The Effect of Prompting Techniques on Level-Targeted Narrative Generation with GPT Models

Ronald William Marbun[1] Makoto Shishido[2]
Tokyo Denki University
24udc92@ms.dendai.ac.jp[1] shishido@mail.dendai.ac.jp[2]

Abstract

Narrative-based learning is an effective approach for language acquisition. However, generating level-appropriate narratives using GPT remains challenging, as the model's sensitivity to input does not guarantee consistent grade-level control. This study investigates whether specific prompting techniques can guide GPT to produce narratives suitable for beginner-level learners. We compare Chain-of-Thought (CoT) and Tree-of-Thought (ToT) prompting strategies, with Zero-Shot prompting as a baseline. Generated narratives are evaluated using lexical profiling and readability metrics, specifically the Flesch–Kincaid Reading Ease and Dale–Chall formulas. Results indicate that prompt engineering, through CoT and ToT, yields only modest improvements in readability and grade-level alignment over the baseline. These findings underscore the need for more specialized prompting methods to reliably generate narratives tailored to specific learner levels.

1 Introduction

Narratives are effective learning materials because they can improve students' reading, speaking, listening, and writing skills while enhancing retention and engagement [1],[2]. Evidence indicates that stories benefit both EFL learners and native speakers [3]. At the same time, narratives can be difficult for learners and crafting level-appropriate stories that balance engagement with linguistic accessibility requires expertise [4].

Large language models (LLMs) have recently advanced in generating fluent, audience-aware text. Their educational use remains debated [5], yet many educators are exploring ways to adapt these tools for instruction. GPT-based models, in particular, are sensitive to input phrasing and context [6], which has motivated prompting strategies to steer outputs toward desired qualities [7],[8]. Nevertheless, automated story generation remains inconsistent, despite its promise for educational content [9].

Because narrative writing entails creativity, the effectiveness of prompt engineering for controlling level and quality is still unclear. We therefore investigate whether prompting strategies can improve the suitability of LLM-generated narratives specifically for CEFR level A2. We focus on A2 because learners at this stage can understand short narratives about familiar, everyday topics but are highly sensitive to lexical load and sentence complexity [11], and small increases in out-of-level vocabulary or clause density can quickly reduce comprehensibility (Nation, 2006).

We compare a zero-shot baseline with Chain-of-Thought (CoT) and Tree-of-Thought (ToT) prompting, using CoT/ToT as planning aids (e.g., constraining vocabulary and setting sentence-length targets) while excluding the reasoning traces from the final output [8],[10]. To ensure a fair comparison, we hold the model and generation parameters constant across conditions. Outputs are assessed using lexical profiling and readability checks.

2 Methodology

2.1 Prompting Method

1. **Zero Shot:** Zero-shot prompting provides GPT with only an instruction, without examples or reasoning steps. The model generates responses based solely on its prior knowledge. In this study, zero-shot prompting serves as the baseline to evaluate GPT's ability to produce grade-targeted narratives without guided reasoning.

2. **Tree-of-Thought (ToT):** This prompt method guides GPT to generate multiple reasoning paths before selecting the most contextually appropriate one. This approach allows the model to explore alternative ideas and evaluate them systematically, leading to more coherent and context-aware outputs [8].

3. **Chain-of-Thought (CoT):** This prompting method encourages GPT to produce step-by-step reasoning before generating a final answer. By explicitly outlining its thought process, the model can generate more logical and coherent narratives [9].

2.2 Narratives Generation & Evaluation

The narratives generation process will happen 100 times per prompting method (totaling 300 times). First, 100 titles are generated randomly using GPT 4.1 (gpt-4.1-2025-04-14). Then using the same model every prompting method will generate exactly 100 narratives.

