Faculty of Education Journal- Alexandria University **Volume 34 (Issue 3) 2024**

A Linguistically Developed Prompt Engineering Parameters Model for Enhancing AI's Generation of Customized ESL Reading Texts

∗ **Dr. Hebatollah M. M. Hegazy**

Received: 19/8/2024 **Accepted: 28/9/2024**

Abstract

The quick progression of artificial intelligence (AI), especially Large Language Models (LLMs) now offers unprecedented opportunities for creating personalized learning experiences in the education field, as in the field of teaching English as a Second Language (ESL). The present study investigates the integration of prompt engineering with linguistic theories as a means of guiding LLMs to produce customized ESL reading material that is both linguistically accurate and pedagogically sound, addressing the diverse needs of ESL learners via building a linguistically-informed and user-friendly model of prompt parameters. Through such model, the study seeks to benefit educators regardless of them being experienced enough in prompt engineering and generative AI. The theoretical framework of this study is based on developing a comprehensive model of prompt engineering that integrates elements from three well-known linguistic theories: Transformational Generative Grammar, Systemic Functional Linguistics, and Global Englishes, along with basic prompt engineering elements. Using a mixed-method approach of quantitative and qualitative analysis, the study evaluates the effectiveness of this model. To test the model, six reading texts at different levels of the Common European Framework of Reference for language proficiency (CEFR) are generated by an LLM chatbot named Microsoft Copilot. These texts serve a variety of purposes and are of different genres. The readability scores of the generated texts are analysed using a combination of three metrics. Alongside this, a detailed qualitative analysis of each text is also undertaken. Together, these have revealed a general alignment between the texts and the targeted CEFR levels as well as their adherence to elements of the employed linguistic theories as requested in the devised prompt for each generated text. This demonstrates the developed model's efficiency in enhancing the AI's ability to produce reading material that is responsive to the diverse language levels and needs of the ESL learners, hence contributing to both creating more suitable learning experiences within ESL pedagogy and endorsing the integration of generative AI with linguistic theories to help teachers satisfy such needs.

Keywords: prompt engineering, large language models, generative AI, transformational generative grammar, systemic functional linguistics, global Englishes, customized ESL reading materials

---

An Assistant Professsor of Linguistics & Translation – Dept. of English Language – Faculty of Education - Alexandria University/ Egypt.
Email: hebatollahheg@alexu.edu.eg

نموذج معايير لهندسة الاستعلامات مطور لغويا لتعزيز توليد الذكاء الاصطناعي لنصوص القراءة المُخَصَّصة في مجال تعليم اللغة الإنجليزية كلغة ثانية

