Investigating OpenAI’s ChatGPT Potentials in Generating Chatbot's Dialogue for English as a Foreign Language Learning

Julio Christian Young, Makoto Shishido
Department of Information, Communication, and Media Design Engineering, Tokyo Denki University, Tokyo, Japan

***Abstract*—Lack of opportunities is a significant hurdle for English as a Foreign Language (EFL) for students during their learning journey. Previous studies have explored the use of chatbots as learning partners to address this issue. However, the success of chatbot implementation depends on the quality of the reference dialogue content, yet research focusing on this subject is still limited. Typically, human experts are involved in creating suitable dialogue materials for students to ensure the quality of such content. Research attempting to utilize artificial intelligence (AI) technologies for generating dialogue practice materials is relatively limited, given the constraints of existing AI systems that may produce incoherent output. This research investigates the potential of leveraging OpenAI's ChatGPT, an AI system known for producing coherent output, to generate reference dialogues for an EFL chatbot system. The study aims to assess the effectiveness of ChatGPT in generating high-quality dialogue materials suitable for EFL students. By employing multiple readability metrics, we analyze the suitability of ChatGPT-generated dialogue materials and determine the target audience that can benefit the most. Our findings indicate that ChatGPT's dialogues are well-suited for students at the Common European Framework of Reference for Languages (CEFR) level A2 (elementary level). These dialogues are easily comprehensible, enabling students at this level to grasp most of the vocabulary used. Furthermore, a substantial portion of the dialogues intended for CEFR B1 (intermediate level) provides ample stimulation for learning new words. The integration of AI-powered chatbots in EFL education shows promise in overcoming limitations and providing valuable learning resources to students.**

***Keywords*—ChatGPT; chatbots as learning partners; EFL chatbot system; dialogue creation**

I. Introduction

English has become the most widely spoken language globally, with approximately 1.5 billion people speaking it as a first, second, or foreign language [1]. As a result, English proficiency is increasingly becoming essential for success in various fields such as academics, business, and international relations. Undoubtedly, English language learning can significantly benefit foreign language students. From an academic standpoint, English is essential in various fields, including science, technology, engineering, and mathematics, where English is the primary language of instruction and research. Moreover, by mastering English, students can increase their chances of success in their later global careers, as it is commonly used in international business.

Despite the benefits, learning a new language can be challenging, and for foreign language students, learning English can be particularly difficult due to several factors [2]–[4]. One significant challenge foreign language students face in learning English is the lack of opportunity to practice speaking the language [4]–[7]. Many previous studies have shown that speaking a language is essential for effective communication and language acquisition [2], [8], [9]. Thus, the lack of speaking practice can lead to a significant barrier in language acquisition, as speaking is essential to build fluency and confidence in using the language. When left alone, this situation can lead to demotivation and further decrease opportunities to practice.

To deal with this situation, using chatbots in language learning has gained increasing research attention in recent years [10]–[15]. Several studies have found that chatbots can be an effective tool for helping EFL students learn the language. One of the main benefits of practicing with a chatbot is that students can gain conversation experience in a safe, low-pressure environment [11]. Furthermore, through the recent advancement of speech recognition and synthetic speech technologies, chatbots can be implemented to simulate real-life conversations [10], [15]. Other than that, a previous study also showed that students often feel less judged when they receive feedback or corrections from the chatbot [10], [11]. Furthermore, chatbots can enable students to practice anywhere and at any time outside the classroom, thus increasing their language exposure [10], [11], [15], [16]. Such flexibility can help them to overcome the challenge of limited opportunities to practice, particularly for students who may not have access to native speakers.

A successful chatbot system for language learning *typically* involves several key components, including a speech recognition (SR) module, audio content, and reference dialogue content. In previous studies, we have covered subjects related to a speech recognition module and audio content for a chatbot system for helping EFL students to learn English [10], [11]. One study evaluated the use of Vosk, an internet-free speech recognition module, and found that limiting the vocabularies recognized by the SR module during runtime improved the system's ability to recognize student speech input, resulting in a more pleasant learning experience [11]. In another study, WaveNet, a deep learning-based speech synthesizer, was evaluated for generating audio content in an EFL chatbot system [10]. While students could distinguish that the content produced by WaveNet sounded less natural than actual human speech, they produced fewer errors when transcribing the WaveNet-generated audio, indicating that it was easier to understand.

While numerous previous studies have extensively explored speech recognition and technology for developing audio content, the chatbot's dialogue content is often still sourced from existing materials produced by humans. With the advancements in artificial intelligence and deep learning technology, it is now possible for machines to generate readable and contextually appropriate content. Using machine-generated content could reduce reliance on human-produced content in the development process, thus reducing the cost and time needed significantly. Therefore, this research aims to evaluate the potential of using artificial intelligence (AI)-generated content for reference dialogue in an EFL chatbot system.

