Prompting Brilliance: Unlocking ChatGPT's Potential to Revolutionize EFL Dialogue Practices

Julio Christian Young[1] Makoto Shishido[2]
Graduate School of Advanced Science and Technology, Tokyo Denki University
[1]{julio.christian.young}@gmail.com [2]{shishido}@mail.dendai.ac.jp

Abstract

This research addresses a critical gap in the literature by investigating the impact of prompt engineering on generating high-quality learning materials for EFL students using the relatively new ChatGPT, an area that has not been extensively explored in previous research. The study analyzes various prompt techniques and their impact on the quality and characteristics of generated dialogues. The findings reveal that explicitly providing specifications through the prompt yields better results in terms of meeting desired low-level criteria (e.g. word counts or lines per dialogue). However, for a more abstract criterion (proficiency level of generated materials) only a limited percentage of dialogue produced satisfied the target. Despite receiving explicit instructions to generate dialogue suitable for CEFR B2 students, most of the resulting materials meet criteria below the B2 level. The unintentional bias towards easier to understand response in the training process may contribute to this limitation.

1 Introduction

In our daily lives, the ability to convey ideas, express emotions, and engage in meaningful dialogue plays a pivotal role in personal and professional growth. By the end of 2022, ChatGPT, a state-of-the-art language model, captivated the world with its astonishing ability to comprehend human input and engage in human-like conversations. Nevertheless, like other large language models, several studies have suggested that we may have yet to unlock the full potential of ChatGPT due to the usage of suboptimal prompts in our communication with the model [1 - 5]. Like how a skilled leader employs effective communication strategies to inspire and guide their team towards success, the well-thought prompt wields a substantial impact on the quality and outcome of Chat's GPT responses. Hence, researchers argued that understanding the impact of prompts on ChatGPT is crucial for producing desired response that suits the task description.

Numerous research has demonstrated the immense potential of how prompt could enhance the performance of large language models across a wide range of tasks [3 - 9]. On top of that, several studies have gone even further, showcasing how a well-designed prompt can achieve relatively similar, and in some cases, even better performance compared to models that have been fine-tuned for specific tasks. As such, the term "prompt engineering" represents the deliberate manipulation and crafting of prompts to unlock the potential of language models. By exploring and employing prompt engineering techniques, we aim to harness ChatGPT's true capabilities in producing suitable dialogue references for EFL students.

A recent study in [4] formalized several prompt techniques to amplify the performance of large language models, such as ChatGPT. By employing techniques like direct task specification, task specification by demonstration, memetic proxy, constraining behavior, and the concept of meta prompting, the model's performance can be significantly elevated. For example, meta prompting could be particularly valuable in reducing the time investment required from human operators when utilizing the model. In the study, meta prompting is described as instructing ChatGPT to generate a set of prompts for a particular task, rather than including the task details within the initial prompt. In a dialogue generation task, rather than manually creating a list of topics for each dialogue, meta-prompting can be used to define a list of topics before the actual dialogue generation process. As expected, this approach can save time and effort typically required for creatively brainstorming dialogue topics. For the definition of other prompt techniques, readers can refer to the next section.

Although there is a growing body of literature on ChatGPT's potential in language education [10 - 15], the exploration of prompt engineering particularly on its role in producing high-quality learning materials for students remains scarce. To address this gap, our study aims to analyze the effectiveness of several learning prompt techniques. By employing these techniques in a combinatorial manner, we will generate EFL dialogue practice materials tailored to a specific learning scenario. Then, quantitative measurements will be used to assess the quality and appropriateness of the generated materials. By doing so, we aim to uncover valuable insights on how to effectively prompt ChatGPT for generating the best EFL dialogue practice materials.

2 Research Methodology

Learning Context and Dialogue Specifications - First, we will define a specific learning context target to measure how different prompt techniques can enhance the dialogue generation process. Consider a situation where the objective is to generate dialogues for fresh undergraduate students to practice in a classroom setting. A pair of students will follow the dialogue and act out a conversation in front of the class to practice their speaking skills. To ensure that all students could practice in front of their peers, and considering the time constraints of the course, the dialogue should be completed in about 2-3 minutes. Thus, to enhance both clarity and manageability within the given time constraint, we aim for the dialogue to consist of 14 lines, each containing about 10 words. As these students have been learning English as their second language from elementary school, lets also set their proficiency level of approximately CEFR B2. Based on the described context, the following dialogue specifications are set.

1. "The dialogue should strictly involve a conversation between two participants."
2. "The generated dialogue should be suitable for EFL students with CEFR B2 level."
3. "The dialogue should consist of exactly 14 lines."
4. "Each line of dialogue should contain approximately 10 words maximum."

