Korean Journal of English Language and Linguistics, Vol 23, December 2023, pp. 1136-1153
DOI: 10.15738/kjell.23..202312.1136

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Young Korean EFL Learners' Perception of Role-Playing Scripts: ChatGPT vs. Textbooks

Sol Kim (Hyohaeng Elementary School) **Seon-Ho Park** (Gyeongin National University of Education)

Received: November 4, 2023
Revised: December 7, 2023
Accepted: December 13, 2023

Sol Kim (First author)
Teacher, Hyohaeng Elementary School
E-mail: solkim@hyohaeng.es.kr

Seon-Ho Park (Corresponding author)
Professor, Dept. of English Education,
Gyeongin National University of Education
E-mail: shpark@ginue.ac.kr

ABSTRACT

Kim, Sol and Seon-Ho Park. 2023. Young Korean EFL learners' perception of role-playing scripts: ChatGPT vs. textbooks. *Korean Journal of English Language and Linguistics* 23, 1136-1153.

This study explores the perceptions of elementary students in South Korea regarding two types of scripts used in reader's theaters: those derived from textbooks and those generated by ChatGPT. The research involved 27 fourth-grade students from Gyeonggi Province. The scripts consisted of six topics, each with dialogues presented in the textbooks and those created by GPT-3.5 to match the language proficiency of 9-year-old EFL learners. Students performed both script types in reader's theaters, and evaluations were conducted based on text flow, storyline attractiveness, English level, and practice process. Surveys were conducted twelve times, and results were analyzed using repeated measures of two-way ANOVA. The study revealed some statistical differences in the storyline attractiveness and English level, aligning with various student opinions. The study underscores the potential of integrating Artificial Intelligence (AI), such as ChatGPT, into English teaching while discussing pedagogical implications and emphasizing the need for differentiated approaches based on students' linguistic abilities. Suggestions for future research involving Chat GPT in elementary English education are provided.

KEYWORDS

ChatGPT, AI, textbook, elementary, EFL, perception, storyline, English level

1. Introduction

Does the advent of AI result in human obsolescence and unemployment? (Steele 2023) Will ChatGPT eventually supplant human teachers? (Ausat et al. 2023) These questions reflect the concerns of educators regarding the rapid advancements in artificial intelligence (Baidoo-Anu and Ansah 2023). These fears extend beyond the mere existence of teachers and delve into questioning how students should be educated. In this regard, Aoun (2018) emphasized the need for higher education to renovate its curriculum to train college graduates to become robot-proof.

Calculators, computers, the internet, and cell phones have transformed people's lifestyles and brought about significant changes in education. Since the introduction of the Turing test requirements (Turing 1950), Large Language Models (LLMs) such as ELIZA (Weizenbaum 1966), the Transformer model (based on attention mechanisms) (Vaswani et al. 2017), BERT (Devlin et al. 2018), GPT-3 (Brown et al. 2020), Galactica (Taylor et al. 2022), Switch-C (Fedus et al. 2022), and BARD (Ram and Pratima Verma 2023) have been developed, pushing the boundaries to function in a human-like manner. These innovations have demonstrated the feasibility of artificial intelligence. However, there are risks of fake information, biased results, deliberate misuse, and the amplification of a hegemonic worldview (Bender et al. 2021). In this rapidly changing era of data inundation, training students to evaluate AI-generated information and make informed judgments has become more significant than ever.

Studies have suggested the use of ChatGPT in educational settings, particularly concerning crafting effective prompts (Baidoo-Anu and Ansah 2023, Bonner et al. 2023). These studies stress the importance of providing context, identity, language proficiency level, and specific details when formulating prompts for GPT. While these insights are valuable, there is a lack of empirical research that explores how teachers and students perceive the content generated by ChatGPT.

Steele (2023) contends that students should engage critically with the content produced by GPT to deepen their *own* understanding (Italics in the original text). In the field of English education in South Korea, research on instructors' perspectives toward GPT usage has partially commenced. However, studies focusing on cultivating a critical attitude based on students' direct usage and experiences with GPT are still in the early stages. Jung et al. (2022) conducted a blind test with college students, who were pre-service teachers, using AI-generated texts produced by HyperCLOVA, a Korean-based LLM. The results indicated no statistically significant difference between AI-generated texts and the original content. The majority of respondents expressed that additional corrections were unnecessary and anticipated that text-generation tools could alleviate the workload of language teachers. Nevertheless, the study was conducted over a relatively short period, and the content was composed in Korean, with a specific focus on the tertiary level. Little has been revealed about the perspective of elementary EFL learners regarding ChatGPT-generated English materials and any potential comparisons between these materials and those created by human instructors.

Based on the awareness of the inadequacy of existing research, this study examines the practical utility of AI-generated scripts compared to textbook scripts for elementary English learners in Korea. To achieve this, students will participate in reader's theater performances using role-playing scripts generated by GPT-3.5 and those provided in textbooks. After experiencing and evaluating the characteristics of both types of scripts, a survey will be conducted to investigate the students' perceptions. Through this, we aim to uncover the educational implications necessary for the effective utilization of AI-generated materials in elementary English education in Korea. Based on these findings, we intend to propose directions for future education that are more efficient and practical.

2. Literature Review

Large Language Models (LLMs) are highly advanced artificial intelligence neural network models that undergo extensive training on textual data, enabling them to engage in human-like conversations and become proficient in the English language through a learning process. These self-supervised LLMs, foundation models, are celebrated for their remarkable versatility (Sejnowski 2023). They can perform a wide range of language-related tasks and swiftly adapt to new linguistic skills, requiring only minimal examples for adaptation. Over time, these models significantly grew in terms of parameters and training data size, giving rise to influential models such as BERT (Devlin et al. 2018), GPT-2, T-LNG, GPT-3 (Brown et al. 2020), GShard (Lepikhin et al. 2020), SWITCH-C (Fedus et al. 2022), and BARD (Ram and Pratima Verma 2023). For this study, GPT-3.5 was the model of choice.

