A Readability-Driven Prompting Framework for Accurate Grade-Specific EFL Narrative Creation

Ronald William Marbun, Makoto Shishido
Media Design Communication Engineering, Tokyo Denki University, Tokyo, Japan

***Abstract*—The integration of Artificial Intelligence (AI) into English as a Foreign Language (EFL) education offers new opportunities for developing adaptive and engaging learning materials. Narrative-based content is central to improving reading comprehension, vocabulary acquisition, and learner motivation. However, maintaining grade-appropriate readability in AI-generated narratives remains a major challenge. This study presents Readability-Driven Prompting (RDP), a novel technique designed to enhance the accuracy and efficiency of large language models in generating grade-level narratives. Using GPT-4o-mini, three prompting strategies—CEFR Keyword-Constrained Prompting (CKCP), Instruction-Based Prompting (IBP), and the proposed RDP—were applied to produce narratives for 7th-grade (A1–A2 CEFR) and 10th-grade (B1–B2 CEFR) learners. The outputs were evaluated using Flesch Reading Ease (FRE), Dale–Chall (DC) readability metrics, lexical analysis, and human assessments. Experimental results indicate that the RDP approach achieves higher alignment with target readability levels and improved lexical appropriateness compared to baseline methods, demonstrating a scalable and effective strategy for generating educational narratives, particularly for beginner-level learners.**

***Keywords*—Artificial Intelligence (AI); English as a Foreign Language (EFL); large language models (LLMs); readability metrics; narrative generation; prompt engineering; educational technology**

I. Introduction

The increasing use of large language models (LLMs) in English as a Foreign Language (EFL) education offers new opportunities to create engaging and adaptive materials. Narrative texts, in particular, are valuable for developing reading comprehension, cultural awareness, and student motivation [1] [2]. However, a significant challenge remains: controlling the grade-level of AI-generated narratives. If these texts do not match the intended readability level, their instructional effectiveness may be compromised.

Providing EFL learners with materials that are appropriate for their grade level is crucial for effective language development [3]. In contrast, materials that are too advanced can be ineffective and may even hinder learners' progress [4]. Crafting narratives precisely targeted to specific grade levels is particularly challenging, as it typically requires the expertise of skilled educators and content specialists [5]. Recent studies have shown that LLMs often struggle to assess or control grammar and vocabulary levels consistently without careful guidance [6], highlighting the need for improved prompting strategies.

Using ChatGPT to support vocabulary learning has been shown to be effective [7], but sustained gains require repeated exposure to level-appropriate texts. Incorporating a high number of low-frequency words may discourage students from engaging with reading materials, underlining the importance of word frequency in instructional texts [8].

Lexical difficulty is critical for EFL readers; eye-tracking research has found that readers spend more time fixating on long, low-frequency words, and that L2 readers tend to skip fewer words than L1 readers [9], [10]. In addition, research has shown that achieving 95% overall understanding of a text is highly effective for language learning [11]. The CEFR is an internationally recognized standard for describing language ability, ranging from beginner (A1) to proficient (C2) levels, and widely used in language curriculum design. This highlights the importance of aligning texts with the target CEFR level, and suggests that the effectiveness of instructional materials can be evaluated through lexical analysis.

Another approach to evaluating text appropriateness is through readability metrics. Flesch and Kincaid introduced the Flesch Reading Ease (FRE) measure, which emphasizes syllable complexity [12]. In contrast, Dale and Chall proposed a formula (DC) that assesses textual complexity based on the frequency of familiar words [13]. They continue to be widely used and adapted in modern educational research and computational linguistics [14], [15]. By combining these metrics, educators can develop learning materials that balance linguistic complexity with student engagement and motivation. However, readability metrics alone are insufficient, particularly when addressing the needs of L2 learners. Therefore, this study evaluates narratives using both lexical analysis and readability metrics.

Effective prompt strategies are important for improving and optimizing ChatGPT's performance in EFL applications [6], [16], [17]. Recently, prompt engineering has become a popular research topic. The fundamental step in prompt engineering is to provide clear instructions and context to the model [6], as a result, Instruction-Based Prompting has become widely used. However, Instruction-Based Prompting alone may not be sufficient for generating narratives or supporting L2 learning [6], [16].

Building on previous work, research has shown that applying CEFR word constraints can enhance the effectiveness of targeted narrative generation [18]. This approach is more effective for producing narratives suitable for specific grade levels. Since GPT models are sensitive to input and perform better with sufficient context, embedding readability theory and CEFR-based constraints into prompts provides clearer guidance with fewer tokens. Furthermore, supplementing prompts with sample stories at the target grade level can yield comparable

results with reduced computational cost. Accordingly, this study introduces Readability-Driven Prompting (RDP), a novel few-shot prompting framework that integrates readability theory, CEFR guidelines, and representative examples to generate grade-appropriate EFL narratives.

We hypothesize that Readability-Driven Prompting (RDP) will be more effective than Instruction-Based Prompting (IBP), and non-inferior to CEFR Keyword Constrained Prompting (CKCP), in generating grade-appropriate EFL narratives for Grade 7 (A1–A2) and Grade 10 (B1–B2). Effectiveness is assessed as follows: a) Readability alignment—RDP is expected to produce a higher proportion of texts falling within pre-registered grade bands on Flesch Reading Ease and Dale–Chall metrics, with smaller deviations from band centers compared to IBP, and no worse than CKCP within a pre-specified non-inferiority margin; b) Controllability across grades—RDP should yield a larger and more consistent separation between Grade 7 and Grade 10 readability levels than IBP, and at least as distinct as CKCP; and c) Lexical targeting—RDP is anticipated to more accurately match CEFR lexical profiles (e.g., more A1–A2 and fewer ≥B1 tokens at Grade 7, more B1–B2 and controlled ≥C1 usage at Grade 10) compared to IBP, and to perform non-inferior to CKCP, with frequency profiles that minimize long, low-frequency words at Grade 7 and introduce them in a controlled manner at Grade 10.