The evaluation of the potential of machine-generated content for an EFL chatbot system will focus on a chatbot implementation by OpenAI, ChatGPT [19]. ChatGPT is a novel chatbot implementation with impressive abilities to return coherent and contextually appropriate responses based on user requests. By leveraging the vast amounts of text data, ChatGPT can generate text in various styles and tones, making it a promising option for generating content suitable for numerous purposes [19]. For EFL content generation, OpenAI has excellent potential to generate English content that could be useful for EFL students.

Therefore, this study aims to evaluate the appropriateness of ChatGPT-produced materials for dialogue practice in language learning. As ChatGPT is a relatively new technology that hasn't been extensively explored in this context, investigating its capabilities becomes necessary. We will utilize ChatGPT to produce a series of dialogue practice materials and employ multiple readability metrics to thoroughly analyze their suitability. By gaining insights into the characteristics of ChatGPT-generated dialogue, we can identify the most appropriate audience to maximize learning benefits. Determining the target audience that can derive the most from these materials will allow us to optimize their use and enhance the effectiveness of language learning experiences.

II. Literature Review

A. Voice-enabled Chatbot for EFL Learning

There are two types of chatbots: text-based and voice-enabled chatbots classified based on their modality. Voice-enabled chatbots have been proposed as a helpful tool for learning and practicing a second language (L2) speaking skills. A study in [12] discovered that L2 learners appreciated the chatbot's capacity to expose them to various conversational expressions and vocabulary and enable repetitive practice. On top of that, L2 learners prefer chatbots over human partners due to their fear of making mistakes and appearing incompetent during interactions with human partners [13].

A recent study by Han [14] showed how Alexa, a general voice-enabled chatbot, could help students by engaging them in conversations to practice their speaking skills. The experimentation indicated that chatbot-assisted learning improved students' pronunciation and language fluency. Moreover, post-questionnaire responses showed that the integration of such chatbot positively impacted students' interest in learning and enhanced their motivation to learn. Similarly, using readily-available chatbots such as Google Assistants [15], [16], and Alexa [17], [18] also led to positive improvements in students' language proficiency. Researchers noted that students felt less embarrassed and anxious when practicing with a chatbot [15], [16]. Furthermore, chatbots promote self-directed foreign language learning outside school settings where native speakers are hard to find [17], [18].

Although general chatbots may seem appealing, several studies have suggested that such system adaptation may be less effective for L2 learners as it may not cater to their specific needs [21], [22]. Therefore, several criteria should be considered when designing a chatbot for language learning, such as language learning potential, learners' suitability, and authenticity [21]. The language learning potential criterion can be further broken down into components like interactional modification and task focus. Secondly, the learners' suitability criterion should consider various factors, such as language proficiency, age, learning style, and individual characteristics. Lastly, the authenticity criterion indicates that the materials presented within the chatbot should imitate real-life tasks that learners are likely to encounter.

A previous study [21] that implemented a task-oriented chatbot for helping students in their learning journey yielded promising results. The chatbot could maintain lengthy English conversations and engage in L2 problem-solving tasks with participants. Researchers noted that this type of speaking experience is hard to provide in regular EFL classes due to class size and time constraints within the curriculum. Similarly, an evaluation of a specifically designed EFL chatbot in [23] demonstrated the significant potential of its adaptation. The study found that the chatbot matched students' learning styles and enabled them to learn ubiquitously, thus making them enjoy their learning experience. Regardless, the study's pre-test and post-test settings revealed no significant improvement in students' Oral Proficiency Interview – Computer (OPIc) scores after the system adoption. The mixed findings in chatbot research indicate a need for further investigation.

B. Readability Metrics for English Materials

Readability metrics are tools used to measure how easily readers can understand a written text. They are commonly used to evaluate the appropriateness of text materials by determining the complexity of the language used within the presenting material based on specific criteria. Flesch Reading Ease [24], Dale-Chall [25], and McAlpine EFLAW [26] are several commonly used readability metrics for English text materials. These metrics consider factors such as syllable counts, mini words counts, or a dictionary of difficult words to calculate a score that reflects the text's difficulty level.

Previous studies showed that these metrics could help assess the appropriateness of text materials for EFL learning [27], [28]. By utilizing these readability metrics, we can evaluate the language complexity of the chatbot's responses and ensure they are appropriate for the target audience. For instance, if the chatbot generates too complex responses for low-level EFL learners, it may stunt their ability to comprehend and learn using the material. However, high-level EFL learners may find the material less challenging and unstimulating if the responses are simple enough. By assessing the readability of chatbot-generated material, researchers and developers can ensure that the language complexity is appropriate for the targeted EFL learner group, thus enhancing the learning experience.