ChatGPT, a Generative Pre-Trained Transformer (GPT), is a general-purpose conversational chatbot unveiled by OpenAI on November 30, 2022 (Kim et al. 2022). GPT is an artificial intelligence language model that employs a deep neural network to generate human-like text based on input data, thus demonstrating natural language understanding and generation capabilities (By ChatGPT-3.5, October 11, 2023). For this study, users input prompts in a conversational language format, often in questions or incomplete sentences, and then ChatGPT responds (Bonner et al. 2023).

Concerns about the implementation of GPT revolve around the potential for generating inaccurate or biased content. These models are trained on internet-based text datasets, which may reflect dominant perspectives and encode biases that could harm marginalized groups (Bender et al. 2021). GPT-3, in its indiscriminate collection of internet data, can inadvertently incorporate hate speech and biased language into its text generation (Godwin-Jones 2021). Moreover, McGuffie and Newhouse (2020) argue that GPT-3 can produce content that emulates interactive, informative, and influential materials, potentially contributing to the radicalization of individuals toward extremist ideologies and behaviors, far-right ideologies and behaviors (Bonner et al. 2023). Another drawback is the absence of human interaction, as ChatGPT cannot replicate the same level of human engagement as a teacher (Ausat et al. 2023). Additionally, they exhibit limited understanding and a restricted capacity for personalized instruction. Because of these problems, educational institutions, including New York City Public Schools, the Los Angeles Unified School District, and universities like Sciences Po in Paris, ban student use of ChatGPT (Castillo et al. 2023).

Another concern toward GPT is the issue of plagiarism. Hassoulas et al. (2023) conducted a blind test to markers consisting of undergraduates and graduates. From the result, only 23% of undergraduates and 19% of graduates could identify the ChatGPT writings. They asserted that banning using generative AI is like stopping internet use. Therefore, they insisted that the priority is to use these tools responsibly to enhance lives, avoiding misuse. Meanwhile, Carlini et al. (2021) pointed out the vulnerability of language models, specifically GPT-2, to attacks aiming to extract hundreds of verbatim texts that include personal identification data. They warned that larger models are more defenseless than the smaller ones.

Despite these drawbacks, ChatGPT has versatile potential for use in education. In a pioneering educational study related to GPT, Zhai (2022) tasked ChatGPT with writing an academic paper titled "Artificial Intelligence for Education." ChatGPT produced a coherent, well-organized, and informative thesis within 2-3 hours, even including partially accurate professional information. The paper proposed primarily three key points: adjusting learning goals to enable students to utilize AI tools, with a focus on enhancing students' creativity and critical thinking skills. To achieve these objectives, GPT suggested that researchers should develop AI-based learning activities that engage students in solving real-world problems. Additionally, Zhai (2022) stressed the importance

of introducing novel assessment formats to evaluate students' creativity and critical thinking-areas in which AI cannot serve as a substitute.

Researchers explored ways to incorporate GPT in education. Bonner et al. (2023) provided specific guidelines on creating prompts suitable for classroom use. They suggested that ChatGPT or other LLMs can reduce the workload and assist in various educational tasks. In language learning, these LLMs can apply their extensive language knowledge in areas such as summarizing, interpreting, consolidating knowledge in specific fields, and presenting information following established genre conventions. The researchers showcased examples of prompts for creating learning materials and assessments. These examples include summarizing text at an appropriate language level, correcting grammar and mechanics, providing feedback, generating writing prompts, creating presentation notes, suggesting lesson ideas, and adapting texts for testing or reading practice. Independent tools like Text Analyzer (https://hub.cathoven.com/) can be used to verify the text's language level, such as CEFR B2.

Pack and Maloney (2023) provided more concrete prompt examples that language teachers can follow (see Figure 1). They found ChatGPT to be a valuable tool for pre-service TESOL teachers in creating materials and assessments, emphasizing that the output is more useful when the user provides specific instructions. Pack and Maloney (2023) proposed several ways to utilize GPT, including generating discussion questions, adapting materials for varying proficiency levels, crafting handouts with explanations and practice exercises, producing listening transcripts, devising assessment rubrics, and assessing students' written work. They acknowledged that if a language assessment aims to gauge students' comprehension of conveyed information, then inaccuracies in the output may be unimportant. However, when the purpose of listening is to inform knowledge, hallucinations in the generated content can become problematic. In both cases, when employing LLMs for assessment, teachers need to be cautious in using the output of AI. Ultimately, their recommendation involves compiling a set of pre-tested prompts that students can use or even creating a custom application or tool that allows students to input text and receive feedback from ChatGPT using only validated prompts.

Example general prompt:

Create a handout for English language learners that helps them learn and practice using passive voice.

Example specific prompt:

You are a professional English language teacher. Create a handout for CEFR B1 English language learners that helps them learn and practice using passive voice. It should begin with a simple explanation of what passive voice is and how it is used in academic writing. Provide a couple of model sentences. Then provide 3 series of practice exercises. Each series should contain 5 exercises and each series should be scaffolded, progressing from easy to difficult. Use the following words throughout the handout: employ, transition, isolated, insist on, preservation, accustomed to, assemble, contradiction, principle, pursue.

Figure 1. General prompt and specific prompt (Pack and Maloney 2023)

Another study supporting the use of ChatGPT in teaching and learning is Baidoo-Anu and Ansah (2023). They argued that GPT can promote personalized and interactive learning. They insisted that GPT can generate prompts for formative assessment activities, providing ongoing feedback to inform teaching and learning. According to

Baidoo-Anu and Ansah (2023), ChatGPT can be a valuable tool for teachers in two key ways. Firstly, it enables them to generate open-ended questions that align with their lesson objectives and success criteria. Secondly, it can produce detailed rubrics that clearly outline the requirements for students to achieve proficiency at varying levels.

Despite these advantages, there are also points to be aware of when using ChatGPT. Steele (2023) suggested that ChatGPT poses potential challenges to three main elements of modern education systems: the assessment of students' knowledge and skills, the accuracy of the information students acquire, and the market value of the skills. Steele (2023) argued that in the digital age, educational institutions have recognized the importance of imparting information literacy, which involves the capacity to assess and evaluate information sources. However, ChatGPT cannot comprehend elements like humor, originality, beauty, surprise, irony, envy, or love because it does not possess true understanding; it operates through imitation.