This study addresses the following research questions:

- Can Readability-Driven Prompting (RDP) improve the alignment of large language model-generated EFL narratives with target grade-level readability metrics?
- How do students perceive the readability, engagement, and vocabulary learning potential of narratives generated using RDP?

The primary contributions of this work are as follows:

- We propose Readability-Driven Prompting (RDP), a novel prompting framework that integrates explicit readability theory, CEFR-based constraints, and in-context grade-level narrative exemplars to guide large language models in producing educationally aligned narratives.
- We provide, to our knowledge, the first systematic experimental comparison of RDP with CEFR Keyword Constrained Prompting (CKCP) and Instruction-Based Prompting (IBP), focusing on both reading metrics and lexical targeting for EFL learners at distinct grade bands.
- We perform a comprehensive analysis—combining established readability formulas, lexical CEFR alignment, and human learner evaluation—demonstrating the efficacy of RDP for generating grade-specific texts in EFL education.

The remainder of this study is organized as follows: Section II reviews related work on prompt engineering and readability control. Section III describes our experimental methodology and evaluation framework. Section IV presents the results of lexical and readability analyses and a human evaluation, followed by discussion. Section V presents the discussion. Finally, Section VI concludes and suggests avenues for future research.

II. Related Work

A. Prompt Engineering on Education

Few-shot and zero-shot prompting strategies have emerged as key techniques for improving reasoning performance, mitigating hallucinations, and regulating the linguistic register in large language model (LLM) outputs [19]. In educational contexts, these methods help ensure that generated content aligns with learners' proficiency levels and pedagogical goals. An important extension, known as Generated Knowledge, enhances model reasoning by providing theoretically relevant information as supplemental input [20]. Incorporating educational or linguistic frameworks—such as readability measures like Flesch–Kincaid or Dale–Chall—within prompts enables the model to produce text that is calibrated for specific grade levels. This approach not only constrains linguistic complexity but also increases the predictability, interpretability, and educational value of the generated instructional materials.

B. Readability Metrics

Readability refers to how easily a text can be understood by its intended audience. Classical formulas, such as the Flesch–Kincaid and Dale–Chall indices, quantify textual difficulty based on variables like sentence length, word familiarity, and syntactic transparency. More recently, LLM-driven or prompt-based methods have been proposed to move beyond these surface-level indicators by capturing latent linguistic features that influence perceived difficulty [14]. Despite these advancements, empirical evidence suggests that model-based approaches remain inconsistent in generating text at predefined complexity levels, largely due to the opaque and stochastic internal workings of LLMs. In contrast, traditional readability measures continue to provide transparent, replicable, and pedagogically meaningful metrics for the development and assessment of educational texts.

From a theoretical standpoint, readability frameworks provide quantifiable, evidence-based parameters for managing text complexity [12] [13]. When such parameters—sentence length, lexical frequency, and semantic transparency—are incorporated directly into prompt formulation, they introduce explicit and interpretable constraints that counterbalance the inherent variability of LLM generation processes [14]. For example, prompts specifying bounded average sentence length or prioritizing high-frequency vocabulary enable more reliable alignment between model outputs and intended learner proficiency levels. The integration of readability theory into prompt engineering thus transforms text generation from an iterative, heuristic endeavor into a structured, evidence-driven methodology for producing educationally appropriate materials.

C. Narrative Generation

Recent advancements in large language models (LLMs), particularly GPT-3 and GPT-4, have demonstrated significant capabilities in narrative generation. With well-designed prompts, these models can produce fluent and coherent narratives that surpass previous automated storytelling systems in both linguistic quality and narrative cohesiveness [15]. In educational

contexts, LLM-based narrative generation has been used to enhance learner engagement, foster narrative awareness, and promote higher-order comprehension and creativity [21]. However, notable limitations persist, especially in maintaining long-range coherence, character consistency, and thematic integration throughout extended texts [14], [15]. As a result, expert curation, human supervision, and post-editing remain necessary to ensure that generated narratives are pedagogically appropriate and aligned with instructional objectives.

D. Research Gap

Despite advances in prompt engineering, readability control, and narrative generation, existing research has yet to develop a robust framework for consistently producing educational narratives aligned with predetermined grade-level and cognitive-developmental criteria [6]. While prompt engineering can influence the stylistic and structural aspects of generated text, the inherent opacity and variability of LLM outputs make it difficult to achieve deterministic control over readability and conceptual depth.

Integrating readability theory within a Generated Knowledge framework offers a promising solution to these challenges. By embedding explicit textual constraints—such as quantified expectations for sentence length, lexical familiarity, and semantic transparency—into the model's reasoning process, it becomes possible to guide generation toward predictable linguistic boundaries. This hybrid approach combines the adaptability of contemporary language models with the stability and interpretability of traditional readability measures.

Such integration not only enhances technical reliability but also reframes educational text generation as an evidence-based practice rooted in literacy and educational measurement theory. It enables the creation of pedagogically aligned narratives that ensure coherence, accessibility, and adherence to instructional goals across diverse learner levels. Consequently, this study addresses a critical gap by investigating how readability-based generated knowledge can systematically regulate LLM narrative production to achieve consistency, transparency, and educational appropriateness—outcomes not yet realized in the current state of research.

III. Methodology

A. Study Design and Overview

This study aims to validate the effectiveness of the proposed Readability-Driven Prompting (RDP) framework for English as a Foreign Language (EFL) narrative generation. RDP is an enhanced prompting method that conditions large language models on readability theory and grade-level exemplars, enabling the production of narratives more closely matched to learner proficiency.