For example, a study in [28] showed how Flesch-Kincaid readability metrics could be used to analyze the difficulty level of English textbooks for Chilean EFL high school students. The study illustrated how readability metrics could be used to recommend adjustments to English teaching materials according to students' level of comprehension. Another study by Gao et al. in [27] also showed the potentiality of several readability metrics as features to predict chatbots' popularity. The study found that very popular and unpopular chatbots have significant readability scores, thus indicating that readability metrics can be a valuable indicator to reflect users' interest in chatbot adoption.

This research will use three different readability metrics to assess the appropriateness of chatbot-generated material for EFL learning. This combination was chosen because each metric employs a different strategy to calculate the readability score. For instance, Flesch Reading Ease considers the syllable count when calculating the material's readability. Differently, Dale-Chall utilizes difficult words not commonly used in everyday language for calculating the difficulty. On the other hand, McAlpine EFLAW computes the readability score by using mini-words in a given text

1) Flesch reading ease: is a tool used to assess the readability of a given text in English. It works based on a formula proposed by Rudolf Flesch in [24]. The Flesch Reading Ease score is between 0 and 100, indicating how easy or difficult it is to understand a text. A higher score indicates that the text is easier to read, while a lower score indicates that the text is more challenging to read. The definition of the Flesch Reading Ease score is given in Formula 1,

$$206.835 - 1.015 \times \left(\frac{W}{S}\right) - 84.6 \times \left(\frac{s}{W}\right) \tag{1}$$

where s, W, and S represent the number of syllables, words, and sentences in the given text, respectively. Then, the interpretation of the score is shown in Table I.

2) Dale-Chall readability formula: this is another readability formula first developed by Edgar Dale and Jeanne Chall in the 1940s adjusted further in 1995 [25]. The formula calculates a text's readability by considering its number of difficult words. The formula defines a difficult word as any word that is not in a list of common words familiar to most fourth-grade students. The formula generates a score that ranges from 0 to 10. A score of 0 indicates that the text is effortless to read, while a score of 10 indicates that the text is challenging to read. The formula for calculating the raw score of the Dale–Chall readability score is given by Formula 2.

TABLE I. FLESCH READING EASE SCORE INTERPRETATION

Score	US School Level	Description
90.00-100.00	5th grade	Very easy to read. Easily understood by average 11 years old students.
80.00-90.00	6th grade	Easy to read. Conversational English for consumers.
70.00-80.00	7th grade	Fairly easy to read.
60.00-70.00	8th and 9th grade	Plain English. Easily understood by 13 to 15 years old students.
50.00-60.00	10th – 12th grade	Fairly difficult to read.
30.00-50.00	College	Difficult to read.
10.00-30.00	College Graduate	Very difficult to read. Best understood by university graduates
0.00-10.00	Professional	Extremely difficult to read. Best understood by university graduates.

$$0.1579 \times \left(\frac{DW}{W} \times 100\right) - 0.0496 \times \left(\frac{W}{S}\right) \tag{2}$$

where W, DW, and S represent the number of words, difficult words in the given text, respectively. The interpretation of the Dale-Chall readability score is given in Table II.

TABLE II. DALE-CHALL SCORE INTERPRETATION

Score (x)	Description
$x < 5.0$	Easily understood by an average 4th grade student or lower.
$5.0 \le x < 6.0$	Easily understood by an average 5th or 6th grade student.
$6.0 \le x < 7.0$	Easily understood by an average 7th or 8th grade student.
$7.0 \le x < 8.0$	Easily understood by an average 9th or 10th grade student.
$8.0 \le x < 9.0$	Easily understood by an average 11th or 12th grade student.
$9.0 \le x$	Easily understood by an average college student

The Dale-Chall readability formula was revised in 1995 to improve its accuracy and reliability. The revision included a new list of 3,000 familiar words compiled based on surveys of fourth-grade students. This new list replaced 769 words on the previous one, which had become outdated over time [25]. This research will use the new version of the Dale-Chall readability formula.

3) McAlpine EFLAW readability formula: is specifically developed to measure the readability of English language materials for non-native speakers of English. It regards mini words as a linguistic feature that can make English texts difficult for non-native speakers to read. Mini words are common words of one, two, or three letters. In the previous study [25], the researcher argued that a cluster of mini words in wordy cliches, colloquial expressions, and phrasal verbs confuse international readers. The McAlpine EFLAW readability score calculation is given by Formula 3.

$$EFLAW\ Score = \frac{W+M}{S} \tag{3}$$

where W, M, and S represent the number of words, mini-words, and sentences in the given text. The interpretation from the McAlpine EFLAW score can be seen in Table III.