Readability refers to the ease with which a reader can comprehend a written text. Readability formulas consider factors such as sentence length, word complexity, and overall structure. These metrics are important for determining whether a text is appropriate for its intended audience. Two readability formulas were employed in this study:

1. **Flesch–Kincaid Reading Ease (FKE)**. Developed by Rudolf Flesch and J. Peter Kincaid in 1948, this metric evaluates text difficulty based on the average sentence length and the average number of syllables per word. The resulting score indicates how easily a passage can be understood by readers. Sometimes longer syllables do not mean harder content and yet a research findings show there is difference of reaction time whenever humans see longer syllables, indicating that syllables do contribute to word difficulties and shifting the reader's attention making it better for learning [13].

2. **Dale–Chall Readability Formula (DC)**. Proposed by Edgar Dale and Jeanne Chall in 1948, this formula measures readability using a list of familiar words and average sentence length to estimate a grade-level score. It primarily assesses the difficulty of the vocabulary used in a text. It operates word difficulty through familiarity lists that strongly correlate with phonological and morphological complexity. In psycholinguistic terms, multisyllabic and less frequent words impose greater processing effort during lexical access and decoding [13].

Lexical ability represents one of the most salient distinctions between native speakers and learners of English as a Foreign Language (EFL). A research finds that emphasize, native speakers typically possess a more extensive and contextually nuanced lexical repertoire, which allows them to process and produce language with greater automaticity and flexibility [13]. In contrast, EFL learners often rely on a more limited vocabulary base, which constrains both comprehension and expressive fluency, particularly when encountering texts with high lexical density or unfamiliar collocations.

Achieving at least 95% lexical coverage of a text constitutes the optimal threshold for effective language comprehension and learning [14]. Below this level, the cumulative effect of unknown words can significantly impede the construction of meaning, reduce inferencing ability, and increase cognitive load during reading. Consequently, instructional methods and text generation techniques that can produce materials approaching or surpassing this 95% coverage threshold are considered more pedagogically suitable for EFL learners. Such methods not only enhance accessibility but also facilitate incremental vocabulary acquisition, enabling learners to bridge the lexical gap between non-native and native proficiency over time.

For the lexical analysis, all textual data underwent a systematic preprocessing pipeline to ensure accuracy and comparability across samples. Initially, the data were normalized and tokenized to segment the text into discrete lexical units. Following tokenization,

nonlinguistic and redundant elements—such as punctuation marks, numerical symbols, and stopwords—were removed through a filtering process to eliminate noise that could distort lexical frequency counts.

The remaining tokens were subsequently lemmatized, reducing inflected forms to their canonical base forms to allow for more consistent lexical comparison. The lemmatized tokens were then cross-referenced against the CEFR-J word list, a lexically graded corpus specifically developed to reflect the vocabulary progression of Japanese EFL learners. In this framework, lexical items categorized within the A1–A2 bands were identified as beginner-level vocabulary, representing foundational linguistic competence. Conversely, words absent from the CEFR-J list were operationally classified as advanced, corresponding to the C1–C2 proficiency range, thereby indicating a higher level of lexical sophistication and reduced pedagogical accessibility for lower-proficiency learners.

2.3 Limitation

This study is limited by its reliance on GPT-4.1, and results may not generalize to other large language models or future versions. Only three prompting techniques Zero-Shot, Chain-of-Thought, and Tree-of-Thought—were evaluated. Alternative or hybrid prompting strategies may produce different results. Evaluation relied on Flesch–Kincaid, Dale–Chall, and CEFR-J lexical profiling, which measure readability and vocabulary but do not assess narrative coherence, creativity, or learner engagement. The use of randomly generated titles may have introduced variability in narrative complexity and topic difficulty. Additionally, the CEFR-J lexicon may misclassify rare or domain-specific words, affecting level-targeting accuracy. Finally, the study focused exclusively on EFL learners in Japan, limiting the generalizability of findings to other learner populations or languages

3 Results and Discussion

3.1 Lexical Analysis

Tables 1 and 2 summarize the lexical analysis outcomes across prompting methods. Table 1 shows mean understanding and unknown word ratios, which reveal minimal differences at the aggregate level (Zero-Shot: 88.50%, Tree-of-Thought: 88.49%, Chain-of-Thought: 89.51%). In this study, "understanding" refers to the proportion of words in each narrative that fall within the A1–A2 CEFR-J levels, representing beginner-level vocabulary for EFL learners.