As such, despite some caveats, recent studies have explored the possibility of ChatGPT in language classrooms and the creation of customized prompts, but there has been a notable dearth of empirical research investigating user perspectives on GPT-generated content. An example of such research is the study conducted by Jung et al. (2022), which delved into the perceptions of pre-service teachers regarding AI-generated text using HyperCLOVA, a Korean-based LLM. They collected survey data from university students majoring in education to evaluate the face validity of AI-generated texts compared to human-written texts and to gauge opinions on the feasibility of AI text generation in educational contexts. The results of this study revealed no statistically significant difference between AI-written and human-written texts. Additionally, the study suggested the possibility of generating AI texts that do not require further modifications, thereby reducing the workload of Korean language teachers. Although there are studies like this, there is a lack of research on applications for young English foreign language (EFL) learners at the elementary level. Moreover, teaching students to assess ChatGPT's output remains rare despite its potential to enhance students' understanding.

The educational impact of authentic play and drama in language learning has been documented in various studies. For instance, Bora (2021) reported significant improvements in syntactic complexity and accuracy among 10 Italian high school students who participated in a 40-hour blended-drama class. Reader's theater, a specific form of drama, is recognized for enhancing reading fluency through repeated readings of the same text. Lekwilai (2014) demonstrated the effectiveness of reader's theater for adult EFL students in Thailand, suggesting its applicability in language classrooms with diverse student proficiency levels. Her study identified inaccurate pronunciation feedback and script selection as crucial factors in the overall practice.

However, crafting suitable scripts for elementary Korean students is a labor-intensive task that demands a considerable amount of time. Textbook scripts undergo rigorous scrutiny by English education experts, ensuring not only their overall quality but also their appropriateness for students at the respective proficiency levels. Developing such high-quality scripts individually for a single English teacher poses a considerable challenge.

Although ChatGPT cannot offer real-time feedback to students, it can assist teachers in planning the play's overall structure, determining line lengths, and selecting appropriate levels of English vocabulary. In this regard, ChatGPT proves valuable for adapting textbook scripts, and tailoring them to address the linguistic needs in grammar and pragmatic usage for the target learners. Despite being described as a 'stochastic parrot' (Bender et al. 2021), ChatGPT excels at generating content. The use of GPT in language classrooms for script creation remains relatively unexplored, although Bonner et al. (2023) have shown success in generating prompts for narrative writing.

To address the gap in current research, this study engages Korean EFL elementary students in the evaluation of two different types of texts: those generated by GPT-3.5 and those created by humans. By prompting students to analyze these texts from their perspectives, this research aims to foster their independent understanding and

provide valuable pedagogical insights in an age of AI technology. This endeavor seeks to bridge the divide between the expanding role of AI and the skills students need to harness its potential. The research questions of this study are as follows:

> How do Korean elementary school students assess and compare role-playing scripts in English textbooks with those generated by ChatGPT?

3. Methodology

3.1 Participants and Classroom Context

Participants in this study were recruited from a leading elementary school in Gyeonggi Province, acknowledged for its innovative approach to education through the integration of artificial intelligence (AI). The school served as a pioneer in AI education. The study involved twenty-seven 9-year-old students enrolled in an EFL classroom. Each student was provided with an individual Chromebook, fostering a technology-friendly learning environment. Participants actively engaged in reader's theater activities and provided evaluations for two distinct script types after each of the twelve performances conducted during the study.

An English diagnostic test provided by EBS (Korean Educational Broadcasting Station) was administered. According to the American Council on the Teaching of Foreign Languages (ACTFL) guidelines for listening and reading proficiency, all students fell within the novice level (Novice High: 4, Novice Mid: 12, Novice Low: 11). Henceforth, Novice High will be denoted as [High], Novice Mid as [Intermediate], and Novice Low as [Low]. Notably, female students exhibited a higher average English proficiency score (M = 50/100) compared to their male counterparts (M = 34/100).

Table 1. The number of students by English proficiency level (*n*)

Gender \ Level	High	Intermediate	Low
Male (13)	-	6	7
Female (14)	4	6	4
Total (27)	4	12	11

At the outset, the primary author introduced the concept of ChatGPT to the students and explained how they would be involved in the reader's theater activities. All students consented to participate in the study. Reader's theater was chosen as the primary activity because the student participants thoroughly enjoyed role-play. However, their English proficiency remained at a novice level according to the ACTFL standards. The students' linguistic abilities were lower than their cognitive capabilities, highlighting the English divide. This makes reader's theater a fit choice as it decreases students' anxiety, as they need to iterate the text aloud (Hautala et al. 2023).

3.2 Generating Scripts Using ChatGPT

ChatGPT (GPT-3.5) generates text by responding to prompts from users. Pack and Maloney (2023) suggested

an approach that involves assigning a role or identity to the GPT, defining a goal or purpose for the chatbot, and then specifying the context and constraints to delimit the scope of the responses. Providing more detailed information in prompts is generally beneficial. It's worth noting that the quality of ChatGPT responses was influenced by the prompts provided (Sejnowski 2023).

Furthermore, due to its characteristic as a LLM, there are variations in the responses generated each time a prompt is submitted (Bonner et al. 2023). Consequently, users may need to submit the prompt multiple times and select the best response. As a result, the researchers underwent several stages to obtain scripts tailored to the English proficiency levels of the students.

We initially tasked ChatGPT to create a *Peter Pan* script without specific prompts: "Would you make an easy role-playing script for primary EFL students? The theme is about *Peter Pan*." The resulting script did not meet the desired linguistic level, and the number of characters and English complexity level were too high. Bonner et al. (2023) demonstrated methods for generating narrative writing through ChatGPT. Their prompt was: "Create a short story introduction based on the following information: character, occupation, setting, action (p. 33)." Subsequently, the authors followed the approach of Bonner et al. (2023) and made modifications to the prompt, including the topic, the number of characters, and the linguistic level (see Figure 2), resulting in a more sophisticated script suitable for the students. Additionally, when the generated scripts exceeded the length of the textbook script, the researchers requested to be shortened. However, despite these concerted efforts, GPT-3.5 struggled to align with the beginner English level of Korean EFL students.