TABLE III. MCALPINE EFLAW SCORE INTERPRETATION

Score (x)	Description
$x \leq 20.49$	very easy to understand
$20.49 < x \leq 25.49$	quite easy to understand
$25.49 < x \leq 29.49$	a little difficult
$29.49 < x$	very confusing

C. Generative Pre-training Transformers, InstructGPT, and ChatGPT

Generative Pre-trained Transformers (GPT) have emerged as a significant advancement in natural language processing (NLP) and have gained immense popularity in recent years [30], [35]. GPT, developed by OpenAI, is a deep learning model based on the Transformer architecture. It is designed to generate coherent and contextually relevant text given a prompt. The model employs a self-attention mechanism, allowing it to capture dependencies between words efficiently [30]. GPT achieved state-of-the-art performance on a wide range of language tasks due to its ability to learn from large amounts of text data. The original GPT model was trained on a dataset containing 40GB of text data from the internet [30]. As of today, OpenAI's ChatGPT is based on the GPT-3.5 model. While there is no publicly available information about the exact amount of data used for training GPT-3.5 specifically, it is worth noting that GPT-3, on which GPT-3.5 is built, was trained on a substantial corpus of text data. GPT-3's training dataset comprised approximately 570GB of text sourced from various types of content, including books, websites, and articles [31]. This extensive and diverse dataset facilitated GPT-3's ability to grasp language patterns and acquire a broad understanding of knowledge and context.

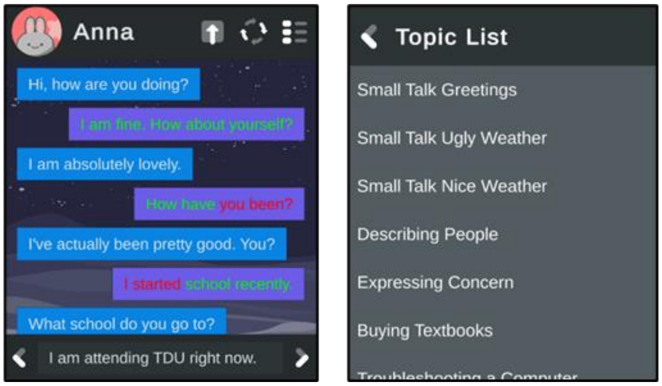
Like any other transformers-based large language model (LLM), GPT training divides into pre-training and fine-tuning stages [31]. A language model is trained on a large corpus of publicly available text data during pre-training. The model learns to predict the next word in a sentence, acquiring a broad understanding of grammar, context, and world knowledge. After pre-training, the model is fine-tuned on specific downstream tasks using supervised learning. The fine-tuning process involves training the model on a narrower dataset with labeled examples. This step allows the model to specialize in a specific task such as language translation, sentiment analysis, or question answering. There is no publicly detailed information available about how ChatGPT was trained. However, the documentation of ChatGPT mentioned that it was using a pre-trained by using a larger LLM than GPT-3 on a more significant amount of data. Then, the model was fine-tuned further to generate detailed responses based on given instructions or demonstrations (InstructGPT), using Reinforcement Learning from Human Feedback (RLHF) [32], [33].

RLHF is a technique used to improve the performance of language models through iterative fine-tuning using human-generated feedback [33]. RLHF involves collecting comparison data where different model responses are ranked by quality. These rankings are used to create a reward model, which guides the model's training using reinforcement learning algorithms such as Proximal Policy Optimization (PPO) [34]. RLHF has been used to refine GPT models, enhancing their output quality and reducing biases—the series of human-in-the-loop iterations allowing the model to generate more coherent responses.

III. RESEARCH METHODOLOGY

A. Tools and Materials

As previously mentioned, this research evaluated the appropriateness of artificial intelligence (AI)-generated dialogue for EFL students using several readability criteria. The generated dialogues are intended as reference dialogue in the mobile application to help students practice their speaking skills. Students can choose a topic using the app and practice their English skills, as shown in Fig. 1.



Hi, how are you doing?

I am fine. How about yourself?

I am absolutely lovely.

How have you been?

I've actually been pretty good. You?

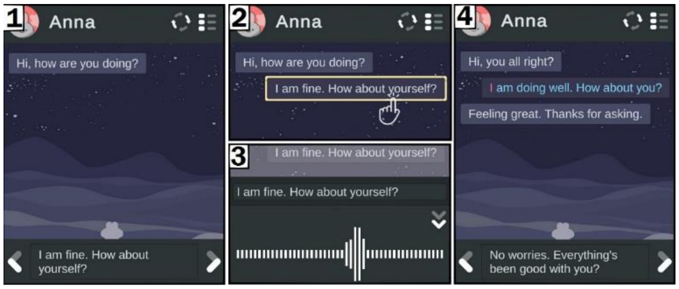
I started school recently.

What school do you go to?

I am attending TDU right now.

Fig. 1. Voice-enabled chatbot mobile application.