Table 1: Lexical Analysis Result (Mean)

Methods	Statistics		
	A1-A2 (mean)	Unknown Words Ratio (mean)	Targets
Zero Shot	88.50%	1.6%	95%
Tree of Thought	88.49%	1.7%	
Chain of Thought	89.51%	1.8%	

Table 2: Lexical Analysis Result (Frequency)

Methods	Understanding	Frequency
Zero Shot	93% - 97%	6 Narratives
	87% - 93%	61 Narratives
	< 87%	33 Narratives
Tree of Thought	93% - 97%	8 Narratives
	87% - 93%	62 Narratives
	< 87%	30 Narratives
Chain of Thought	93% - 97%	9 Narratives
	87% - 93%	68 Narratives
	< 87%	23 Narratives

However, categorizing narratives into understanding bands (Table 2) highlights the impact of prompting strategies. Both Chain-of-Thought and Tree-of-Thought methods slightly increased the number of narratives in the higher understanding bands (87–93% and 93–97%) compared to Zero-Shot. For instance, Chain-of-Thought produced nine narratives in the 93–97% band, compared to six for Zero-Shot. This pattern indicates that structured prompting can enhance the proportion of narratives composed predominantly of beginner-level vocabulary, even when mean scores appear similar. These results suggest that refining prompting methods may further improve the generation of level-targeted materials for EFL learners.

3.2 Readability Analysis

Table 3: Readability Analysis (Mean)

Methods	Evaluation	Statistics	
		Mean	Target
Zero Shot	FKE	80.33	70-100
	DC	6.39	<= 6.9
Tree of Thought	FKE	73.38	70-100
	DC	6.12	<= 6.9
Chain of Thought	FKE	78.61	70-100
	DC	6.01	<= 6.9

Table 4: Readability Analysis (Frequency)

Methods	Evaluation	Statistics	
		Value	Frequency
Zero Shot	FKE	< 70	4 Narratives
	DC	> 6.9	19 Narratives
Tree of Thought	FKE	< 70	26 Narratives
	DC	> 6.9	6 Narratives
Chain of Thought	FKE	< 70	9 Narratives
	DC	> 6.9	4 Narratives

Tables 3 and 4 present the readability results. Table 3 shows mean Flesch–Kincaid Ease (FKE) and Dale–Chall (DC) scores, which fall within target ranges for all methods (FKE: 73.38–80.33; DC: 6.01–6.39), confirming that GPT-4.1 can generally produce beginner-level narratives.

Frequency analysis (Table 4) reveals differences among prompting methods. Zero-Shot produced more narratives outside the desired readability ranges when using DC meanwhile it creates more desired readability range when using FKE (FKE < 70: 4 narratives; DC > 6.9: 19 narratives) while Chain-of-Thought create more desired value of DC. (FKE < 70: 9; DC > 6.9: 4) and Tree-of-Thought (FKE < 70: 26; DC > 6.9: 6) is underperforming when evaluated using FKE but doing pretty good on DC. These results indicate that structured prompting influences the distribution of readability scores, increasing the likelihood of narratives falling within the target range. Consequently, more precise and targeted prompting may further enhance the readability of beginner-level EFL narratives.

4 Conclusion

This study investigated the effect of prompting techniques Zero-Shot, Chain-of-Thought, and Tree-of-Thought on generating beginner-level narratives for EFL learners. Lexical analysis showed that structured prompting slightly increased the proportion of narratives composed of A1–A2 level vocabulary, while readability analysis demonstrated that Chain-of-Thought and Tree-of-Thought methods produced more narratives within the target Flesch–Kincaid and Dale–Chall ranges. These findings suggest that structured prompting can improve the alignment of generated narratives with intended proficiency levels, even when mean scores show minimal differences.