Create a short role-play script based on the following information:
1) Topic: Peter Pan
2) Character: Peter Pan, Wendy, Tinker Bell, Captain Hook
3) Linguistic level: Novice (ACTFL)
4) Students' age: 9

Figure 2. Generating a *Peter Pan* script with specific prompts

A simple strategy was providing ChatGPT with textbook examples ("You can refer to the following example"). This approach yielded adaptable responses. However, one drawback was that ChatGPT sometimes produced scripts that closely resembled the original examples. It often paraphrased expressions and added extra lines to align more closely with the provided literature. The GPT scripts exhibited a length approximately twice that of the textbook scripts. The responses were refined several times (Pack and Maloney 2023, Sejnowski 2023).

Ultimately, the researchers provided six topics (*Peter Pan, Happy Prince, The Shoemaker and the Elves, Gingerbread Man, The Honest Woodcutter*, and *The Prince and the Pauper*) of scripts starting from March to July. There were two types of scripts for each topic: generated by ChatGPT and in the textbook. For each topic, students performed with the textbook script first and then practiced with GPT scripts.

3.3 Data Collection

Data were collected from the following sources: survey results, interviews, classroom observations, and teaching logs. This triangulation provided a deeper and more accurate understanding of the research context. Survey responses were gathered on 12 occasions and served as the primary source. Students used a five-point Likert scale (1 for 'strongly disagree' to 5 for 'strongly agree') to assess the scripts after each presentation. These assessments covered various aspects, including the natural flow, the level of interest generated, the appropriateness of the

English language level, and the effectiveness of the peer practice process. Additionally, students had the opportunity to express their opinions through open-ended responses. Individual in-depth semi-structured interviews were conducted in the Korean language with focus group students (high-level girl, intermediate-level boy, and low-level boy). The interviews took place on May 23 and May 30, 2023, in the middle of the study. These interviews aimed to elicit reflections on their practice experiences. Questions for these interviews were meticulously designed and conducted individually after the third and sixth sets of reader's theater. During these interviews, students shared their preferences, reasons for selecting specific scripts, perceived strengths and weaknesses of each script type and provided suggestions for potential enhancements. All interviews were carried out in Korean, the student's native language, and were recorded and transcribed verbatim. The interviews' duration varied between 15 and 30 minutes. Teacher observations were supplemented by video recordings of three performances, further enriching the qualitative data collected during the study.

3.4 Data Analysis

Repeated measures analysis of variance (ANOVA) was conducted on data collected through six repeated measurements, using individual learners as the unit of analysis. Additionally, we performed paired sample analysis based on individual response counts. The study utilized a mixed-method approach, incorporating content analysis techniques as described by Patton (2014). The primary objective was to identify recurring themes and patterns in students' perceptions through the analysis of open-ended responses and interview transcripts. To develop the initial coding scheme, the researchers carefully reviewed teaching logs to supplement the survey and interview data, following a systematic process. Each section was then isolated for coding purposes, and codes with similar underlying meanings were grouped into common themes. Finally, these themes were categorized into two distinct categories. This iterative analytical process involved transitioning between teaching logs, student survey results, interviews, and prior research. The researchers extracted representative responses to document and capture recurring patterns of meaning within the students' statements. To ensure the privacy of the students, their names were anonymized using numbers (e.g., B1, G2).

4. Findings

4.1 Attractiveness of the Storyline

For the text flow of the scripts, no statistical difference was found between the two types, which aligns with Jung et al. (2022) and Hautala et al. (2023). Both types scored in the 'adequate' to 'satisfactory' range, with textbook scripts at 3.58 and GPT scripts at 3.54 (p = .620). Though not statistically meaningful, some female students explained that GPT scripts successfully captured the original storyline of the topic. A high-level learner, G10 found GPT scripts to be more appealing due to their similarity with the original text. This similarity was considered a significant factor in enhancing the attractiveness of GPT scripts.

> It was fun and like watching a movie! What GPT made had a smooth flow of words. The storyline is almost the same. GPT usually makes scripts that are close to the original text. The least fun one was, well, there almost was not one but *The Happy Prince*, the textbook script. It was different from the original story. In the textbook scripts, the character suddenly gets happy or sad without enough

> explanation. Usually, we don't abruptly say "I'm happy!" or "I'm so sad." Hence it was not enjoyable. I believe that GPT did a better job of making the natural flow. (G10, High, interview conducted on May 23, 2023)

A low-level learner, G7 also mentioned the charm of reflecting the original text. Despite some difficulties in English studies, her reading proficiency in her mother tongue (Korean) was excellent, allowing her to read thick Korean novels. She pointed out that GPT scripts are lengthier and exhibit a discrete text flow. Another low-level student, B11 insisted that the storylines in GPT scripts were more comprehensible than those in the textbooks, a viewpoint akin to G7 and G10. There is a moderate correlation ($r = 0.44$, $p < .05$) between the text flow and the level of interest. It implies that the natural flow of the text could have played a role in enhancing the storyline.

According to the student responses, 'fun' was the most frequently cited word, appearing 156 times. The appeal of the storyline was assessed by employing repeated measures of two-way ANOVA, with script type as the variable, to examine the interaction between the two factors. The main effect of script type was not found to be significant ($F = 1.461$, $p = .233$, $\eta^2 = .032$). However, the main effect of storyline attractiveness reached significance ($F = 6.099$, $p < .001$, $\eta^2 = .122$). Notably, a significant interaction effect between text type and storyline attractiveness was observed ($F = 6.483$, $p < .001$, $\eta^2 = .128$), indicating a meaningful interaction between text types and the attractiveness of the storyline.