However, our preliminary experimentation revealed that when asked to generate a dialogue, ChatGPT naturally produces dialogues for two participants, even without explicitly mentioning the number of participants. Therefore, we will disregard the first criterion then proceed with other specifications.

Prompt Strategies for Dialogue Generation - To generate dialogues that are suitable for our target audience at the CEFR B2 level, it is crucial to specify appropriate topics for the bot. Rather than relying solely on our own brainstorming process to determine suitable topics, we requested ChatGPT to generate a list of topics. By employing the prompt, "Generate 30 topics that suitable for CEFR B2 students dialogue practice" we obtained a set of topics from ChatGPT.

To initiate the experiment, we will start by utilizing a Base Prompt as shown in Table 2. This prompt aims to simulate a typical request made by teachers to ChatGPT for the generation of dialogue practice materials. We intentionally omitted specific information about the target audience's proficiency level and other dialogue specifications to highlight the significant differences in dialogue results that can occur when such details are not provided. Later, we will compare materials generated using different prompt techniques with Base Prompt resulting materials.

Table 1. Summary of prompt techniques used.

Technique	Prompt Text
Base Prompt	*Hi ChatGPT, please help me to generate a dialogue between A & B. The dialogue topic is {topic_name}. The dialogue will be used for dialogue practice between two students.*
Direct Task Specification	*Do a task with the following details.* *Task: "Generate a dialogue between A & B"* *Topic: "{topic_name}"*
Task Demonstration	*Do a task with the following details.* *Task: "Generate a dialogue between A & B"* *Topic: "{topic_name}"* *Example: "A: Hey, B, have you got a minute? I've got a small favor to ask.* *B: Go on then.* *... (another 11 lines of dialogues)* *A: Great! Thanks, B!"*
Mimetic Proxy	*Do a task with the following details.* *Act as: "EFL teacher"* *Task: "Generate a dialogue between A & B"* *Topic: "{topic_name}"*
Constraining Behaviour	*Do a task with the following details.* *Task: "Generate a dialogue between A & B"* *Topic: "{0}"* *Audience English Level: "CEFR B2"* *Criteria of the line of dialogue: "about 10 words max"* *Total number of lines: "14 lines"*
**Note: {topic_name} will be replaced by prior generated topics	

The first prompt technique that will be evaluated is Direct Task Specification (DTS). Instead of using a polite and conversational prompt like in the base prompt, we will employ a direct specification as shown in Table 2. This prompt aims to minimize ambiguity and enhance the model's understanding of the intended dialogue generation task. By employing DTS, we expect to observe improved dialogue results.

Despite DTS, we also experimented with other techniques as can be seen in Table 2. For such techniques we have adopted a DTS-like format. In the Task Demonstration (TD) prompt, the dialogue included within is specifically designed for a CEFR B2 practice. By providing such an example, we aim to guide ChatGPT in generating responses with similar criteria. As for Mimetic Proxy prompt, the additional information of "Role: EFL teachers" is included to guide ChatGPT in adopting a specific behavior and response style consistent with the designated role. By explicitly specifying the role of EFL teachers, we aim to encourage ChatGPT to generate dialogues that reflect the language, tone, and instructional approach typically employed by EFL teachers. In the case of the Constraining Behavior prompt, we take a more direct approach by explicitly defining all the specifications and guidelines that ChatGPT needs to adhere to during the dialogue generation process.

Additionally, besides evaluating each prompt technique individually, we have designed two additional prompts. The first approach combines all techniques, while the second approach omits task demonstration (TD) from the combination. The second prompt intentionally excluded the TD technique, based on the assumption that the inclusion of TD may introduce confusion to the model as produce a lengthier input. For clarity, a summary of the experiment setup is depicted in Figure 1.

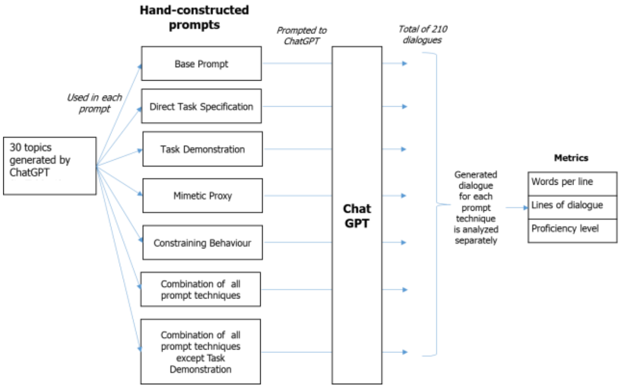
Hand-constructed prompts
Prompted to ChatGPT
Total of 210 dialogues
Used in each prompt
30 topics generated by ChatGPT
Base Prompt
Direct Task Specification
Task Demonstration
Mimetic Proxy
Constraining Behaviour
Combination of all prompt techniques
Combination of all prompt techniques except Task Demonstration
Chat GPT
Generated dialogue for each prompt technique is analyzed separately
Metrics
Words per line
Lines of dialogue
Proficiency level