To establish comparative validity, RDP was evaluated against two reference prompting strategies: Instruction-Based Prompting (IBP), which represents minimally guided text generation, and CEFR Keyword Constrained Prompting (CKCP), which applies strict lexical constraints.

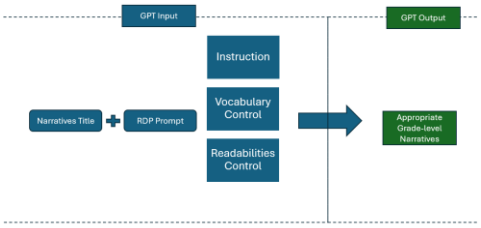

Fig. 1. Block diagram of the proposed readability-driven prompting (RDP) framework.

The proposed RDP methodology (see Fig. 1) integrates explicit readability instructions (using Flesch–Kincaid Grade Level and Dale–Chall metrics), CEFR-aligned vocabulary references, and example narratives directly within the prompt. By embedding theoretical constraints and representative exemplars, RDP aims to achieve high readability alignment and reduced computational cost compared to traditional word-list-based approaches.

B. Targeted Grade Levels

In line with common curricular standards that map lower-secondary EFL learners to CEFR levels A1–A2 and upper-secondary learners to B1–B2 [22], [23], we operationalize early EFL as Grade 7 and intermediate EFL as Grade 10. In CEFR terms, Grade 10 typically aligns with level B1. Accordingly, we compare RDP against CEFR Keyword Constrained Prompting (CKCP) and Instruction-Based Prompting (IBP) using GPT-4o-mini with default temperatures (t = 1) at two anchor levels commonly targeted in secondary EFL curricula: 7th grade (A1–A2 CEFR) and 10th grade (B1–B2 CEFR).

Thus, 7th and 10th grade serve as representative anchor points for lower- and upper-secondary EFL reading. Seventh grade typically corresponds to early–low intermediate proficiency (A2 CEFR), where learners possess sufficient decoding skills and core vocabulary for short narratives but remain sensitive to sentence length and unfamiliar lexis. In contrast, tenth grade aligns with upper-intermediate proficiency (B1 CEFR), where learners can manage more complex syntax and a broader academic vocabulary.

Fig. 2 illustrates the research objectives. The study aims to generate narratives for both beginner and intermediate EFL learners. Narratives will be classified as appropriate for beginners if they are suitable for 7th graders (CEFR A1–A2 level) and as appropriate for intermediates, if they are suitable for 10th graders (CEFR B1–B2 level).

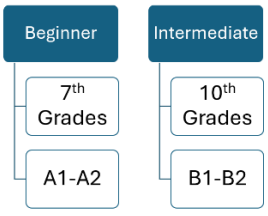

Fig. 2. Research target representative.

C. Dataset Generation

For this research, we produced 100 narratives for each combination of prompting method and grade level. Specifically, for each of the three methods (IBP, CKCP, RDP) at both 7th and 10th grade targets, we generated 100 narratives using GPT-4o-mini under identical decoding settings. Every narrative title and theme is unique covering common EFL themes (e.g., school life, family, travel, science) with a range between 200 and 300 words. To isolate prompting effects while holding topic constant, each title was instantiated once for every prompting method (IBP, CKCP, RDP) at each target grade (7th, 10th). In total, this yielded 600 narratives (3 methods × 2 grades × 100 narratives).

D. Evaluation

This research evaluates the generated narratives using three methods: readability analysis, lexical analysis, and survey-based assessment. For readability evaluation, two complementary metrics were employed. Table I presents the Flesch–Kincaid Reading Ease (FRE) scores and corresponding grade interpretations. FRE scores range from 0 to 100, with higher values indicating easier readability. Unlike the Flesch–Kincaid Grade Level (FKGL), which directly maps text complexity to a U.S. grade level, FRE provides a broader indication of textual accessibility across proficiency levels. In this study, narratives targeting the 7th-grade level were required to achieve FRE scores between 70 and 80, while narratives targeting the 10th-grade level were expected to fall within the range of 50 to 60.

TABLE. I. FLESCH-KINCAID EASE MEASUREMENT

Flesch-Kincaid Reading Ease (FRE)	
Point	Grade
90-100	Very easy (5th grade; suitable for children)
80-89	Easy (6th grade)
70-79	Fairly easy (7th grade)
60-69	Standard (8th-9th grade; plain English)
50-59	Fairly difficult (10th-12th grade)
30-49	Difficult (college level)
0-29	Very difficult (college graduate level)

TABLE. II. DALE-CHALL MEASUREMENT

Dale-Chall (DC)	
Point	Grade
<= 4.9	Easily understood by 4th grade or lower
5.0-5.9	Easy (6th grade)
6.0-6.9	Fairly easy (7th grade)
7.0-7.9	Standard (8th-9th grade; plain English)
8.0-8.9	Fairly difficult (10th-12th grade)
9.0-9.9	Difficult (college level)
10+	Very difficult (college graduate level)

Table II presents the Dale–Chall formula (DC) scores and corresponding grade interpretations. The DC formula measures text complexity based on the proportion of unfamiliar words relative to a predefined list of commonly used words. In this study, narratives targeting the 7th-grade level were required to achieve DC scores between 6.0 and 6.9, while narratives targeting the 10th-grade level were expected to fall within the range of 7.0 to 7.9.