Table 2. Repeated measures of two-way ANOVA results for storyline attractiveness

Source	*df*	*SS*	*MS*	*F*	*p*
Script type (T)	1	5.232	5.232	1.461	.233
Attractiveness (A)	5	3.449	3.449	6.099	<.001***
T×A	1	3.667	3.667	6.483	<.001***

Note. T×A: Type×Attractiveness, ***p < .001: significant effect by 2-way ANOVA

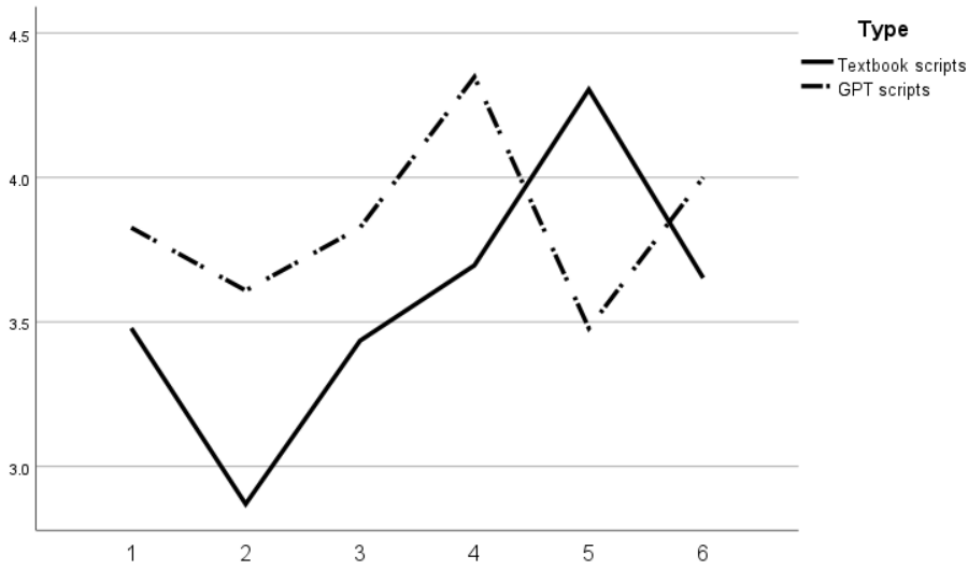

Figure 3. Repeated measures analysis of variance results for storyline attractiveness

In Figure 3, textbook scripts are depicted with a solid line, while GPT scripts are illustrated with a dotted line. Overall, students exhibited positive perspectives toward the GPT scripts, with the exception of the fifth session. Notably, students highly appraised the fourth GPT script, *Gingerbread Man*, as the most entertaining. In the textbook script, the Gingerbread Man typically queries other characters with simple questions like 'What are you doing?' or 'What is he/she doing?' The responses are given in the present continuous tense (e.g., 'I'm singing' or 'He is sleeping'). Even during encounters with the wolf, the exchanges remain straightforward, with the wolf stating, 'I'm eating a cookie!' In contrast, ChatGPT generated more intricate versions of the play by introducing a third-turn follow-up response from the Gingerbread Man. For instance, when inquiring about the turtle's activities, the counterpart responds, 'He is taking a nap. Turtle loves to sleep and rest.' In reaction, the Gingerbread Man adds, 'Oh, I should be quiet, so I don't disturb him.' ChatGPT enhanced the plot, making it more vibrant and the dialogue flow more naturally, simulating real-time conversation. Despite the presence of unfamiliar words (e.g., cheerful, gracefully, nap, and disturb), students did not find the script intimidating, rating it as a 'manageable level,' and expressing enjoyment.

However, during the fifth topic, GPT script was evaluated less interesting than the textbook script. The theme of the fifth session was *The Honest Woodcutter*. We posited that the level of English proficiency in the scripts might have influenced the level of interest. For instance, an intermediate-level student, B4 expressed that "This GPT script (*The Honest Woodcutter*) was too difficult, thus not interesting. However, the textbook version was easy and fun." This response suggests that the difficulty of the English language may have diminished the appeal of the GPT scripts, although the correlation is weak ($r = .15$, $p < .05$), a point to be discussed in the next section.

The following responses are from high-level groups who found longer and more concrete scripts in ChatGPT enjoyable, which aligns with their advanced English proficiency. When we consider a high-level group consisting of female students, these comments somewhat reflect female student opinions.

> The role of a boy was quite simple, and I would like to explore a different role next time. I am interested in trying something more challenging, and oddly enough, longer scripts tend to catch my attention more. (G3, High)

> The role-play from the textbook was too easy. I liked what GPT wrote better because it had more stuff to say. (G11, High)

Notably, the low-level group exhibited a preference for GPT scripts. Some of them expressed that if a script was interesting enough, it was worth trying. The following excerpts are from low-level students. When asked to rate the GPT script, they commonly answered that GPT scripts were exciting, albeit somewhat demanding.

> It was long and a bit tricky, but it was still fun. B2 played Tinkerbell, but he sounded like an old man. That was funny. The GPT script was long, but I enjoyed it because I got to talk a lot with my friends. And I loved it when G14 said I spoke English well – that made me happy! (G8, Low)

> I read this story in a book at home, so I knew it quite well. The last time, the script was shorter and simpler than today, but today it was cooler. There were lots of lines. The sentences were long. I want to do something harder next time, not just a short and easy script. (B11, Low)

> The previous one had fewer lines. Consequently, the storyline was not very interesting. However, today there were many lines, and the content was more fun. (B2, Low)

Regardless of their linguistic proficiency levels, most students were eager to embrace challenges if the story was engaging. Contrary to stereotypes, low-proficiency students did not shy away from difficulties. Instead, they actively sought out more captivating scripts that would serve as motivation. Several explanations may account for this phenomenon. First, one of the challenges with EFL students is the incongruity between their cognitive abilities and their English proficiency levels. Many students possess higher academic levels than their English proficiency, which could explain their preference for GPT scripts. They highly valued engaging content and the emotional aspects that included more nonverbal cues. As mentioned, the two boys emphasized that GPT scripts were more stimulating because they had abundant lines. Low-level students also pursued challenges, even if it required extra effort. Second, English textbook scripts in Korea have limitations, such as word count per sentence, sentence structure, and vocabulary range. However, ChatGPT has no constraints. It provides more intricate sentence structures, employs a real-time vocabulary that surpasses the textbook, and generates longer, more complex story scripts. This quality may appeal to students, motivating the low-level group and supporting them to participate in the task.

The following responses support this result. One of the intermediate-level students, B10, remarked, "It is worth trying something more difficult if it's more enjoyable" (Interview conducted on May 30, 2023). During performances, he often mimicked female voices, eliciting laughter from peers. He explained that attempting an interesting script was worth it. Not being reprimanded by the teacher for an unsuccessful performance, they are safe to try something new and hard. Similar responses were echoed by other intermediate-level students.