Figure 1. Summary of The Experiment Setup

Assessing Dialogue Suitability for CEFR Levels - When evaluating the appropriateness of generated dialogues for a specific CEFR (Common European Framework of Reference for Languages) level, there is no direct implementation that can precisely measure the CEFR level based on the input text. Therefore, we employ a several readability metrics to assess the suitability of the generated dialogues. These metrics include Flesch Reading Ease, SMOG, Coleman Liau, Automated Readability Index, Dale Chall, Linsear Write, and Gunning Fog. By subjecting the dialogues to these metrics, we generate scores that capture various aspects of text complexity. Subsequently, a consensus-based approach is employed to ascertain the optimal target audience for each dialogue. Then, we map the target audience level for its CEFR level equivalence. Specifically, we consider school grades below the sixth grade, junior high school, and senior high school and higher education as equivalent to CEFR A2. B1, and B2 respectively.

3 Results

The following table presents the experimentation results, showcasing the performance of different prompt techniques for generating EFL dialogue. In the table, we will use the following abbreviations to represent the performance results for different prompt techniques.

1. **BP**: Base Prompt
2. **DS**: Direct Task Specification
3. **TD**: Task Demonstration
4. **MP**: Mimetic Proxy
5. **CB**: Constraining Behavior
6. **C1**: All Prompt Combinations
7. **C2**: All combinations except Task Demonstration

Table 2. Experiment Results

	BP	**DS**	**TD**	**MP**	**CB**	**C1**	**C2**
Words/ Line							
- Average	19.2	20.2	16.8	19.4	**9.4**	**9.8**	**9.8**
- Median	18	19	16	17	**9**	**9**	**9**
- Below 10 (%)	17.1	15.6	18.5	17.5	**69.3**	62.9	63.5
Lines of Dialogue							
-Exactly 14 (Count)	6	1	4	4	6	7	**10**
Exactly 14 (%)	20	3	13.3	13.3	20	23.3	**33.3**
Proficiency Level							
School Grade (Average)	6.0	**6.4**	5.9	6.1	5.8	5.8	5.5
A2 (Count)	16	17	18	16	**20**	18	18

B1 (Count)	**14**	10	10	12	8	10	11
B2 (Count)	0	**3**	2	2	2	2	1

Table 4 presents various quantitative metrics according to the dialogue specifications for different prompt techniques. The results from CB prompt indicate ChatGPT's understanding of the maximum words per line specification. In contrast, in other cases where the specification is not explicitly mentioned, the average number of words per line is significantly higher. However, it is important to note that at best (The CB case), only 69.3% of the dialogues generated met this specified criterion. Furthermore, as additional specifications are introduced through different techniques (A1 and A2), we observe that ChatGPT starts producing more dialogue with lines of dialogue more than 10 words. This suggests that ChatGPT may start to compromise on the maximum words per line criteria to fulfill other specified criteria.

Meanwhile, the results for words per line in the TD prompt indicate that ChatGPT struggles to learn the implicit criteria provided in the example dialogue. Despite each line of the example dialogue having a word count below 10, the resulting dialogues from the TD prompt exhibit an average words per line that is significantly higher (with a median of 16 and only 18.54% of lines containing fewer than 10 words). These findings highlight how ChatGPT struggles to fully grasp and replicate the implicit criteria from an example dialogue.

Similarly, the higher percentage of dialogues that met 14 lines of dialogue criteria in CB, C1 and C2 prompts demonstrate ChatGPT understanding towards explicit dialogue specifications. Interestingly, while the CB prompt yielded the best dialogue results in terms of word per line criterion, more dialogues from A2 prompt met number of lines of dialogue criterion (**33.3%**). This result might suggest that ChatGPT might prioritize the specified criteria in the prompt differently when extra context provided (role as EFL teacher in A2 prompt). Nonetheless, even the most effective prompt employed in the experiment only led to 10 dialogues that met the specific line count. This implies that ChatGPT may better comprehend lower-level criteria better than the higher one.

Moreover, by looking at distribution of the appropriateness of proficiency level from the generated materials across all prompts, none of the prompt techniques used were able to generate dialogues suitable for the intended proficiency level. The dialogues produced by the prompt technique that achieved the best results, DS, only reached an intended school grade of 6.43, which aligns with sixth grade to first junior high school students (A2 - B1 level). Other prompt techniques resulted in dialogues that were even further below the intended proficiency level. Despite any techniques used, ChatGPT failed to learn and replicate the implicit requirements. Out of the resulting dialogues, only about one or two dialogues were identified as potentially suitable for CEFR B2. This number is even smaller than DS prompt that doesn't contain any information regarding the intended target audience. Therefore, we concluded that ChatGPT couldn't understand the concept of proficiency level.