This research also evaluates the generated narratives using lexical analysis. The narratives undergo several text processing steps, as shown in Fig. 3. The main processes are as follows:

- Normalization: To ensure consistency in format, all raw text is converted to lowercase, punctuation is removed, Unicode is standardized, and extraneous characters such as hyphens are eliminated.
- Tokenization: To conduct lexical analysis, each word must be identified and categorized. This process separates individual words and assigns parts of speech (e.g., noun, adjective, verb). SpaCy is used for tokenization in this study.
- Filtering: Certain tokens, such as character names, URLs, numbers, acronyms, and hashtags, do not contribute to text difficulty and should be excluded. SpaCy, in addition to our word list, is also used to filter out these elements to ensure proper lexical analysis.
- Lemmatizing: English words can appear in many forms. To accurately assess difficulty, each word is reduced to its dictionary form. The NLTK library is employed for the lemmatization process.

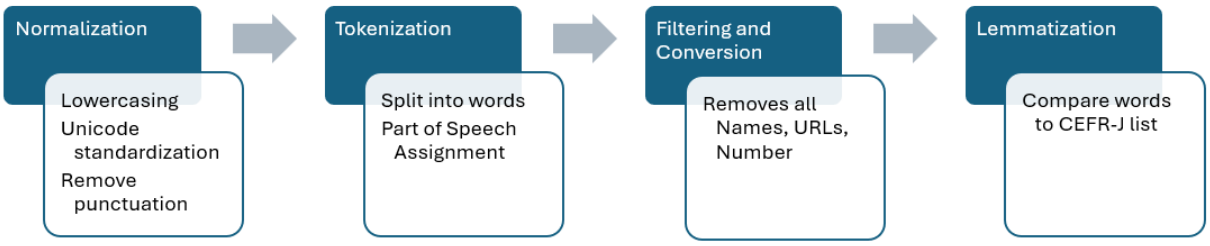

Fig. 3. Lexical analysis text processing pipeline.

After text processing is complete, all lemmatized tokens are compared against the CEFR-J dataset. Each token is counted and its difficulty level is determined based on its CEFR classification. Tokens not present in the CEFR-J dataset are considered as above C1 level, since most non-relevant items have already been filtered out. We will refer this as Unknown Words.

For learning to be effective, students need to understand at least 95% of the content they read [11]. Accordingly, for the beginner level (A1–A2), only words within the A1–A2 bands are counted, while for the intermediate level (B1–B2), words within the A1–B1 bands are considered. The effectiveness of each method is determined by how closely the proportion of appropriate-level vocabulary approaches the 95% threshold.

The final evaluation method involved a human evaluation survey. Participants were beginner-level Japanese learners of English with an interest in improving their language skills. The evaluation was conducted through a storytelling game, developed for smartphones using the Unity engine. Thirty students from Tokyo Denki University participated in the study. Participants were selected based on their English proficiency and voluntarily agreed to try the game and assess the quality of the narratives for learning purposes.

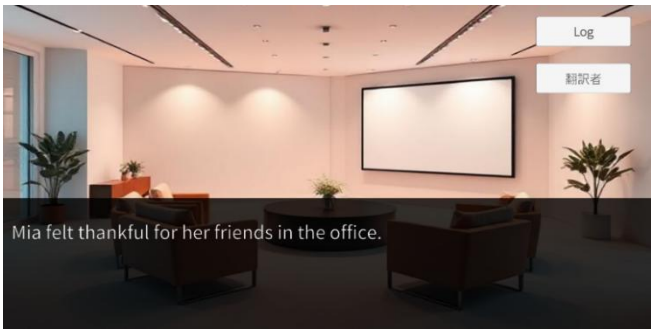

Fig. 4. Example of game narratives.

The process begins with student self-introduction and course selection, which function as initial engagement steps. Learners then proceed through narratives generated using the RDP method. Fig. 4 illustrates how these narratives are implemented and represented within the game environment. During the narrative progression, students are prompted with comprehension questions that must be answered to continue. If a student is unsure about the meaning of a word, a translation button is available, providing an explanation of the word's meaning and its usage within the narrative context.

The game served dual purposes: acting both as a storyteller and as a learning tool. At the beginning of the game, participants selected their preferred topics. Based on these preferences, a new narrative was generated using the RDP prompting method. The generated narrative was then presented in the game, and participants were asked to read the text before answering questions designed to assess comprehension and engagement. After completing these questions, participants received immediate feedback on their performance and a review of their responses. Finally, they were prompted to complete a survey evaluating the narrative's quality, their level of engagement, and the perceived learning value of the experience.

The survey comprised six items designed to assess participants' perceptions of the generated narratives:

1) The generated narratives were easy to understand: This item measures how easily students could comprehend the narratives.

2) The English in the narratives felt natural: This item assesses how natural the English in the narratives was compared to that found in a textbook.

3) The generated narratives were interesting: This item evaluates how engaged students felt while reading the stories.

4) The words used in the narratives are not too difficult: This item measures the perceived difficulty of individual words used in the narratives.

5) I was able to learn a new word from these narratives: This item assesses whether students felt they learned new vocabulary from the texts.

6) I would like to read more narratives like this: This item measures students' interest in reading similar stories in the future.

All survey questions were administered in Japanese. The items and descriptions were translated using ChatGPT 4.1 and subsequently proofread by a native speaker to ensure accuracy and clarity.

There are several limitations to the human experiment. First, only beginner-level materials were evaluated, so intermediate-level narratives were not assessed. Second, the evaluation focused exclusively on narratives generated using the RDP method.

E. Prompting Strategies

This research focuses on three prompting strategies:

- CEFR Keyword Constrained Prompting (CKCP): Prompts are designed to generate narratives using only words from designated CEFR levels, following approaches established in prior research [18]. CKCP operates by constraining the model to the vocabulary specified within the prompt. For 7th grade, only A1–A2 words are used, while for 10th grade, A1–B2 words are included. As a result, token usage increases substantially for higher grade levels. The Prompting method is represented in the Algorithm 1.