However, the results also indicate that further refinement of prompting strategies is needed to consistently achieve the optimal 95% comprehension threshold for vocabulary and ensure readability. Overall, this research highlights the potential of prompt engineering to generate more effective, level-targeted EFL learning materials and provides a foundation for future work exploring advanced prompting strategies, model variations, and broader evaluation metrics.

5 Acknowledgements

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

References

[1] Andrew Simmons. **Whoever Tells the Best Story Wins: How to Use Your Own Stories to Communicate with Power and Impact**. AMACOM, 2007.

[2] M. Padmavathamma. Learning through stories. **Learning Curve**, pp. 40–43, 2023. https://doi.org/10.12968/prtu.2013.1.17.18.

[3] Phil Hiver, Shannon Zhou, Soroush Tahmouresi, Yu Sang, and Mostafa Papi. Why stories matter:

Exploring learner engagement and metacognition through narratives of the L2 learning experience. **System**, 2020. https://doi.org/10.1016/j.system.2020.102260.

[4] Gabriella Melzi, Alyssa R. Schick, and Christine Wuest. Stories beyond Books: Teacher Storytelling Supports Children's Literacy Skills. **Early Education and Development**, Vol. 34, No. 2, pp. 485–505, 2022. https://doi.org/10.1080/10409289.2021.2024749.

[5] Ricardo Michel-Villarreal, Edgar Vilalta-Perdomo, Diana E. Salinas-Navarro, Ricardo Thierry-Aguilera, and Fotini S. Gerardou. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. **Education Sciences**, Vol. 13, No. 9, p. 856, 2023. https://doi.org/10.3390/educsci13090856.

[6] Vipula Rawte, Prachi Priya, S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Amit P. Sheth, and Amitava Das. Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness. **arXiv preprint**, 2023. https://doi.org/10.48550/arXiv.2309.11064.

[7] Jiachang Liu, Alon Alon, Xiangru Tang, Sean Welleck, Peter West, Ronan Le Bras, and Hannaneh Hajishirzi. Generated Knowledge Prompting for Commonsense Reasoning. **Proceedings of the Annual Meeting of the Association for Computational Linguistics**, 2022. https://doi.org/10.18653/v1/2022.acl-long.225.

[8] Shanshan Cao, Jinyang Zhang, Jiaqi Shi, Xinyu Lv, Zhiyuan Yao, Qi Tian, and Lei Hou. Probabilistic Tree-of-thought Reasoning for Answering Knowledge-intensive Complex Questions. **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.835.

[9] Tony S. Wang and Andrew S. Gordon. Playing Story Creation Games with Large Language Models: Experiments with GPT-3.5. **Lecture Notes in Computer Science**, 2023. https://doi.org/10.1007/978-3-031-47658-7_28.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Edward H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. **Proceedings of the 36th International Conference on Neural Information Processing Systems**, pp. 24824–24837, 2022.

[11] Council of Europe. **Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume**. Council of Europe Publishing, 2020.

[12] Aisha A. Almabruk. Syllabic Length Effect in Visual Word Recognition. **International Journal of Applied Linguistics & English Literature**, Vol. 12, No. 1, pp. 73–78, 2023. https://doi.org/10.7575/aiac.ijalel.v.12n.1p.73.

[13] Mahsa Nasseri and Paul Thompson. Lexical Density and Diversity in Dissertation Abstracts: Revisiting English L1 vs. L2 Text Differences. **Assessing Writing**, Vol. 47, p. 100511, 2021. https://doi.org/10.1016/j.asw.2020.100511.

[14] I. S. P. Nation. How Large a Vocabulary Is Needed for Reading and Listening? **Canadian Modern Language Review**, Vol. 63, No. 1, pp. 59–82, 2006. https://doi.org/10.3138/cmlr.63.1.59.