Enhancing Narrative Generation in ESL: Tailored Prompting for Proficiency-Specific Learning

Ronald William Marbun, Tokyo Denki University, Japan
Makoto Shishido, Tokyo Denki University, Japan

The European Conference on Education 2025
Official Conference Proceedings

Abstract

Narratives can be powerful tools for teaching English as a Second Language (ESL), but they must be carefully tailored to specific proficiency levels to be effective. This paper evaluates the capability of large language models (LLMs) to generate level-specific narratives and introduces a novel prompting method designed to enhance narrative generation for distinct educational levels. The method leverages English profiling frameworks, such as the Commoerbs, nouns) based on target levels. Usinn European Framework of Reference (CEFR), by introducing and restricting word forms (e.g., vg the Flesch-Kincaid and Dale-Chall readability frameworks, the study adopts a quantitative approach to assess efficacy. Narratives were generated for two proficiency levels: 7th grade (elementary) and 10th grade (intermediate). To evaluate the efficiency of the proposed method, it was compared to the widely used Instruction and Role-based Zero-Shot Prompting approach. Additionally, the study examined its performance when integrated with complementary techniques, such as the Tree of Thought method. Results demonstrate that the novel method improves narrative performance by 63% for elementary levels and 20% for intermediate levels. While the Tree of Thought method did not enhance efficiency, it contributed to a better balance of difficult word usage.

Keywords: narrative generation, ESL, LLM, educational technology, language education, prompting technique, GPT

iafor
The International Academic Forum
www.iafor.org

Introduction

To achieve fluency in English, one must master four key aspects: reading, speaking, writing, and listening (Ali, 2022). In order to master these skills, students must actively practice their English. However, effective practice depends on access to appropriate learning materials. Stories are an excellent learning resource because they engage all four aspects of language acquisition (Simmons, 2007) and are particularly effective in capturing the reader's interest (Wang & Lee, 2007). Stories also provide valuable context for the reader and are especially effective in sticking in people's memories. In fact, even adults can still recall stories told to them when they were young, highlighting the lasting impact of storytelling.

Large Language Models (LLM) are a disruptive technology that has gained popularity, particularly in education, due to their potential. GPT is the most popular LLM and has proven to be an effective tool in the educational world. While GPT cannot fully replace the role of a teacher, it serves as a valuable tool in language education (Chan & Tsi, 2023), and GPT can be used to create learning material. The generated learning materials have been shown to match or even surpass those found in school databases (Namilae & Leddo, 2024) while adding other advantages such as personalized learning (Hatmanto & Sari, 2023). Because of its ability to generate unlimited personalized content, GPT can be leveraged to create tailored stories for language learners. But in order to ensure contextually appropriate output, effective prompting techniques are essential.

Designing a narrative suitable for a specific audience level is challenging and requires a properly trained educator (Melzi et al., 2023). Furthermore, the narrative is created for ESL students, which means they have limited proficiency in the English language. Giving ESL students material that is difficult to read may overwhelm learners and discourage them because their English proficiency may still be at an elementary or middle school level, requiring carefully tailored resources.

Several frameworks exist to grade English proficiency, such as the Common European Framework of Reference (CEFR). This framework has been proven effective in text classification (Gaillat et al., 2022) and verb classification (Milton & Alexiou, 2020), which leads to determining learners' abilities. These profiling methods will play a critical role in boosting LLM potential and providing more context for better segmentation targeting. The knowledge from the CEFR will be used as the context and knowledge for narrative generation. The readability of the generated stories will be assessed using established metrics, such as the Flesch-Kincaid and Dale-Chall formulas.

This study will also examine the adaptability of prompting methods in conjunction with other prompting techniques. Specifically, it will compare the effectiveness of the Tree of Thought method with CEFR-based profiling in narrative generation. Tree of Thought is selected because it enhances creativity by allowing models to explore multiple options (Yao et al., 2023), making it a suitable approach for refining narrative content.