Algorithm 1: CKCP Prompt

You are a story creator assistant that specializes for {targeted_grade}th grade children. Your task is to craft a compelling story based on a user-provided title and a predefined list of verbs and nouns. Follow these instructions carefully:
Instructions
1. **Input Requirements**:
- The user provides only a title.
- You will also be given a list of verbs,nouns, pronoun, adverb and adjective. Use these words exclusively to create the story.

2. **Constraints**:
- Use only the verbs, nouns, adjective, adverb from the provided list.

- Articles (e.g., \"the,\" \"a\") and basic prepositions (e.g., \"in,\" \"on\") are allowed for readability.
- Verbs can be used in various tenses (past, present, future) as needed.
- Make it suitable for 7th grades
- Please use simple word and frequently used word only

3. **Output Format**:

- Make sure it was between 200-300 words
List of words:
Verbs : [a1-a2 CEFR J] or [a1-b1 CEFR J]
Adverbs : [a1-a2 CEFR J] or [a1-b1 CEFR J]
Nouns : [a1-a2 CEFR J] or [a1-b1 CEFR J]
Adjective : [a1-a2 CEFR J] or [a1-b1 CEFR J]

- Readability Driven Prompting: Prompts include two to three example stories at the target grade level, selected based on Flesch Reading Ease (FRE) and Dale–Chall (DC) scores falling within the desired thresholds. The prompt references both FRE and DC theories and provides explicit instructions to generate narratives within these specified readability ranges. However, we observed that GPT sometimes struggles to select appropriate vocabulary for each theory, especially at lower grade levels. To address this, selected words from the CEFR list with proven impact at specific grades were also referenced in the prompt. The example stories were taken from CKCP outputs that met our criteria for readability and lexical alignment. The Prompting method is represented in the Algorithm 2.

Algorithm 2: RDP Prompt

You are a children's story writer specializing in {targeted_grade}th-grade material.

Follow these rules strictly:
Vocabulary:

- Use only CEFR A1-A2 words.
- Avoid uncommon, academic, or abstract words.
- Use concrete, everyday language.
- Include slightly longer (2–3 syllable) words when appropriate to raise FRE.
- Avoid words above {targeted_grade}reading level (Dale-Chall).
- Example Forbidden words: {hard_word}
- Use simple alternatives: {alternative_word}

Sentence Structure:

- Mix short and compound sentences to increase words per sentence.
- Each sentence should ideally be 8–12 words when possible.
- Use conjunctions like 'and', 'but', 'so', 'because', 'while' to combine ideas.
- Add short descriptive phrases to expand sentences naturally.
- Keep sentences clear, active, and readable.
- Avoid complex subordinating clauses beyond simple compounds.

Story Guidelines:

- Word count: 200–300 words.

- Include curiosity, teamwork, problem-solving, or small adventures to engage readers.
- Keep paragraphs short.
- Include a clear beginning, middle, and end.
- Introduce small challenges or conflicts, then resolve them.

Readability Targets:

- Flesch-Kincaid Ease (FRE) {targeted_range}
- Dale-Chall Score {targeted_range}
- After writing, check FRE and DC internally. Redo if scores are outside range. Do not include scores in the story output.

Example:

[Stories example]

- Instruction-based Prompting: In this approach, GPT-4o-mini is instructed to generate a narrative for a specific grade and CEFR level without providing examples, theoretical guidance, or CEFR word constraints. The prompt simply requests the model to create a narrative tailored to the designated grade, specifying that the text

should be appropriate for the targeted difficulty level. The Prompting method is represented in the Algorithm 3.

Algorithm 3: IBP Prompt

You are a story creator assistant that specializes for {target_grade}. Follow these instructions carefully:

Instructions:

1. **Input Requirements**:

- The user provides only a title.

2. **Task**:

- Create a narratives for ESL. Mainly targeted {target_level} grade US grade or CEFR {target_cefr} level

3. **Output Format**:

- Make sure it was between 200-300 words

IV. Results

A. Lexical Analysis

Beginner narrative lexical analysis is presented in Table III. Although none of the methods achieved the 95% target, the results indicate that the proposed RDP technique performed the best, achieving 86.17% A1–A2 vocabulary usage. This was followed by CKCP at 84.84% and IBP at 82.89%. These findings suggest that narrative difficulty can be effectively adjusted using prompting techniques and RDP shown to increase the narrative qualities for a specific target level.

TABLE. III. Beginner Narrative Lexical Analysis

Method	Unknown Words	A1-A2 Words Ratio	Diff w/ target	Words Avg.
IBP	3%	82.89%	-12.11%	263 words
CKCP	2%	84.84%	-10.16%	287 words
RDP	2%	86.17%	-8.83%	336 words

Regarding unknown words, RDP demonstrates strong performance, producing only 2% unknown words, compared with 3% for IBP and matching the performance of CKCP. Notably, RDP also generates longer narratives, with an average length of 336 words, compared to 263 words for IBP and 287 words for CKCP. Achieving a low proportion of unknown words in longer texts is generally more challenging, as increased text length raises the likelihood of introducing unfamiliar vocabulary. Therefore, this result indicates that RDP achieves a higher level of lexical control and overall performance than the other methods.

TABLE. IV. Intermediate Narrative Lexical Analysis

Method	Unknown Words	A1-B1 Words	Diff w/ target	Words Avg.
IBP	5%	89.68%	-5.32%	267 words
CKCP	4.33%	90.41%	-4.59%	307 words
RDP	5.71%	88.43%	-6.67%	327 words

Intermediate narrative lexical analysis is presented in Table IV. Although RDP remains relatively close to the 95% target at the intermediate level, its performance shows a slight decline compared to IBP. IBP achieves 89.68% A1–B1 vocabulary usage, while RDP reaches 88.43%. Among the three methods, CKCP performs best, coming closest to the target with 90.41% A1–B1 vocabulary usage.