> The textbook script was short and boring, but the GPT script was long and fun. I wish the lines would get even longer next time. (B7, Intermediate)

> I kept repeating the same speech, and it was banal. Today, the script is longer than before and better. (G6, Intermediate)

To gain deeper insights into students' perceptions of more engaging scripts, we invited them to select scripts they found more fascinating. Overall, 17 students (62.96%) preferred GPT scripts as more appealing. Among 13 male students, seven (53.85%) opted for GPT scripts, while ten (71.42%) out of 14 girls chose GPT scripts. When examined by linguistic proficiency, three (75%) of high-level, seven (58.33%) of intermediate-level, and seven (63.64%) of low-level students favored GPT scripts over the textbook for their enchanting storyline.

4.2 Perceived English Level of the Scripts

In the responses, the word 'easy' appeared the second most frequently (71 times), following 'fun' (156 times), implying that ease influenced overall practice and performance. Students found that the manuscript difficulty of the textbook was more manageable than that of the GPT scripts in general.

Despite finding the GPT-generated script attractive, the students observed that the vocabulary was beyond their range. There was a substantial disparity between the students' present English level and the advanced English level in the GPT-generated one. Though the high-level group expressed that both types of scripts are manageable, (M =

5 for textbooks, $M = 4.875$ for GPT scripts), the other students with low-level or intermediate-level expressed that GPT scripts are far more challenging than the textbook scripts.

Perceived text English difficulty was examined using repeated measures two-way ANOVA with script type as the variable, analyzing the interaction between the two factors. The main effect of script type was found to be significant ($F = 7.090$, $p < .05$, $\eta^2 = .139$), while the main effect of English level reached significance as well ($F = 2.402$, $p < .05$, $\eta^2 = .052$). Notably, a significant interaction effect between text type and English level was observed ($F = 5.830$, $p < .001$, $\eta^2 = .117$), indicating a meaningful interaction between text types and English level.

Table 3. Repeated measures two-way ANOVA results for English level

Source	*df*	*SS*	*MS*	*F*	*p*
Script type (T)	1	47.087	47.087	7.090	.011*
English level (L)	5	4.783	.957	2.402	.038*
T×L	5	11.609	2.322	5.830	<.001***

Note. T×L: Type×English level, *p < .05, ***p < .001: significant effect by 2-way ANOVA

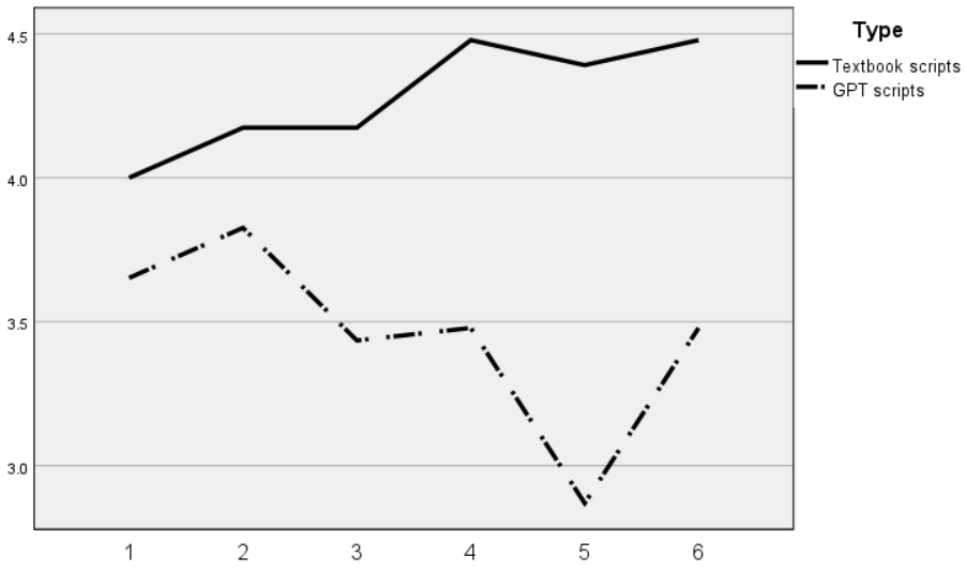

Figure 4. Repeated measures analysis of variance results for English level

In Figure 4, textbook scripts are indicated by a solid line, while GPT scripts are delineated with a dotted line. The distinction between the lines is evident throughout all topics and is particularly highlighted in the fifth session, featuring the story of *The Honest Woodcutter*. The GPT script contained words that surpassed their English curriculum, such as 'chop, axe, shiny, either, wooden handle, or fell into'. Furthermore, the length of the sentences was longer in the GPT scripts. In the textbook, sentences are at most 7 words long. In contrast, in the GPT script, some sentences consist of nine words (e.g., 'As a reward, you can keep all three axes!') or even 16 words (e.g., 'I was honest with a water fairy in the forest, and she gave them to me!'). Students appeared to be stressed by the combination of challenging vocabulary and lengthy sentences. The following excerpt shows distinct student dynamics. While students appeared more relaxed and autonomous during textbook practice sessions, they

exhibited tension and anxiety while practicing the GPT script.

> During practicing the textbook skit, students were frequently seen with smiles or laughter, showcasing high engagement across all proficiency spectrums. Some teams debated nonverbal cues, discussing the types of gestures they would use and the directions of the moves. The overall atmosphere exuded a relaxed and relaxed vibe. However, when they rehearsed with scripts generated by GPT-3.5, there were rough facial expressions. Students seemed somewhat tense, lacking the same sense of ease as before. This unease was evident during the performance, with some students displaying heightened sensitivity. While errors in the textbook script prompted laughter, the GPT-generated led to a somewhat accusatory tone, indicating a difference in response. (Teaching log obtained on June 22, 2023)

To investigate this further, we asked students which type of script they would choose if they could not receive additional assistance in reading English vocabulary from the teacher. To this question, 21 students answered textbook scripts over GPT scripts. Six students who favored GPT scripts were three high-level girls, two intermediate-level boys, and one low-level girl. Even if the storyline of the GPT script was enchanting, students struggled to read or enunciate the script lines. This difficulty was manifest even in a high-level English student. G4, the only high-level student who chose the textbook scripts over GPT scripts, explained her preference due to their simplicity and shortness. In an open-ended response, she emphasized that the easy one facilitated relaxed practice with classmates.