ChatGPT in Education

ChatGPT is a large language model (LLM) that utilizes natural language processing (NLP) technologies to engage in real-time interactions and provide personalized responses (OpenAI, 2022). Recent studies have explored ChatGPT’s applications in education, particularly in second-language learning, where it has shown promising results through unique approaches.

For instance, ChatGPT is used as a conversation partner (Young & Shishido, 2023) and as a teaching assistant specifically to prepare materials for learning (Mikeladze, 2023). Furthermore, GPT-personalized material is very good and has been proven to be usable in the education world (Pataranutaporn et al., 2021).

However, despite these advancements, research indicates that optimizing prompt design remains a significant challenge. One study suggests that while GPT-4o exhibits strong conversational capabilities, its impact on real-life communication still requires further investigation, particularly in assessing grammar and vocabulary accuracy (Lo et al., 2024). These findings highlight the need for structured prompting techniques to improve ChatGPT's effectiveness in language education, ensuring that its outputs align with specific learning objectives.

Narrative Generation

Studies have explored how models like GPT can create coherent and engaging narratives, particularly in interactive storytelling and game development. For instance, Wang and Gordon (2023) leveraged GPT to generate dynamic storylines for a simple Story Creation Game by using GPT-3.5. This study shows that the stories generated by GPT incorporated all elements of the story and were particularly easy to read. Similarly, studies by Tian et al. (2024) attempted to compare human-generated plots by dividing them into various plot points, and this study showed that GPT has poor pacing issues and is focused on positive endings. Despite that, both studies demonstrated the potential of narrative generation using GPT.

While much of the existing research focuses on plot coherence and structure, its application in language education, particularly for ESL learners, remains underexplored. Additionally, prior studies indicate that while GPT is a powerful storytelling tool, its outputs can be diverse and unpredictable (Lo et al., 2024). These challenges highlight the need for novel prompting techniques to ensure that AI-generated narratives align with specific educational objectives, making them more effective for structured learning environments.

Prompt Engineering

Prompt engineering is a technique used to refine AI-generated outputs by providing structured input that guides the model's responses (Schmidt et al., 2024). It plays a crucial role in optimizing language models, ensuring that they generate contextually appropriate and high-quality results (Giray, 2023). While some research suggests that role-based prompting is less relevant for creative writing, it has been shown to improve performance compared to zero-shot prompting (Kong et al., 2024).

Despite these findings, the potential of role-based prompting in language learning remains underexplored, particularly when targeting specific proficiency levels. Additionally, Tree-of-Thought (ToT) prompting, which enhances model reasoning by exploring multiple possibilities, has demonstrated improvements in AI-generated outputs (Yao et al., 2023). Our previous research indicates that while prompting techniques significantly influence narrative generation, their effectiveness in adapting AI-generated content to specific educational levels has been limited. This limitation underscores the need for new prompting frameworks that align with structured learning objectives, ensuring that AI-generated narratives effectively

support language acquisition. To address this gap, we will evaluate the compatibility of various prompting techniques for educational applications.

To further refine prompt design, this research will incorporate meta-prompting, a technique that autonomously generates prompts by leveraging the LLM itself (Tamenaoul et al., 2024). Meta-prompting has proven effective as it enables the model to act as both a conductor and an expert, facilitating a more structured and adaptive approach to prompt optimization (Suzgun & Kalai, 2024).

Keyword Targeting

Focusing on specific keywords is an effective method for enhancing learning, particularly when introducing new concepts. Various techniques have explored the use of keywords as a means of structuring learning experiences. For example, research utilizing the Zone of Proximal Development (ZPD) has demonstrated the importance of targeting new vocabulary within the learner's developmental range (Zaretsky, 2021). Similarly, the Spaced Repetition System (SRS) has been employed to reinforce key vocabulary through frequent review, ensuring retention over time (M & G, 2024). Additionally, vocabulary grading frameworks, such as those based on the Common European Framework of Reference for Languages (CEFR), offer structured approaches to categorizing words according to learner proficiency levels.