RDP also produces the highest proportion of unknown words, at 5.71%, compared with 5.0% for IBP and 4.33% for CKCP. In addition, RDP generates longer narratives, averaging 327 words, whereas IBP and CKCP produce 267 and 307 words, respectively. The combination of increased text length and greater lexical variation may contribute to the higher incidence of unknown words and the slight reduction in vocabulary alignment observed at the intermediate level. Overall, these results indicate a modest decline in RDP's performance when applied to intermediate-level narrative generation.

B. Readability Analysis

TABLE. V. 7TH Grade Focused Readability Analysis

Method	Flesch-Kincaid Ease	Dale-Chall	Flesch-Kincaid Target	Dale-Chall Target
IBP	68.63	7.05	70-79	6.0-6.9
CKCP	69.18	6.65	70-79	6.0-6.9
RDP	77.28	6.80	70-79	6.0-6.9

7th grade focused readability analysis is presented in Table V. RDP achieves a Flesch–Kincaid Reading Ease (FKE) score of 77.28, which falls within the targeted range. In contrast, CKCP does not meet the target, obtaining a score of 68.63, and IBP similarly falls short with a score of 68.63. These results indicate that RDP outperforms the other prompting methods in aligning narrative readability with the intended Flesch–Kincaid Ease target.

Regarding the Dale–Chall formula, both RDP and CKCP fall within the target band. RDP achieves a score of 6.80, while CKCP obtains 6.65. IBP fails to meet the target with a score of 7.05, suggesting that it produces narratives suitable for a higher grade level. These findings indicate that RDP achieves comparable performance to CKCP in terms of word familiarity control.

TABLE. VI. 10TH Grade Focused Readability Analysis

Method	Flesch-Kincaid Ease	Dale-Chall	Flesch-Kincaid Target	Dale-Chall Target
IBP	60.35	7.77	50-59	8.0-8.9
CKCP	60.31	7.99	50-59	8.0-8.9
RDP	58.98	7.61	50-59	8.0-8.9

10th grade focused readability analysis is presented in Table VI. RDP achieves a Flesch Reading Ease (FRE) score of 58.98, making it the only method that falls within the targeted range. In comparison, IBP and CKCP obtain scores of 60.35 and 60.31, respectively, both of which fall outside the desired range. These results indicate that RDP performs better in controlling overall text complexity as measured by the FRE formula.

With respect to the Dale–Chall formula, none of the methods reach the target range. RDP achieves a score of 7.61, while IBP and CKCP obtain scores of 7.77 and 7.99, respectively. The similarity of these scores suggests comparable performance across methods in terms of word familiarity at the 10th-grade level.

C. Human Evaluation

The first question aimed to assess the overall difficulty of the narratives. The mean score of 4.43 indicates that most participants agreed that the narratives were easy to understand. The second question evaluated whether the narratives felt natural, with a mean of 4.4, showing general agreement that the text was fluent and natural. The third question measured the narratives' level of interest, and the mean of 4.26 suggests that most participants found the stories engaging. The fourth question focused on word difficulty, yielding a slightly lower mean of 3.63. This indicates that while participants generally understood the vocabulary used by the RDP method, some words were challenging for certain users. The fifth question assessed whether participants could learn new words from the narratives, with a mean of 4.23, suggesting that the texts effectively supported incidental vocabulary learning. Finally, the sixth question measured engagement and willingness to read more narratives, resulting in a mean of 4.43, indicating that participants were motivated and interested in continued reading (see Table VII).

TABLE. VII. QUESTIONNAIRE RESULT

Question	1	2	3	4	5	Mean	SD
The narratives generated were easy to understand.	0	0	4	9	17	4.43	0.73
The English in the narratives felt natural.	0	0	6	6	18	4.4	0.81
The generated narratives were interesting.	0	1	2	15	12	4.26	0.73
The words used in the narratives are understandable.	2	4	6	9	9	3.63	1.24
I was able to learn a new word from these narratives	0	1	5	10	14	4.23	0.85
I would like to read more narratives like this	0	0	2	13	15	4.43	0.63

V. DISCUSSION

A. Lexical Analysis

The lexical analysis employed in this study is not proposed as a novel methodological contribution; rather, it serves as an evaluation lens for assessing the effectiveness of different prompting strategies. When examining the RDP results in comparison with the other methods, the incorporation of readability constraints—such as the Dale–Chall and Flesch–Kincaid formulas—clearly enhances GPT's ability to generate narratives that align more closely with targeted grade-level expectations. These findings are consistent with prior studies indicating that models such as GPT are unreliable when generating narratives without guidance [24], [25], and that they struggle to accurately assess grammatical and lexical difficulty [6]. By explicitly integrating readability theory, RDP helps mitigate these limitations and demonstrates stronger performance than the other prompting methods.

However, this effectiveness is primarily observed at the beginner level. At the intermediate level, RDP does not outperform IBP or CKCP and instead exhibits a slight decline in performance. This reduction may indicate a form of model confusion. At higher proficiency levels, RDP attempts to satisfy multiple theoretical constraints while simultaneously allowing greater lexical freedom, which can challenge the model's ability to balance competing objectives. As GPT naturally introduces increased vocabulary variation at intermediate levels, the additional constraints may interfere with one another, reducing the model's ability to consistently maintain vocabulary within the A1–B1 range.

While gains are modest at the intermediate level, the consistent beginner-level improvements are pedagogically meaningful, as early-stage readability mismatches have a disproportionate impact on learner motivation and comprehension.