> I found the textbook script more engaging because the GPT script was longer and had more difficult words than I thought. I think I did better with the textbook script compared to today. (G4, High)

Intermediate-level students displayed difficulties with GPT scripts. An intermediate-level student, B9 highlighted that the textbook scripts were shorter than those generated by ChatGPT. He explained that reading textbook scripts with friends resulted in much laughter and created a relaxed atmosphere (Interview conducted on May 30, 2023). Other intermediate-level students showed similar opinions.

> The stuff made by GPT seems quite hard. I kept messing up. I felt like my confidence went down. (G1, Intermediate)

> The script created by GPT has a natural flow of writing. However, it was a bit challenging because English is somewhat difficult. (G14, Intermediate)

> You can memorize lines in textbook skits. However, the dialogues in GPT scripts are often too long and contain unknown words. I am clueless about how to pronounce some of the words. (B4, Intermediate)

Students with lower proficiency levels experienced a sense of relief when using textbook scripts. These scripts featured lines consisting of fewer than five words, specifically designed for easy recall, which significantly boosted their confidence. This kind of simple text increased their overall self-assurance. In general, other students also exhibited increased confidence during their performances of the textbook skits.

> The textbook script is easy and fun. (B6, Low)
>
> This one was easy. I liked it. (G9, Low)
>
> I said everything from memory. It was a lot of fun! (B11, Low)

In the meantime, the study found no statistically significant differences in the practice process between textbook and GPT scripts. Although there is no strong correlation between English proficiency and the practice process ($r = .24$, $p < .05$), some responses suggested a modest influence of English level on overall rehearsal dynamics.

> Memorizing the lines was easy, and my friends were good at English. Thus, the practice was easy. (B1, Intermediate)
>
> GPT-English was tough, and the practice process went rough. It would be more fun if it were a bit easier. (G14, Intermediate)
>
> It used to be okay, but today (GPT) was a total mess. It was painful. My friends were not paying attention and were ignorant of what to do. They didn't even know their turn. That made me frustrated and annoyed. (G1, Intermediate)

Through practicing and performing six sets of scripts, students were aware of the captivating storyline of the GPT script. Conversely, aside from high-level students, several students evaluated their English proficiency in GPT scripts as inadequate. For instance, an intermediate-level student, G5 emphasized the importance of offering pronunciation assistance (Sardegna and McGregor 2013). She said this kind of help can alleviate discomfort and enhance student confidence. This observation remained consistent across both genders, indicating that GPT requires detailed scaffolding to meet the specific needs of students. While GPT, in its current state, may not be fully prepared for independent use, it can effectively address diverse student needs when guided by classroom teachers.

5. Conclusion

This study explores the perceptions of young Korean EFL students regarding scripts generated by GPT-3.5 in comparison to textbook scripts. The participants engaged in reader's theater performances over six sessions. Following each performance, students were requested to assess two script types (totaling twelve scripts) and share their opinions. The overall text flow of the two types showed no statistically significant difference, as revealed by repeated measures of two-way ANOVA. Moreover, although storyline attractiveness exhibited no statistically significant difference, students consistently rated GPT scripts as more interesting for five times. The only exception was when the English level of the GPT scripts was too challenging, leading students to find textbook scripts more enjoyable. Concerning the English proficiency of the scripts, textbook scripts outperformed GPT scripts, indicating a statistically meaningful difference. Notably, low-level students expressed a preference for and perceived greater engagement with GPT scripts over textbook scripts, despite their difficulty. This suggests that while students found

GPT scripts challenging, they were more motivated to work on them, making them worth trying. Meanwhile, some students evaluated textbook scripts as more suitable in terms of English proficiency, facilitating smooth practice and creating a responsive atmosphere.

The research findings have provided several educational implications for the future use of AI in English classrooms. First, when crafting scripts using ChatGPT, individual teacher expertise is still required to fine-tune the linguistic level of the results just above the students' proficiency. For instance, even when prompted with the expected linguistic level, ChatGPT may fail to adapt its English proficiency. In addition, novice level span does not seamlessly align with the specific students that each teacher encounters. In this study, students found GPT scripts more engaging than textbook scripts across five topics, except when they considered the English difficulty level of the GPT script to be too challenging. In such cases, students preferred the textbook script for greater acclimatization. They enjoyed incorporating novel vocabulary that would enhance the realism of the masterpiece (e.g., 'Sprinkle some pixie dust on us' from *Peter Pan*). However, when students perceived the script as too challenging, they became less attentive. This implies that, despite ChatGPT's capacity to generate numerous educational resources, the challenge of aligning the English proficiency level just beyond the students' linguistic abilities persists. This underscores the need for teachers to carefully deliberate on setting the English level slightly above that of the learners (Krashen and Terrell 1983) and safeguard the overall appeal of the generated content.

Second, it demonstrates that ChatGPT can assist low-level group learners not only with linguistic support but also based on their cognitive capacity and their willingness to participate in classroom activities. Empowering them highlights the possibility of changes in their classroom dynamics as they can make more meaningful contributions to the learning process. Unlike the traditional classroom setting where they often depended on assistance from human peers, they now have a more equitable opportunity to compete with their fellow students. When provided with guidance on how to approach unfamiliar English vocabulary, learners were more inclined to tackle more challenging texts. Achieving this is relatively straightforward when using an app or another type of AI. In addition, students found ChatGPT-generated scripts to be both arduous and captivating. This may be because these scripts push the boundaries and provide a more authentic reflection of the real-life language that students require. As Tarone and Swain (1995) emphasized, learners seek real languages that can effectively express themselves because the languages they encounter in textbooks often fall short of capturing their true intentions. Despite demanding vocabulary, students were willing to choose GPT scripts because of their charms.