The app provided a range of topics that students could use to practice their listening, reading, and speaking simultaneously. After selecting a topic, the app will load the reference dialogue on the selected topic. The app will always start the conversation using TTS technology by converting the first line into the reference dialogue. Then, to reply to the conversation, the student can choose one of three text options in the reference dialogue. Based on their choice, the app will engage them in a read-a-loud activity and evaluate their pronunciation using SR technology. By comparing the student's text choice and the SR transcription result, the app will re-render some text parts in red if they are not present in the transcription; otherwise, they will be in green. The conversation between a student and the bot will continue if there is still a line of dialogue in the reference dialogue. Fig. 2 depicts the interaction between the student and the app.



1 Hi, how are you doing?

I am fine. How about yourself?

2 Hi, how are you doing?

I am fine. How about yourself?

3 I am fine. How about yourself?

I am fine. How about yourself?

4 Hi, you all right?

I am doing well. How about you?

Feeling great. Thanks for asking.

No worries. Everything's been good with you?

Fig. 2. Students' interactions with the application.

A successful learning outcome in a dialogue practice can only be achieved when students are willing to keep practicing the dialogue repeatedly. Students can learn more about what they are trying to understand with each repetition. Moreover, when stimuli are learned by repetition, they are remembered better and retained for a longer time. Dialogue variability is an essential factor affecting students' motivation to keep practicing. If the reference dialogues need to be more varied, students may feel bored having to practice using them repeatedly. The dialogue's difficulty level is another crucial factor affecting the learning process's success. Dialogues that are too difficult will undoubtedly lower the confidence level of the students and decrease their motivation to learn. Teachers generally spend much time and effort creating teaching materials that fulfill those two criteria.

Therefore, this research evaluates the possibility of using AI-generated materials as a reference dialogue within the app. The reference dialogues were generated by using OpenAI's ChatGPT. The dialogues were produced by inputting the following prompt to the bot: "Please help me to make a dialogue to help EFL students to practice their English. The dialogue is between A and B. A is an undergraduate student at ABC University. B is an exchange student from Italy. The topic is {{topic name}}", where {{topic name}} was selected from Table IV.

TABLE IV. LIST OF TOPICS GENERATED BY CHATGPT

Topic (1st – 5th)	Topic (6th – 10th)	Topic (11th – 15th)
Greet new exchange student	Fermented foods	Learn programming
Lunch Invitation	Sumo wrestling	Summer's vacation
Play arcade on weekends	Coin Laundry	Traveling to Kyoto
Foods and hobbies	Favorite snacks	Buying new clothes
Learn to use chopsticks	Sightseeing in Tokyo	Last week in Tokyo

In the prompt above, the lines "The dialogue is between A and B. A is an undergraduate student at Tokyo Denki University. B is an exchange student from Italy" are intended to give context to the AI so it could create a livelier dialogue related to students. Furthermore, a series of topics in Table IV means to test whether the ChatGPT can produce various topics for students to practice. On top of that, we asked ChatGPT to give two or three alternative lines of dialogue for each line in the produced dialogue. Later, in the experimentation, using a single dialogue from ChatGPT, 30 unique combinations will be generated. Therefore, 450 unique sample combinations of dialogues will be analyzed.

B. Metrics and Measurements

Based on each ChatGPT-generated dialogue, an analysis process was first carried out using three readability metrics: Flesch Reading Ease, McAlpine EFLAW, and Dale-Chall readability metrics. The Flesch Reading Ease metric is intended to measure the difficulty level of a dialogue by considering the ratio of polysyllables in all words in the dialogue. The more polysyllables there are, the more complex the dialogue is assumed to be according to this metric. On the other hand, the McAlpine EFLAW metric is used to consider mini-words in dialogue. The more mini-words used, the metric assumes more complex it is. Lastly, the Dale-Chall metric considers a list of difficult words compiled from previous studies.

The usage of these three metrics aims to cover the flaws of each metric with the other two metrics. Since the Flesch Reading Ease metric only considers the number of polysyllables in its calculation process, sentences with multiple mini-words will be considered easy to understand. Therefore, the McAlpine EFLAW readability score calculation process is carried out to complement the weakness of the metric. Then, the Dale-Chall metric is also used to determine the difficulty level of the text based on words that have few polysyllables but are challenging to understand, such as "abide," "deem," and "quail".

Based on the scores generated by each metric, a process of interpreting the difficulty level of the dialogue is carried out. The interpretation will be made by first visualizing the distribution of the difficulty level of the generated dialogue. From the visualization results, an analysis is carried out to determine the complexity of the ChatGPT-generated dialogues.

IV. RESULTS

Based on 450 dialogue samples that aim to simulate conversations between the bot and the students in the application, the Flesch Reading Ease score for each sample was first calculated. Then, through the resulting scores, a visualization was carried out to show the scores' central tendency and distribution from all samples. Fig. 3 shows the distribution of scores from all samples.