Overall, the lexical analysis indicates that RDP is well-suited for early-grade narrative generation but encounters difficulties at the intermediate level. CKCP performs consistently at the beginner level and achieves the strongest results for intermediate narratives, while IBP demonstrates improved performance as the target grade level increases.

B. Readability Analysis

RDP demonstrates strong performance in controlling word complexity, as it is the only method that consistently falls within the targeted Flesch–Kincaid Reading Ease range of 70–79 for beginner-level narratives and 50–59 for intermediate-level narratives. This indicates that RDP effectively improves word complexity control across proficiency levels.

RDP also achieves favorable results on the Dale–Chall formula at the beginner level, falling within the 6.0–6.9 target range for beginner narratives. However, despite explicitly specifying the target grade and the model's theoretical understanding of the Dale–Chall formula, the generated outputs do not consistently meet the desired range for 10th-grade narratives. Notably, CKCP comes closest to the Dale–Chall target, with only a 0.01 deviation from the ideal band, whereas RDP exhibits the largest deviation at 0.39. This contrast highlights the different ways in which the prompting methods manage word familiarity and overall narrative complexity.

Overall, unlike the lexical analysis results, the readability analysis reveals that CKCP and RDP exhibit distinct and complementary strengths. CKCP aligns more closely with Dale–Chall expectations, while RDP performs best on the Flesch–Kincaid scale. These findings suggest that generating grade-appropriate narratives may require different prompting strategies depending on which readability dimension, structural complexity or word familiarity, is prioritized.

C. Human Evaluation

Overall, the human evaluation results demonstrate that narratives generated using the RDP method received positive feedback and performed well as learning materials. Participants generally found the texts understandable, natural, and engaging,

while also providing opportunities for learning new words. These results suggest that RDP is effective for creating targeted, learner-friendly narratives that balance readability, vocabulary, and interest.

VI. Conclusion

This study investigated the effectiveness of three prompting techniques: IBP, CKCP, and RDP. Three types of analysis were conducted. The first was lexical analysis, which showed that RDP performed best for early-grade narratives by producing a higher proportion of A1–A2 words and closely aligning with the intended vocabulary range. However, at the intermediate level, RDP's performance declined, likely due to the model's need to balance multiple constraints, which sometimes resulted in the use of less familiar words. CKCP demonstrated consistent performance across both levels, while IBP improved as the target grade increased.

Second, the readability analysis further highlighted the complementary strengths of the prompting techniques. RDP generated narratives that aligned closely with Flesch–Kincaid targets at both the 7th and 10th grade levels, indicating effective control over sentence structure and syllable complexity. In contrast, CKCP more closely matched the Dale–Chall thresholds, demonstrating stronger control over word familiarity. These findings suggest that different prompting strategies may be required depending on which dimension of readability is prioritized.

Finally, a human evaluation survey was conducted. The results confirmed that narratives generated by RDP were generally understandable, natural, engaging, and conducive to vocabulary learning. Participants reported positive experiences and expressed willingness to read more narratives, indicating that RDP can effectively produce learner-friendly texts. However, slightly lower scores for word understandability correspond with the lexical analysis, suggesting that some vocabulary may still pose challenges for certain learners.

In summary, the results indicate that RDP is a promising approach for generating grade-targeted narratives, particularly for beginner learners and when controlling for structural readability. CKCP and IBP also provide complementary benefits, depending on the target proficiency level and the specific readability metric being prioritized. Overall, these findings demonstrate that carefully designed prompting techniques can enhance large language model outputs for educational purposes by balancing lexical appropriateness, readability, and learner engagement.

Our findings suggest several promising avenues for future research. First, given that RDP did not outperform CKCP at intermediate proficiency levels, future work should explore adaptive prompt strategies that dynamically balance readability constraints as task complexity increases. This might involve reinforcement learning or prompt chaining to better harmonize competing metrics. Second, while this study focused on narrative generation for EFL Japanese learners, extending RDP to other languages, learner backgrounds, and genres (e.g., expository or dialogic texts) would help assess its generalizability and cultural adaptability. Finally, the current human evaluation was limited to beginner-level narratives with a small sample size; future studies should involve a broader range of proficiency levels and larger, possibly classroom-based longitudinal evaluations to evaluate educational impact.

Acknowledgments

This work was partially supported by the Research Institute for Science and Technology of Tokyo Denki University, Grant Number Q25D-10 / Japan

References

[1] J. Reeve, R. M. Ryan, S. H. Cheon, L. Matos, and H. Kaplan, *Supporting Students' Motivation: Strategies for Success*, 1st ed. London, U.K.: Routledge, 2022. [Online]. Available: https://doi.org/10.4324/9781003091738.

[2] A. Zainal, A. Gustina, R. Risnawaty, I. Hassan, R. Rt. Bai, and M. Febriani Sya, "The comparative effect of using original short stories and local short stories as two types of cultural sources on Indonesian EFL learners' reading comprehension," Int. J. Soc., Cult. Lang., vol. 10, no. 1, pp. 143–152, 2022. [Online]. Available: https://doi.org/10.22034/ijscl.2021.247370.

[3] L. P. Cabanilla García, L. F. Pereddo Hidalgo, and T. G. Pineda Guzmán, "Enhancing EFL Young Learners' Vocabulary Through Online Extensive Reading (ER) and Visual Strategies: An Action Research Study," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 8, no. 6, pp. 10227-10244, 2025. [Online]. Available: https://doi.org/10.37811/cl_rcm.v8i6.15682.

[4] Y.-H. Yang, H.-C. Chu, and W.-T. Tseng, "Text difficulty in extensive reading: Reading comprehension and reading motivation," Reading in a Foreign Language, vol. 33, no. 1, pp. 78–102, 2021. [Online]. Available: https://doi.org/10.64152/10125/67394.