Methodology

The goal of this study is to introduce a prompting technique that targets a specific level of learning and to evaluate its impact on text readability while testing its adaptability with another prompting technique. Tree of Thought prompting technique will be used and mainly focused to enhance the story. To measure readability, this research employs two established formulas: Flesch-Kincaid Ease (FKEF) and Dale-Chall Formula (DCF). FKEF primarily considers syllable count (Flesch, 1948), where a lower FKEF score indicates a more difficult-to-read text. In contrast, DCF focuses on word difficulty (Dale & Chall, 1948), where a higher DCF score signifies increased complexity. Both formulas are used to ensure a comprehensive readability assessment, as syllable count alone does not always correlate with difficulty.

Prior research indicates that GPT's default output aligns mostly with college-level readability based on FKEF and DCF. Thus, this study focuses on targeting 7th-grade (early level) and 10th-grade (intermediate level) texts.

The data collection will follow these methods:

1. *Title Generation:* 100 story titles will be generated randomly using GPT-4o.
2. *Vocabulary Assignment:* Each title will be paired with a grade-appropriate keyword that must be introduced to the reader selected from a list of keywords from CEFR collected from CEFR-J list.
3. *Keyword Restriction:* List of Verb, Adverb, Adjective will be listed for restriction to ensure GPT fully maximizes CEFR Keyword. Suitability will be determined using CEFR levels (A1 for early level, B2 for intermediate level).
4. *Text Generation:* GPT-4o was chosen for its ability to generate structured and coherent text, making it well-suited for evaluating readability across different educational levels. Each story will be generated three times using distinct approaches:

(1) Zero-Shot prompting, (2) the new prompting method (PCEFR), and (3) PCEFR combined with Tree of Thought (PCToT).

5. *Readability Evaluation:* The generated narratives will be analyzed using Flesch-Kincaid Ease (FKEF) and Dale-Chall Formula (DCF). The results will be compared against a baseline (e.g., GPT-4o default output) to assess the impact of the proposed prompting technique. And to measure its performance, every readability pass will be calculated and represented in percentage.

To illustrate the process of title generation and keyword assignment, Figure 1 presents an example. Each title is assigned verbs at two levels to apply ZPD.

Figure 1
Example of the Data

Title: A girl discovers a hidden talent for dance that surprises everyone.
Early level verb: dance, find, enjoy
Intermediate level verb: surprise, demonstrate, express

Then based on the information on Figure 1 prompting will be designed according to their approach:

1. *Zero Shot:* A role-play based prompting and instruction. The instruction is focused on creating a narrative and introducing verbs based on Figure 1 to teach new verbs for students.
2. *PCEFR:* PCEFR using the same approach with Zero Shot. The difference is every CEFR keylist (A1 for beginner and A1-B1 for intermediate) is listed by using prompt and adding constraint to the instruction to not include any words that are not included in the CEFR keylist.
3. *PCToT:* PCToT using the same approach with PCEFR. Because this one focused on checking PCEFR adaptability with other prompting methods, the instruction is using the Tree of Thought approach for the 3 main parts of the story (Introduction, Key points, Ending).

Limitation

This research presents several notable limitations that must be considered:

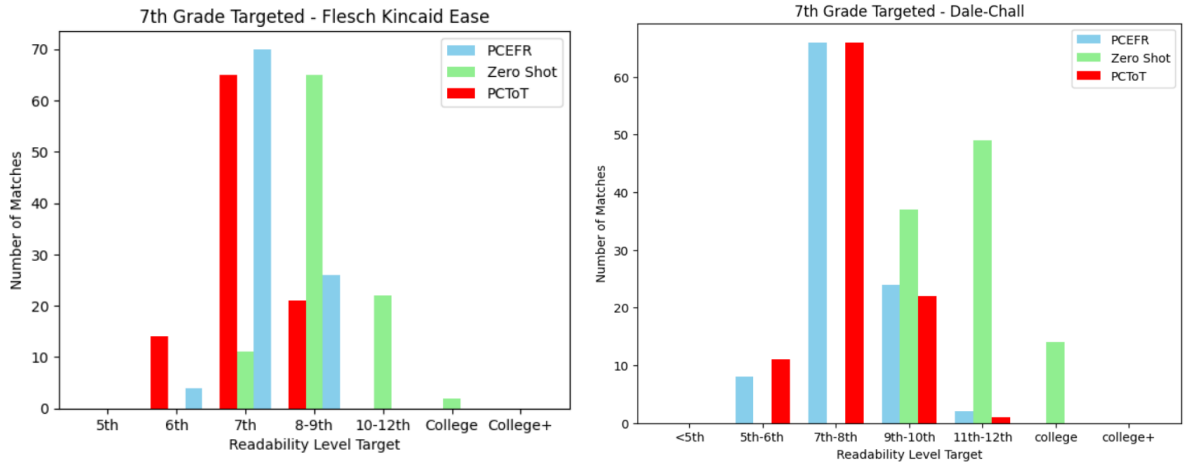
1. *User Interest Not Measured:* The research does not account for variations in user interest, which could influence the effectiveness of the GPT-level targeting.
2. *Limitation to GPT-4o Mini:* The study focuses exclusively on GPT-4o, which may not fully represent the potential outputs or limitations of other language models.
3. *CEFR Level Restrictions:* The study specifically targets CEFR A1 (7th grade) and B1 (10th grade) language levels, which may limit the applicability of the findings to higher or lower proficiency levels.

Discussion

7th Grade

Figure 2

7th Grade Focused Prompt Results



A significant improvement is observed in both readability metrics when targeting the 7th grade specifically. As shown in Figure 2, Zero-Shot only generates a small number of 7th-grade-level texts based on FKEF and none based on DCF. Without any specific context, GPT tends to rely on syllable count as the primary factor for difficulty, which results in most content being classified at 11th-12th grade level in DCF, with some even reaching college-level difficulty. In contrast, the PCEFR results in a wider spread of 7th-grade-level content, demonstrating its effectiveness in targeting the desired readability level. PCToT also has similar performance and has almost no significant differences in data distribution even though it was focused on the plot writing. It indicates that PCToT is suitable to target beginner narratives.

Table 1

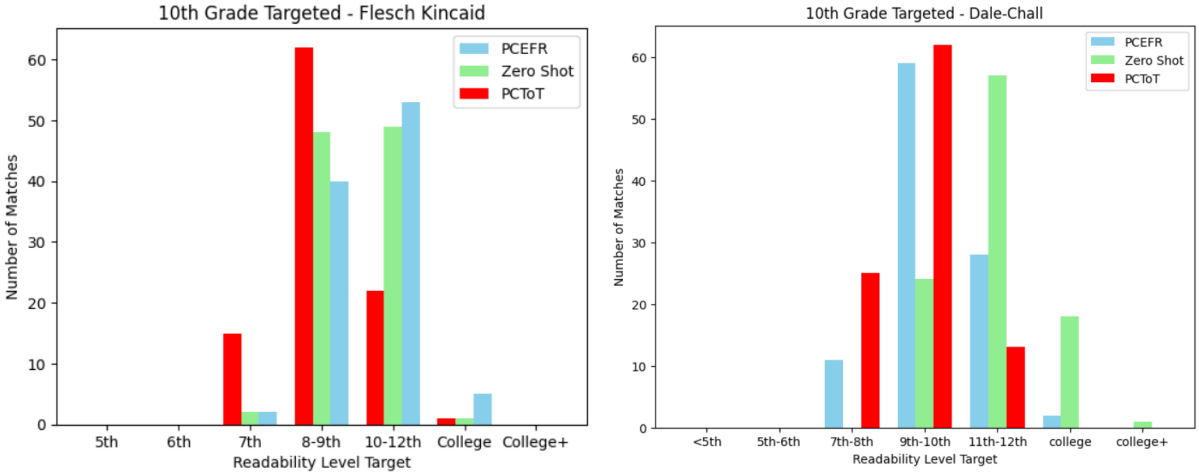
7th Grade Performance Table

Method	FKEF 7th Grade	DCF 7th Grade	Performance Value
Zero Shot	11	0	5.5%
PCEFR	70	66	67.5%
PCToT	65	66	65.5%

