher-led, human-in-the-loop literacy practices. Second, it provides practice-grounded prompt and design specifications that may help teachers preserve vocabulary quality, cognitive challenge, multimodal alignment, and cultural relevance when adapting storybooks into vocabulary materials.

The study is limited by its qualitative multiple-case design and the purposive selection of four school contexts. Its findings therefore support analytic transferability to comparable literacy settings rather than statistical generalization. Participants' prior experience with GenAI may also have influenced how they created, evaluated, revised, and mediated AI-generated materials. In addition, the study did not directly assess children's long-term vocabulary growth, inferencing development, or reading outcomes, nor did it compare curated and uncurated materials through intervention-based testing. Future research could examine relationships among these material-quality dimensions using larger datasets and longitudinal or mixed-methods designs across more diverse school contexts. For similar primary EFL contexts, the findings point to the need for professional development in pedagogical curation, prompt design, cultural appropriateness checking, and responsible GenAI use, while parent-facing guidance may help encourage dialogic meaning-making alongside app-based vocabulary activities at home.

Acknowledgements

We would like to express our sincere gratitude and appreciation to the Indonesia Endowment Fund for Education (LPDP) Scholarship of the Ministry of Finance of the Republic of Indonesia, which has facilitated and provided funding support for this research.

References

Akgun, S., & Greenhow, C. (2021). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431–440. https://doi.org/10.1007/S43681-021-00096-7

Bédi, B., Beedar, H., Chiera, B., Chua, C., Geneix-Rabault, S., Kreusch, V., Maizonniaux, C., Rayner, M., Rendina, S., Ryan-Cooper, E., Sukhyi, V., Vargova, I., Wright, S., Yao, C., & Zviel-Girshin, R. (2025). AI-generated tailor-made pedagogical picture books: How close are we? *Education Sciences*, *15*(12), 1704. https://doi.org/10.3390/educsci15121704

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Cordes, A.-K., Egert, F., & Hartig, F. (2023). Fostering child language with short-term digital storybook interventions. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *55*(4), 155–168. https://doi.org/10.1026/0049-8637/a000264

de Koning, B. B., Wassenburg, S. I., Ganushchak, L. Y., Krijnen, E., & van Steensel, R. (2020). Inferencing questions embedded in a children's book help children make more inferences. *First Language*, *40*(2), 172–191. https://doi.org/10.1177/0142723719894770

Ferdiansyah, S. (2024). Collaborative genre-based digital storytelling of English as a foreign language: a case of an Indonesian primary school. *Education 3–13*, *52*(8), 1594–1601. https://doi.org/10.1080/03004279.2023.2205437

Fu, X., & Zhang, L. J. (2026). Integrating GenAI into EAP course materials design: A case study of the "English academic writing in the digital environment" course. *Journal of English for Academic Purposes*, *83*, 101752. https://doi.org/10.1016/j.jeap.2026.101752

Hastomo, T., Widiati, U., Ivone, F. M., Zen, E. L., Hasbi, M., & Khulel, B. (2025). AI-powered conversational agents and intercultural learning: Insights from Indonesian EFL students. *Intercultural Communication Education*, *8*(1), 103217. https://doi.org/10.29140/ice.v8n1.103127

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Jain, A. (2025). Artificial intelligence and the future of educational content: A cautionary tale. *Journal of Educators Online*, *22*(4). https://doi.org/10.9743/JEO.2025.22.4.7

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI and education: Dynamic personalization of pupils' school learning material with ChatGPT. *Frontiers in Education, 9*, Article 1288723. https://doi.org/10.3389/feduc.2024.1288723

Khan, M. A., Kurbonova, O., Abdullaev, D., Radie, A. H., & Basim, N. (2024). Is AI-assisted assessment liable to evaluate young learners? Parents support, teacher support, immunity, and resilience are in focus in testing vocabulary learning. *Language Testing in Asia*, *14*(1), 48. https://doi.org/10.1186/s40468-024-00324-x

Kristiawan, D., Carter, C., & Picard, M. (2022). Impact of CALL professional development for EFL materials on teacher agency and technological pedagogical content knowledge (TPACK) in Indonesian Islamic schools. *Teaching English with Technology*, *22*(3–4), 20–42. https://eric.ed.gov/?id = EJ1367619

Lee, S. H. (2020). Learning vocabulary from e-book reading and recorded word explanation for low-income elementary students with and without reading difficulties. *Reading and Writing*, *33*(3), 691–717. https://doi.org/10.1007/s11145-019-09983-2

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning. *Review of Educational Research*, *80*(3), 300–335. https://doi.org/10.3102/0034654310377087

Melzi, G., Schick, A. R., & Wuest, C. (2023). Stories beyond books: Teacher storytelling supports children's literacy skills. *Early Education and Development*, *34*(2), 485–505. https://doi.org/10.1080/10409289.2021.2024749

Moore, W., Hammond, L., & Fetherston, T. (2014). Strengthening vocabulary for literacy: an analysis of the use of explicit instruction techniques to improve word learning from story book read-alouds. *Australian Journal of Learning Difficulties*, *19*(2), 153–172. https://doi.org/10.1080/19404158.2014.964992

Ngo, T. N., & Fians, G. (2026). Between prompting strategies and ethical concerns: How EAP teachers develop materials with Generative AI. *Journal of English for Academic Purposes*, *81*, 101680. https://doi.org/10.1016/j.jeap.2026.101680

Nguyen, A., Kremantzis, M., Essien, A., Petrounias, I., & Hosseini, S. (2024). Editorial: Enhancing student engagement through artificial intelligence (AI): Understanding the basics, opportunities, and challenges. *Journal of University Teaching and Learning Practice, 21*(06). https://doi.org/10.53761/caraaq92

Price, L. H., Bradley, B. A., & Smith, J. M. (2012). A comparison of preschool teachers' talk during storybook and information book read-alouds. *Early Childhood Research Quarterly*, *27*(3), 426–440. https://doi.org/10.1016/j.ecresq.2012.02.003

Ridder, H.-G. (2017). The theory contribution of case study research designs. *Business Research*, *10*(2), 281–305. https://doi.org/10.1007/s40685-017-0045-z

Spencer, E. J., Goldstein, H., & Kaminski, R. (2012). Teaching vocabulary in storybooks. *Young Exceptional Children*, *15*(1), 18–32. https://doi.org/10.1177/1096250611435367

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), Article pgae346. 10.1093/pnasnexus/pgae346.

van den Berg, L., & Klapwijk, N. (2020). The impact of second-language storybook reading on the vocabulary acquisition of grade 1 learners. *Language Matters*, *51*(1), 63–85. https://doi.org/10.1080/10228195.2019.1657488

VanWynsberghe, R., & Khan, S. (2007). Redefining case study. *International Journal of Qualitative Methods*, *6*(2), 80–94. https://doi.org/10.1177/160940690700600208

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*(1), 15. https://doi.org/10.1186/s41239-024-00448-3

Xiao, F., Zou, E. W., Lin, J., Li, Z., & Yang, D. (2025). Parent-led vs. AI-guided dialogic reading: Evidence from a randomized controlled trial in children's e-book context. *British Journal of Educational Technology*, *56*(5), 1784–1813. https://doi.org/10.1111/bjet.13615

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education*, *161*, 104059. https://doi.org/10.1016/j.compedu.2020.104059