د.هبة الله محمود محمد حجازي

المستخلص

لقد أتاحت التطورات السريعة في مجال الذكاء الاصطناعي AI ، وخاصة في مجال النماذج اللغوية الكبيرة LLMs، إمكانيات جديدة للتعلم الشخصي في مجال التعليم، وخاصة في مجال تعليم اللغة الإنجليزية كلغة ثانية ESL. وتستكشف هذه الدراسة إمكانات دمج هندسة الاستعلامات مع النظريات اللغوية لتحسين قدرة النماذج اللغوية الكبيرة على توليد نصوص قراءة مخصصة في مجال تعلم اللغة الإنجليزية كلغة ثانية. تهدف الدراسة إلى تطوير وتقييم نموذج معرّف مسبقا لمعايير هندسة الاستعلامات بإمكانه أن يوجه النماذج اللغوية الكبيرة لإنتاج نصوص دقيقة لغويا وسليمة من الناحية التربوية، وتلبي الاحتياجات المتنوعة لطلاب اللغة الإنجليزية كلغة ثانية. ومن خلال تبسيط هندسة الاستعلامات من خلال نموذج معرّف مسبقا، تسعى الدراسة إلى تعزيز إمكانية استخدام المعلمين لتقنيات الذكاء الاصطناعي، بغض النظر عن مستوى خبرتهم. أما بالنسبة للإطار النظري، فقد تم تطوير نموذج شامل لمعايير الاستعلامات، يدمج عناصر من ثلاث نظريات لغوية بارزة وهي: النحو التوليدي التحويلي TGG، واللغويات الوظيفية النظامية SFL، والإنجليزيات العالمية GEs، مع تقنيات الذكاء الاصطناعي التوليدي من خلال المبادئ الأساسية لهندسة الاستعلامات. وتستخدم الدراسة منهجية مختلطة، تجمع بين التحليل الكمي والكيفي للبيانات لتقييم فعالية النموذج. وتتضمن البيانات ستة نصوص قراءة مختلفة لطلاب اللغة الإنجليزية كلغة ثانية، تستهدف مستويات لغوية مختلفة من الإطار المرجعي الأوروبي الموحد للغات CEFR وتغطي أنماطا مختلفة من النصوص، وقد تم توليد هذه النصوص بواسطة "مايكروسوفت كوبايلوت" Microsoft Copilot، وهو روبوت دردشة مدعوم بتقنيات النماذج اللغوية الكبيرة المتقدمة. وقد كشفت نتائج تحليل درجة سهولة قراءة النصوص المولدة، باستخدام مزيج من ثلاث مقاييس، إلى جانب التحليل الكيفي المفصل لكل نص، عن توافق عام بين النصوص المولدة ومستويات اللغة الخاصة بالإطار المرجعي الأوروبي الموحد للغات، والتزامها بمبادئ النظريات اللغوية المستخدمة، مما يدل على قدرة الذكاء الاصطناعي على تكييف تعقيد اللغة وفقا للمعايير المحددة، وعلى دمجه الفعال للخصائص اللغوية والعناصر الأسلوبية والأمثلة الحساسة ثقافيًا المطلوبة في الإستعلامات المصممة. وهذا بدوره يُظهر قدرة النموذج المطور اللغويا على إنتاج نصوص دقيقة لغويا وذات صلة وظيفيا وملائمة ثقافيًا، وبالتالي إمكاناته في إثراء تدريس اللغة الإنجليزية كلغة ثانية، ويساهم أيضًا في دعم دمج قوة الذكاء الاصطناعي مع رؤى النظريات اللغوية لمساعدة المعلمين في إنشاء تجارب تعليمية مُخصصة ومثيرة للاهتمام وذات صلة ثقافيًا لطلاب اللغة الإنجليزية كلغة ثانية حول العالم.

الكلمات المفتاحية: هندسة الاستعلامات، النماذج اللغوية الكبيرة، الذكاء الاصطناعي التوليدي، النحو التوليدي التحويلي، اللغويات الوظيفية النظامية، الإنجليزيات العالمية، نصوص قراءة مخصصة في مجال تعلم اللغة الإنجليزية كلغة ثانية

1. Introduction

The development of generative artificial intelligence (AI), particularly large language models (LLMs) has increased the potential to create customized material for various purposes (Li et al., 2023), and what enhances such LLMs' performance is actually prompt engineering. Research studies have underlined that effectively engineered prompts can considerably improve LLMs' performance in generating specified conversational responses and outputs, like those required in radiology (Russe et al., 2024). More specifically, research on AI language learning tools, as evident in Barrett and Pack (2023), proposes that prompt engineering could help create personalized learning experiences for students of English as a second language (ESL). Nevertheless, prompt engineering may present a challenge, particularly for non-experts who struggle to devise effective prompts. Whereas existing research does not address the creation of a predefined model of prompt parameters for ESL material generation, it underscores the capability of prompt engineering in guiding LLMs' output, as stated by Li et al. (2023). A model could be designed by combining theories of language and the principles of prompt engineering so that LLMs could be guided to produce customized ESL materials. This will involve an integrated approach to meeting the basic needs of language learning and the particular demands of ESL education. Thus, research into this area is essential in order to develop and assess such a model, which would in turn unleash AI’s full potential in producing personalized customized ESL reading texts across different genres, types and registers.

Customization, in ESL education, is central in delivering effective learning experiences. Since ChatGPT has started to be prominent in education, a strong possibility for building and adapting language models to augment ESL learning has become apparent (Leong et al., 2023). Both educators and learners can work on content that suit their specific linguistic needs as well as preferred learning styles via employing generative AI to produce customized ESL reading materials, which, in turn, endorses more effective language acquisition, proficiency, and engagement through affording targeted support (Fryer et al., 2020; Young & Shishido, 2023). Furthermore, in Ochieng's (2023) words, AI can help create diverse and engaging reading materials that satisfy students' interests and real-world experiences. Such personalization boosts the learning process by causing it to be more relevant and stimulating. However, educators, as stressed by Labruna et al. (2023), must seriously evaluate the quality and accuracy of AI-generated content to ensure it meets educational objectives and standards.