Additionally, Table 1 presents the performance of the prompting technique. PCEFR achieves an accuracy of 67.5%, making it the best-performing method for 7th-grade targeting. Zero-Shot, by comparison, performs poorly, with only 5.5% of its outputs at the 7th-grade level. Adding PCToT leads to similar results, with an accuracy of 65.5%.

10th Grade

Figure 3
10th Grade Focused Prompt Results



Unlike the 7th grade, Zero-Shot is able to generate a substantial amount of 10th-grade-level material. This result is expected, as previous research suggests that GPT performs better when generating content for higher education levels. However, similar to the 7th grade, GPT tends to disregard word familiarity, which results in a low DCF score. As shown in the DCF results, most content is generated at the 11th-grade level or higher, with some even reaching college-level difficulty.

PCEFR, however, results in a more balanced distribution of readability scores, with similar spreads in both FKEF and DCF. It indicates that PCEFR did enhance narratives for the intermediate level for the word familiarity. The word complexity has a similar performance with Zero Shot with no significant difference in distribution. As for PCToT, it increases the DCF precision but lowers the FKEF precision. This suggests that PCToT enhances word familiarity more than the word complexity.

Table 2
10th Grade Performance Table

Method	FKEF 10th Grade	DCF 10th Grade	Performance Value
Zero Shot	49	24	36%
PCEFR	53	59	56%
PCToT	22	62	42%

As seen in Table 2, the PCToT achieves the highest performance, with an accuracy of 56%. Zero shot has 36% performance because of the word complexity and PCToT has 42% performance. It shows that PCEFR is performing better than other prompting techniques but PCToT is underperforming on word complexity. PCToT somehow creates much easier word complexity narratives instead.

Conclusion

Using the proposed technique shows a good result if compared with a simple Zero-shot prompt. In this research we find that GPT if we use simple prompting the GPT will disregard word familiarity. Using CEFR as the knowledge and context will help GPT to determine and to target specific levels of learning better achieving higher than 50%. It increased word complexity and word familiarity precision for specific grades. Furthermore in this research we find our prompting method is adaptable with Tree of Thought although it has issues on intermediate level Flesch Kincaid Ease specifically. However the research disregards the plot therefore further research needed in that area.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

This paper acknowledges the use of generative AI technologies during the writing process. OpenAI's GPT-4o mini model was employed solely to enhance the readability and clarity of the Introduction and Related Work sections. The AI was not used to generate original research content or ideas. All AI-assisted outputs were carefully reviewed, edited, and validated by the author to ensure academic integrity and alignment with the paper’s research contributions.

References

Ali, H. H. H. (2022). The Importance of the Four English Language Skills: Reading, Writing, Speaking, and Listening in Teaching Iraqi Learners. *Humanitarian and Natural Sciences Journal, 3*(2). https://doi.org/10.53796/hnsj3210

Chan, C. K. Y., & Tsi, L. H. Y. (2023). The AI Revolution in Education: Will AI Replace or Assist Teachers in Higher Education? *ArXiv*. https://doi.org/10.48550/arXiv.2305.01185

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin, 27*(1), 11–28. http://www.jstor.org/stable/1473169

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2022). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL, 34*(2), 130–146. https://doi.org/10.1017/S095834402100029X