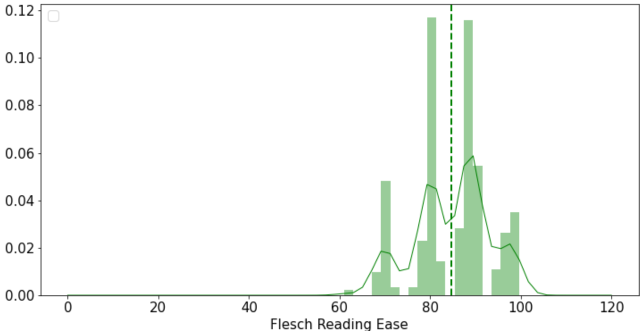


Fig. 3. Flesch reading ease scores' distribution from all samples.

Fig. 3 shows that the sample dialogues have a score distribution ranging from 60 to 100, with most samples having scores in the range of 80-90. Hence, it can be concluded that most of the simulated dialogues are easily comprehensible to sixth-grade elementary school students. Additionally, since there is a small yet significant portion of samples with scores between 60 and 80, they can also serve as sufficient stimuli for junior high school students. However, the generated materials may not be suitable for senior high school students or students in higher education, as they could easily comprehend such materials, thus not providing an appropriate level of challenge for their learning. This interpretation can be further extended for EFL students by referring to the previous study [29]. Since most samples have scores between 80 and 90, students with a Common European Framework of Reference for Languages (CEFR) level of A2 (elementary level) will benefit the most when using the materials. While the materials could still be suitable for students with CEFR levels A1 (beginner) and B1 (intermediate), students with levels B2 (upper intermediate) to C2 (advanced) may find the materials less challenging and too simple.

Next, the Dale-Chall readability scores were calculated for all sample dialogues. Fig. 4 illustrates the distribution of the resulting scores across all samples.

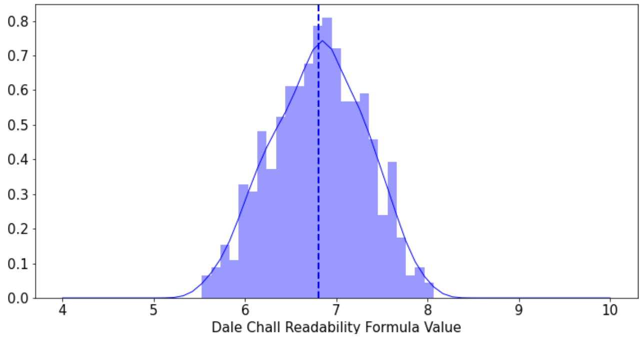


Fig. 4. Dale-Chall readability formula values' distribution for entire samples.

Similarly to the previous interpretation, based on the distribution of resulting scores shown in Fig. 4, it can be argued that the generated materials are most suitable for sixth-grade elementary school students or students in the early years of junior high school (CEFR A2 and B1). Moreover, the absence of samples with Dale-Chall scores above 8.0 confirms that the generated materials are unsuitable for students with CEFR levels B2 to C2. Finally, the McAlpine EFLAW score was calculated for each simulated conversation. The visualization of the score distribution can be observed in Fig. 5. Referring to the resulting scores in Fig. 5, as none of them have a score below 20, it can be interpreted that the resulting materials do not extensively utilize mini-words that could confuse EFL students when consuming them.

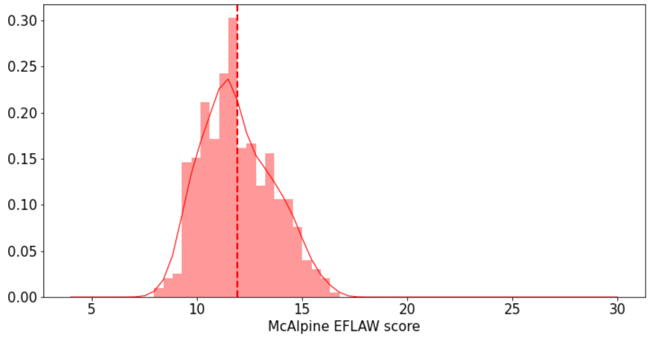


Fig. 5. McAlpine EFLAW scores' distribution for all samples.

V. Discussions

Based on the experimentation results, several conclusions can be drawn regarding the suitability of ChatGPT-generated materials as EFL chatbot reference dialogues. Firstly, the minimal McAlpine EFLAW score observed in all simulated conversations suggests that the dialogues generated by ChatGPT do not contain excessive use of mini-words. This indicates that wordy clichés, colloquial expressions, and phrasal verbs, which could potentially confuse international readers, were avoided in the resulting dialogue [19]. The consistently low McAlpine EFLAW scores across all simulated dialogues indicate that EFL students can easily comprehend and understand the content. These findings provide confidence in the appropriateness of ChatGPT-generated materials as reference dialogues for EFL chatbot systems.