Third, English teachers utilizing GPT should engage in continuous interaction with GPT until it produces content that meets specific expectations. When using GPT, it is important to provide prompts that give the AI a clear sense of identity, the linguistic level of the user, and contextual information (Pack and Maloney 2023). However, even with detailed prompts, GPT may not generate the desired response right away. This implies that despite its potential, users should have a clear goal for the outcomes AI can produce and be patient in providing further prompts (Sejnowski 2023). This suggests that language instructors and learners need to engage in a 'negotiation of meaning' process (Van der Zwaard and Bannink 2014) with GPT. In particular, the challenge arises in that most Korean elementary EFL learners may belong to the novice level. When the authors requested GPT to generate scripts for Korean EFL learners, it often struggled to produce appropriate content. Consequently, in this study, we resorted to using textbook scripts as an example. However, ChatGPT often replicated the scripts and added a few additional lines. This work was slightly superior to textbooks in terms of plot, but it is worth noting that it lacked creativity and originality.

Despite the discovery of significant issues, this study has inherent limitations due to the relatively small number of participants, which may not fully represent the overall spectrum of aspects such as students' English levels, and age range. In particular, the limited number of students with a certain proficiency level made it challenging to

execute more diverse trials of GPT-generated scripts and verify their educational effectiveness. Nevertheless, the results of this study shed a small ray of light on the potential practical use of AI in elementary English classrooms, an area that is not widely adopted in the current Korean elementary education curriculum. Additionally, this study revealed the potential to leverage GPT for more flexible and engaging teaching and learning experiences, emphasizing the need to operate it considering diverse student proficiency levels. In line with the guidelines provided by the Korean national curriculum, the study confirmed that utilizing GPT in teaching and learning processes could further stimulate students' creativity and interest. As Steele (2023) noted, "ChatGPT functions as a synthesizer and a mimic, not a thinker." However, students should aim to be critical thinkers rather than mere imitators. For future research, education experts should pay attention to developing ways to formulate effective prompts for ChatGPT and other AI tools. Furthermore, examining the questions students generate when using AI can enhance our understanding of young students.

References

Aoun, J. E. 2018. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. MIT Press.

Ausat, A. M. A., B. Massang, M. Efendi, N. Nofirman and Y. Riady. 2023. Can chat GPT replace the role of the teacher in the classroom: A fundamental analysis. *Journal on Education* 5(4), 16100-16106.

Baidoo-Anu, D. and L. Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7(1), 52-62.

Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. Association for Computing Machinery.

Bonner, E., R. Lege and E. Frazier. 2023. Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology* 23(1), 23-41.

Bora, S. F. 2021. Taking literature off page! The effectiveness of a blended drama approach for enhancing L2 oral accuracy, pronunciation and complexity. *Language Teaching Research*, available online at https://doi.org/10.1177/13621688211043490

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, ... and D. Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877-1901.

Carlini, N., F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, ... and C. Raffel. 2021. Extracting training data from large language models. In 30th *USENIX Security Symposium (USENIX Security 21)* 2633-2650.

Castillo, A., G. Silva, J. Arocutipa, H. Berrios, M. Rodriguez, G. Reyes, ... and J. Arias-Gonzáles. 2023. Effect of Chat GPT on the digitized learning process of university students. *Journal of Namibian Studies: History Politics Culture* 33, 1-15.

Devlin, J., M. W. Chang, K. Lee and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Fedus, W., B. Zoph and N. Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research* 23(1), 5232-5270.

Godwin-Jones, R. 2021. Big data and language learning: Opportunities and challenges. *Language Learning & Technology* 25(1), 4-19.

Hassoulas, A., N. Powell, L. Roberts, K. Umla-Runge, L. Gray and M. Coffey. 2023. Investigating marker accuracy in differentiating between university scripts written by students and those produced using ChatGPT. *Journal of Applied Learning & Teaching* 6*(*2), 1-7.

Hautala, J., M. Ronimus and E. Junttila. 2023. Readers' theater projects for special education: A randomized controlled study. *Scandinavian Journal of Educational Research* 67(5), 663-678.

Jung, H., Y. Lee and D. Shin. 2022. A survey study on pre-service teachers' perceptions of AI-generated texts. *Bilingual Research* 90, 193-217.

Kim, H., H. Yang, D. Shin and J. Lee. 2022. Design principles and architecture of a second language learning chatbot. *Language Learning & Technology* 26(1), 1-18.

Krashen, S. D. and T. D. Terrell. 1983. *The Natural Approach. Language Acquisition in the Classroom.* Oxford: Pergamon Press Ltd.

Lekwilai, P. 2014. Reader's theater: An alternative tool to develop reading fluency among Thai EFL learners. *Journal of Language Teaching and Learning in Thailand* 48, 89-111.

Lepikhin, D., H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, ... and Z. Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668.*

McGuffie, K. and A. Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. Technical Report. Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey.

Pack, A. and J. Maloney. 2023. Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. *Teaching English with Technology* 23(2), 4-24.

Patton, M. Q. 2014. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. Sage publications.

Ram, B. and P. V. Pratima Verma. 2023. Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI. *World Journal of Advanced Engineering Technology and Sciences* 8(1), 258-261.

Sardegna, V. G. and A. McGregor. 2013. Scaffolding students' self-regulated efforts for effective pronunciation practice. In J. Levis and K. LeVelle, eds., *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*, 193-214. Ames, IA: Iowa State University.

Sejnowski, T. J. 2023. Large language models and the reverse turing test. *Neural Computation* 35(3), 309-342.

Steele, J. L. 2023. To gpt or not gpt? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence* 5, 100160.

Tarone, E. and M. Swain. 1995. A sociolinguistic perspective on second language use in immersion classrooms. *The Modern Language Journal* 79(2), 166-178.

Taylor, R., M., Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, ... and R. Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Turing, A. M. 1950. I.—Computing machinery and intelligence. *Mind* 59(236), 433-460.

Van der Zwaard, R. and A. Bannink. 2014. Video call or chat? Negotiation of meaning and issues of face in telecollaboration. *System* 44*,* 137-148.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30, 1-11.

Weizenbaum, J. 1966. ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1), 36-45.

Zhai, X. 2022. ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.

Examples in: English
Applicable Languages: English
Applicable Level: Elementary, Secondary, Tertiary