Giray, L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering, 51*, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Hatmanto, E. D., & Sari, M. I. (2023). Aligning Theory and Practice: Leveraging Chat GPT for Effective English Language Teaching and Learning. *E3S Web of Conferences, 440*, 05001. https://doi.org/10.1051/e3sconf/202344005001

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). Better Zero-Shot Reasoning with Role-Play Prompting. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, 1*, 4099–4113. https://doi.org/10.18653/v1/2024.naacl-long.228

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. yung. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: a systematic review of empirical studies. *Smart Learning Environments, 11*(1). https://doi.org/10.1186/s40561-024-00342-5

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

Mikeladze, T. (2023). Creating teaching materials with ChatGPT. IRCEELT – 2023 13th International Research Conference on Education, Language and Literature.

Milton, J., & Alexiou, T. (2020). Vocabulary Size Assessment: Assessing the Vocabulary Needs of Learners in Relation to Their CEFR Goals. In Vocabulary in Curriculum Planning: Needs, Strategies and Tools. https://doi.org/10.1007/978-3-030-48663-1_2

M, I., & G, A. (2024). Factors Affecting in Consciously Improving Vocabulary with a Spaced Repetition System. *International Journal of System of Systems Engineering, 14*(3). https://doi.org/10.1504/ijsse.2024.10055993

Namilae, A., & Leddo, J. (2024). Comparing the Effectiveness of Chat GPT and Teacher-generated Content for Teaching Students. *International Journal of Social Science and Economic Research, 09*(07), 2554–2564. https://doi.org/10.46609/ijsser.2024.v09i07.031

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., & Sra, M. (2021). AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence, 3*, 1013–1022. https://doi.org/10.1038/s42256-021-00417-9

Schmidt, D. C., Spencer-Smith, J., Fu, Q., & White, J. (2024). Towards a Catalog of Prompt Patterns to Enhance the Discipline of Prompt Engineering. *ACM SIGAda Ada Letters, 43*(2). https://doi.org/https://doi.org/10.1145/3672359.3672364

Simmons, A. (2007). *Whoever tells the best story wins: How to use your own stories to communicate with power and impact*. Amacom.

Suzgun, M., & Kalai, A. T. (2024). Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding. *ArXiv*. https://doi.org/10.48550/arXiv.2401.12954

Tamenaoul, H., Hamlaoui, M. El, & Nassar, M. (2024). Prompt Engineering: User Prompt Meta Model for GPT Based Models. In Y. Farhaoui, A. Hussain, T. Saba, H. Taherdoost, A. Verma (Eds.), *Artificial Intelligence, Data Science and Applications*. ICAISE 2023. Lecture Notes in Networks and Systems, 838. Springer, Cham. https://doi.org/10.1007/978-3-031-48573-2_61

Tian, Y., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., May, J., & Peng, N. (2024). Are Large Language Models Capable of Generating Human-Level Narratives? 1. https://doi.org/10.48550/arXiv.2407.13248

Wang, F., Lee, S. (2007). Storytelling is the Bridge. *The International Journal of Foreign Language Teaching, 3*(2), 30–35.

Wang, T. S., & Gordon, A. S. (2023). Playing Story Creation Games with Large Language Models: Experiments with GPT-3.5. In L. Holloway-Attaway, J. T. Murray (Eds.), *Interactive Storytelling (ICIDS 2023)*. Lecture Notes in Computer Science, vol 14384. Springer, Cham. https://doi.org/10.1007/978-3-031-47658-7_28

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 11809–11822).

Young, J. C., & Shishido, M. (2023). Investigating OpenAI's ChatGPT Potentials in Generating Chatbot's Dialogue for English as a Foreign Language Learning. *International Journal of Advanced Computer Science and Applications, 14*(6). https://doi.org/10.14569/IJACSA.2023.0140607

Zaretsky, V. K. (2021). One More Time on the Zone of Proximal Development. *Cultural-Historical Psychology, 17*(2), 37–49. https://doi.org/10.17759/chp.2021170204