Additionally, the resulting Flesch Ease Reading scores indicate that most ChatGPT-generated materials are most suitable for students with CEFR levels A2 [20]. By referring further to the definition of CEFR level A2, the generated materials will be most appropriate to be used by students who exhibit the following characteristics.

- Vocabulary: Understand most everyday words and phrases related to personal information and basic needs; and many words and phrases related to hobbies, travel, and work.
- Grammar: Understand simple grammatical structures (e.g., present and past tenses) and basic question forms.
- Reading: Able to read short and simple texts, such as simple stories, with the help of a dictionary.
- Writing: Write basic sentences and short texts about personal experiences or daily routines.
- Listening: Understand simple and direct information in everyday conversations or short speeches on familiar topics.
- Speaking: Engage in basic conversations, and ask and answer questions about personal details, preferences, requests, or suggestions.

This interpretation was further supported by the resulting Dale-Chall scores obtained from the simulated dialogues. Although the Dale-Chall score calculation considers different criteria than the Flesch Reading Ease formula, a similar interpretation was reached.

VI. Conclusion

In this research, we investigate the potential of ChatGPT to generate reference dialogues to help EFL students improve their English proficiency. The reference dialogues might be helpful for an EFL chatbot in mobile applications considering more limited computing resources available on mobile devices. The underlying justification stems from the fact that simulating a deterministic conversation flow involves significantly fewer computational resources than running a complex Question and Answer Generation model. However, as users may feel bored practicing using the same lines of dialogue repeatedly, each line might need alternative replies to make the conversation more varied. Therefore, based on a dialogue generated by ChatGPT, alternative replies are created by asking the model to rephrase each line within the dialogue.

Moreover, we conducted an analysis using multiple readability metrics to determine the optimal target audience for the ChatGPT-generated materials. Only a few mini-words in the generated materials suggest they are free from wordy clichés, colloquial expressions, and phrasal verbs that could confuse EFL students. Furthermore, the resulting Flesch Ease Reading scores further affirm that the produced dialogues are most suitable for supporting students with CEFR A2. Likewise, the resulting Dale-Chall scores also support the same conclusion. The produced dialogues are well-suited for students with CEFR A2 proficiency, as they can comprehend most of the words used. Furthermore, a substantial portion of the dialogues intended for CEFR B1 can provide the CEFR A2 students with great stimulus to learn new words.

VII. Future Work

In future work, it would be valuable to investigate the potential of ChatGPT in generating reference dialogues for different target audiences, particularly those with CEFR B2 proficiency or above. This would involve exploring the adaptability of ChatGPT's dialogue generation capabilities to cater to the specific language needs and complexities of higher-level English learners. By expanding the scope of the study to include higher proficiency levels, we can assess the effectiveness of ChatGPT-generated materials in supporting the language learning journey of a wider range of EFL students.

Additionally, it would be beneficial to explore and experiment with different prompting techniques to further enhance the variety and quality of the dialogue generated by ChatGPT. By utilizing innovative techniques, such as direct task specification, task demonstration or mimetic proxy, we can potentially influence the generated dialogues to align more closely with the desired characteristics and objectives for different target audiences.

References

[1] M. Szmigiera, “The most spoken languages worldwide in 2022,” 2022, [Online]. Available: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

[2] H. Malik, M. A. Humaira, A. N. Komari, I. Fathurrochman, and I. Jayanto, “Identification of barriers and challenges to teaching English at an early age in Indonesia: an international publication analysis study,” Linguist. Cult. Rev., vol. 5, no. 1, pp. 217–229, 2021.

[3] O. F. Hibatullah, “The Challenges of international EFL students to learn English in a non-English speaking country,” J. Foreign Lang. Teach. Learn., vol. 4, no. 2, pp. 88–105, 2019.

[4] S. S. Khan and M. Takkac, “Motivational Factors for Learning English as a Second Language Acquisition in Canada.,” High. Educ. Stud., vol. 11, no. 1, pp. 160–170, 2021.

[5] M. Kurniawan and E. H. Radia, “A Situational Analysis of English Language Learning among Eastern Indonesian Students,” in 1st Yogyakarta International Conference on Educational Management/Administration and Pedagogy (YICEMAP 2017), 2017, pp. 1–6.

[6] P. Rosanda, E. Zehner, and W. Pensuksan, “The potentials and challenges of Indonesian nurses to use English in the hospital: A case study in a newly internationally accredited hospital in Indonesia,” Linguist. J. Linguist. Lang. Teach., vol. 4, no. 1, pp. 1–16, 2019.