[5] G. Melzi, A. R. Schick, and C. Wuest, "Stories Beyond Books: Teacher Storytelling Supports Children's Literacy Skills," *Early Education and Development*, 2022 [Online]. Available: https://doi.org/10.1080/10409289.2021.2024749.

[6] C. K. Lo, P. L. H. Yu, S. Xu, D. T. K. Ng, and M. S. Y. Jong, "Exploring the Application of ChatGPT in ESL/EFL Education and Related Research Issues: A Systematic Review of Empirical Studies," *Smart Learning Environments*, vol. 11, no. 1, 2024. [Online]. Available: https://doi.org/10.1186/s40561-024-00342-5.

[7] B. V. Durazno Abril, K. M. Díaz Carlosama, and J. F. Zambrano Pachay, "Exploring the Effect of ChatGPT as an Educational Tool to Improve Vocabulary Acquisition in the English Language," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 9, no. 1, pp. 4874-4886, 2025. [Online]. Available: https://doi.org/10.37811/cl_rcm.v9i1.16190.

[8] Z. Y. Zeng, L.-J. Kuo, L. Chen, J.-A. Lin, and H. Shen, "Vocabulary Instruction for English Learners: A Systematic Review Connecting Theories, Research, and Practices," Education Sciences, vol. 15, no. 3, Art. 262, 2025. [Online]. Available: https://doi.org/10.3390/educsci15030262.

[9] X. Hu and V. Aryadoust, "A systematic review of eye-tracking technology in second language research," Languages, vol. 9, no. 4, Art. no. 141, 2024. [Online]. Available: https://doi.org/10.3390/languages9040141.

[10] A. Pellicer-Sánchez, S. Webb, and A. Wang, "How does lexical coverage affect the processing of L2 texts?" Appl. Linguist., vol. 45, no. 6, pp. 953–972, 2024. [Online]. Available: https://doi.org/10.1093/applin/amae062 .

[11] B. Melani, S. Willian, K. Apgrianto, and H. Lail, "Vocabulary coverage and reading comprehension of university EFL learners," in *Proc. Thirteenth Conf. Appl. Linguistics*, 2021, pp. --. [Online]. Available: https://doi.org/10.2991/assehr.k.210427.010.

[12] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Flesch Reading Ease, and Flesch-Kincaid Grade Level)," Research Branch Report No. 8-75, U.S. Navy, 1975.

[13] E. Dale and J. S. Chall, "A Formula for Predicting Readability," *Educational Research Bulletin*, vol. 27, no. 1, pp. 11–20, 1948.

[14] M. Trott and P. Rivière, "Measuring and Modifying the Readability of English Texts with GPT-4," Proceedings of the Workshop on Natural

Language Generation and Evaluation, pp. 23–35, 2024. [Online]. Available: https://aclanthology.org/2024.tsar-1.13.pdf.

[15] P. Martínez, A. Ramos, and L. Moreno, "Exploring Large Language Models to Generate Easy-to-Read Content," Frontiers in Computer Science, vol. 6, Article 1394705, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1394705/full.

[16] J. Preschern, M. Hunter, I. Born-Lechleitner, and C. Kobylak, "Developing prompt engineering as a 21st-century skill: The impact of structured ChatGPT instruction in EFL education," GILE J. Skills Dev., vol. 5, no. 3, pp. 87–108, 2025. [Online]. Available: https://doi.org/10.52398/gjsd.2025.v5.i3.pp87-108.

[17] J. Han et al., "RECIPE: How to integrate ChatGPT into EFL writing education," in Proc. Tenth ACM Conf. Learning @ Scale (L@S '23), New York, NY, USA: ACM, 2023, pp. 416–420. [Online]. Available: https://doi.org/10.1145/3573051.3596200.

[18] R. Marbun and M. Shishido, "Enhancing Narrative Generation in ESL: Tailored Prompting for Proficiency-Specific Learning," in *The European Conference on Education 2025: Official Conference Proceedings*, pp. 145-155, 2025. [Online]. Available: https://doi.org/10.22492/issn.2188-1162.2025.13.

[19] G. Ramesh, M. Sahil, S. A. Palan, et al., "A review on NLP zero-shot and few-shot learning: Methods and applications," Discover Applied Sciences, vol. 7, Art. no. 966, 2025. [Online]. Available: https://doi.org/10.1007/s42452-025-07225-5.

[20] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, and H. Hajishirzi, "Generated knowledge prompting for commonsense reasoning," arXiv preprint arXiv:2110.08387, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2110.08387.

[21] D. Rooein, P. Röttger, A. Shaitarova, and D. Hovy, "Beyond Flesch–Kincaid: Prompt-based metrics improve difficulty classification of educational texts," in Proc. 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico, 2024, pp. 54–67. [Online]. Available: https://aclanthology.org/2024.bea-1.5/.

[22] Eurydice, "Key Data on Teaching Languages at School in Europe – 2023 Edition," Publications Office of the European Union, Luxembourg, 2023. [Online]. Available: https://eurydice.eacea.ec.europa.eu/publications/key-data-teaching-languages-school-europe-2023-edition.

[23] Ministry of Education, Culture, Sports, Science and Technology (MEXT), "Course of Study for Foreign Languages (English)," Japan, 2017/2018.

[24] P. Sahoo, A. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," 2024. [Online]. Available: https://doi.org/10.13140/RG.2.2.13032.65286.

[25] T. S. Wang and A. S. Gordon, "Playing Story Creation Games with Large Language Models: Experiments with GPT-3.5," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14372, pp. 381–393, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-47658-7_28.