4 Conclusion

In this research, we explored different prompt techniques to manipulate responses from ChatGPT. The findings indicate that using prompt techniques can influence the quality of the generated dialogues but only for a lower-level criteria. When the maximum words per line specification was explicitly provided, ChatGPT could produce more materials that met required criterion. However, it is important to note that not all generated dialogue lines adhered to this criterion (with at max only 69.35% compliance). Moreover, the TD prompt, which aimed implicitly instruct the criteria from an example dialogue, posed challenges for ChatGPT. The resulting dialogues exhibited a higher average words per line count, indicating the struggle to grasp and replicate the implicit criteria accurately. Nonetheless, when compared to the results from a regular prompt, the TD prompt performed slightly better.

Meanwhile, for higher-level criteria, despite the prompt techniques used, none could generate a satisfying result. ChatGPT failed to produce dialogues at the intended proficiency even when the explicit instruction was given. Such limitations could be attributed to the training process of ChatGPT. As ChatGPT is trained to facilitate conversations with a diverse audience, it may unintentionally receive positive reinforcement for providing more straightforward, easily comprehensible responses. Consequently, this unintentional inclination towards simplicity could limit ChatGPT's ability to offer more advanced materials to higher proficiency students.

References

[1] K. Busch, A. Rochlitzer, D. Sola, and H. Leopold, “Just Tell Me: Prompt Engineering in Business Process Management,” in Enterprise, Business-Process and Information Systems Modeling, 2023, pp. 3–11.

[2] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” ACM Comput. Surv., vol. 55, no. 9, Jan. 2023, doi: 10.1145/3560815.

[3] P. Denny, V. Kumar, and N. Giacaman, “Conversing with Copilot: Exploring Prompt Engineering for Solving CS1 Problems Using Natural Language,” in Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1, 2023, pp. 1136–1142. doi: 10.1145/3545945.3569823.

[4] L. Reynolds and K. McDonell, “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,” in Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021. doi: 10.1145/3411763.3451760.

[5] C. E. Short and J. C. Short, “The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation,” J. Bus. Ventur. Insights, vol. 19, p. e00388, 2023, doi: https://doi.org/10.1016/j.jbvi.2023.e00388.

[6] J. Seo et al., “Plain Template Insertion: Korean-Prompt-Based Engineering for Few-Shot Learners,” IEEE Access, vol. 10, pp. 107587–107597, 2022, doi: 10.1109/ACCESS.2022.3213027.

[7] C. W. F. Mayer, S. Ludwig, and S. Brandt, “Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models,” J. Res. Technol. Educ., vol. 55, no. 1, pp. 125–141, 2023, doi: 10.1080/15391523.2022.2142872.

[8] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, “Calibrate Before Use: Improving Few-shot Performance of Language Models,” in Proceedings of the 38th International Conference on Machine Learning, 2021, vol. 139, pp. 12697–12706.

[9] A. D. White et al., “Assessment of chemistry knowledge in large language models that generate code,” Digit. Discov., vol. 2, no. 2, pp. 368–376, 2023.

[10] J. K. M. Ali, M. A. A. Shamsan, T. A. Hezam, and A. A. Q. Mohammed, “Impact of ChatGPT on Learning Motivation: Teachers and Students’ Voices,” J. English Stud. Arab. Felix, vol. 2, no. 1, pp. 41–49, 2023, doi: 10.56540/jesaf.v2i1.51.

[11] M. S. S. Moqbel and A. M. T. Al-Kadi, “Foreign language learning assessment in the age of ChatGPT: A theoretical account,” *J. English Stud. Arab. Felix*, vol. 2, no. 1, pp. 71–84, 2023.

[12] T. Adiguzel, M. H. Kaya, and F. K. Cansu, “Revolutionizing education with AI: Exploring the transformative potential of ChatGPT,” *Contemp. Educ. Technol.*, vol. 15, no. 3, p. ep429, 2023.

[13] J. Qadir, “Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education,” in *2023 IEEE Global Engineering Education Conference (EDUCON)*, 2023, pp. 1–9. doi: 10.1109/EDUCON54358.2023.10125121.

[14] F. R. Baskara and F. X. Mukarto, “Exploring the Implications of ChatGPT for Language Learning in Higher Education,” *IJELTAL (Indonesian J. English Lang. Teach. Appl. Linguist.*, vol. 7, no. 2, pp. 343–358, 2023.

[15] W. C. H. Hong, “The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research,” *J. Educ. Technol. Innov.*, vol. 5, no. 1, 2023.