[7] D. Xing and B. Bolden, “Exploring oral English learning motivation in Chinese international students with low oral English proficiency,” J. Int. Students, vol. 9, no. 3, pp. 834–855, 2019.

[8] P. S. Rao, “The importance of speaking skills in English classrooms,” Alford Counc. Int. English Lit. J., vol. 2, no. 2, pp. 6–18, 2019.

[9] S. Akhter, R. Haidov, A. M. Rana, and A. H. Qureshi, “Exploring the significance of speaking skill for EFL learners,” PalArch’s J. Archaeol. Egypt/Egyptology, vol. 17, no. 9, pp. 6019–6030, 2020.

[10] M. Shishido, “Evaluating e-learning system for English conversation practice with speech recognition and future development using AI Introducing the E - Leaning system with speech recognition,” in Proceedings of EdMedia + Innovate Learning, 2019, pp. 213–218.

[11] L. K. Fryer, D. Coniam, R. Carpenter, and D. Lăpușneanu, “Bots for language learning now: Current and future directions,” Lang. Learn. Technol., vol. 24, no. 2, pp. 8–22, 2020.

[12] J. Jeon, “Exploring AI chatbot affordances in the EFL classroom: Young learners’ experiences and perspectives,” Comput. Assist. Lang. Learn., pp. 1–26, 2021.

[13] N.-Y. Kim, “A study on the use of artificial intelligence chatbots for improving English grammar skills,” J. Digit. Converg., vol. 17, no. 8, pp. 37–46, 2019.

[14] D. Bailey, A. Southam, and J. Costley, “Digital storytelling with chatbots: mapping L2 participation and perception patterns,” Interact. Technol. Smart Educ., vol. 18, no. 1, pp. 85–103, 2020, doi: 10.1108/ITSE-08-2020-0170.

[15] D.-E. Han, “The Effects of Voice-based AI Chatbots on Korean EFL Middle School Students’ Speaking Competence and Affective Domains,” Asia-pacific J. Converg. Res. Interchang., vol. 6, no. 7, pp. 71–80, 2020, doi: 10.47116/apjcri.2020.07.07.

[16] N. Kim, “Chatbots and Korean EFL Students ’ English Vocabulary Learning,” J. Digit. Converg., vol. 16, no. 2, pp. 1–7, 2018.

[17] J. C. Young and M. Shishido, “Evaluating WaveNet Synthetic Speech for English as Second Language Listening Activities,” in Proceedings of 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), 2022.

[18] J. C. Young and M. Shishido, “Evaluation of Offline Automated Speech Recognition for English as Second Language Learning Application,” in Proceedings of EdMedia + Innovate Learning Online 2022, 2022, pp. 19–25. [Online]. Available: https://www.learntechlib.org/p/221652/

[19] OpenAI, “ChatGPT.” 2022. [Online]. Available: https://openai.com/blog/chatgpt.

[20] L. K. Fryer, K. Nakao, and A. Thompson, “Chatbot learning partners: Connecting learning experiences, interest and competence,” Comput. Human Behav., vol. 93, pp. 279–289, 2019.

[21] H. Kim, H. Yang, D. Shin, and J. H. Lee, “Design principles and architecture of a second language learning chatbot,” Lang. Learn. Technol., vol. 26, no. 1, pp. 1–18, 2022.

[22] J. Lee and Y. Hwang, “A Meta-analysis of the Effects of Using AI Chatbot in Korean EFL Education,” 영어영문학연구, vol. 48, no. 1, pp. 213–243, 2022.

[23] M. Shishido, “Developing and Evaluating an E-learning Material for Speaking Practice with the Latest AI Technology,” in The IAFOR International Conference on Education – Hawaii 2021, 2021. [Online]. Available: https://doi.org/10.22492/issn.2189-1036.2021.5

[24] R. Flesch, “A new readability yardstick.,” J. Appl. Psychol., vol. 32, no. 3, p. 221, 1948.

[25] J. S. Chall and E. Dale, Readability revisited: The new Dale-Chall readability formula. Brookline Books, 1995.

[26] R. McAlpine, Global English for global business. Longman Auckland, NZ, 1997.

[27] M. Gao, X. Liu, A. Xu, and R. Akkiraju, “Chatbot or Chat-Blocker: Predicting chatbot popularity before deployment,” in Designing Interactive Systems Conference 2021, 2021, pp. 1458–1469.

[28] B. Cárcamo Morales, “Readability and types of questions in Chilean EFL high school textbooks,” Tesol J., vol. 11, no. 2, p. e498, 2020.

[29] D. Yao, “A Comparative Study of Test Takers’ Performance on Computer-Based Test and Paper-Based Test across Different CEFR Levels.,” English Lang. Teach., vol. 13, no. 1, pp. 124–133, 2020.

[30] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[31] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[32] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[33] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[34] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.