Indeed, chatbots are valuable language-learning tools especially for those studying English as a Second Language (Petrovic & Jovanovic, 2020). Studies have shown that they also improve reading comprehension of ESL and other language skills (Jiang, 2022) and positively affect vocabulary acquisition, sentence structure, spelling and pronunciation (Mohamed & Alian, 2023). In addition, Devlin et al. (2019) and Radford et al. (2019) argued that contemporary natural language processing (NLP) models can be utilized to improve the capabilities of chatbots in providing personalized learning experiences and feedback for ESL students.

Studies have shown that LLMs like ChatGPT can generate texts of high quality and various styles and for various purposes (Labruna et al., 2023; Ochieng, 2023; Young & Shishido, 2023). Studies have also revealed the effectiveness of prompting techniques in guiding LLMs to produce specific outputs (Li et al., 2023; Woo et al., 2023). Morevoer, from a linguistic perspective, theories, such as Transformational Generative Grammar (TGG), Systemic Functional Linguistics (SFL), and Global Englishes (GEs), (see Section 4.2), offer valuable insights into the structure, function, and sociocultural context of language. However, to my best knowledge, despite the potential of LLMs, precious knowledge from language theories, and importance of prompt engineering techniques, these elements have not been systematically combined in current research to create a model of prompt parameters that can efficaciously instruct LLMs in generating customized ESL reading materials to cater to the specific language needs of ESL learners. The present research aims to address this knowledge gap by developing and evaluating such predefined model that integrates basic prompt engineering elements with insights from language theories to enhance LLMs' generation of ESL reading texts that address the linguistic needs of the ESL learners and save the teachers' time and effort in searching for and finding reading materials suitable for their learners' language proficiency levels, thus contributing to improving the efficiency and effectiveness of ESL teaching and learning.

2. Research Objectives

This study aims to develop and evaluate a predefined model of prompt parameters that can enhance the generation of customized ESL reading material by LLMs. To achieve this aim, the study seeks to achieve the following sub-objectives:

2.1 Model Development

- Identify and integrate relevant elements from prominent language theories, TGG, SFL, and GEs, into a structured model of prompt parameters.
- Incorporate basic elements of effective prompting along with elements from the three language theories to guide LLMs, particularly the utilized Microsoft Copilot as an example LLM, in generating customized ESL reading material.

- Develop a clear and user-friendly framework for applying the predefined model of prompt parameters, making it accessible to educators with varying levels of expertise in prompt engineering.

2.2 Model Evaluation

- Generate a diverse set of ESL reading texts using the developed model, targeting different language proficiency levels based on the Common European Framework of Reference (CEFR) (see Section 6.2.1.3).
- Assess the readability of the generated texts using multiple readability metrics, Flesch-Kincaid Reading Ease (FKRE), Flesch-Kincaid Grade Level (FKGL), and Gunning Fog Index (GFI), (see Section 6.2.1), to determine their suitability for the intended CEFR levels.
- Analyze the extent to which the generated texts incorporate the specified prompt parameters derived from language theories and basic prompt engineering, evaluating the LLM's understanding and execution of the prompts.
- Explore the versatility of the developed model by generating texts across different text types and registers, demonstrating its adaptability to various pedagogical needs.

3. Research Questions

The main research question is: how can the integration of prompt engineering and language theories be leveraged to create a predefined model of prompt parameters that enhances the generation of customized ESL reading material by AI LLMs? To effectively address the research objectives, the researcher has broken it down into the following sub-questions:

3.1 Model Development

1. Which elements from TGG, SFL, and GEs are most relevant for guiding LLMs in generating customized ESL reading materials?
2. What basic prompt engineering elements can be integrated with these linguistic elements to create a structured and effective model of prompt parameters?
3. What practical considerations should be addressed in the model development process to ensure its usability and accessibility for educators?

3.2 Model Evaluation

1. To what extent can the developed model guide LLMs, e.g. Microsoft Copilot, used in the current study, in generating ESL reading texts that align with different CEFR proficiency levels?
2. Do the readability scores of the generated texts, as measured by multiple readability metrics, correspond to the intended CEFR levels?
3. How effectively does the employed LLM incorporate the specified prompt parameters derived from the targeted language theories and basic prompt engineering into the generated texts?
4. How versatile is the developed model in generating customized reading materials across different text types and registers?

4. Theoretical Framework

This section establishes the theoretical groundwork for the current study by underscoring the intersection of generative AI and linguistics, as elucidated below:

4.1 Generative AI

Generative AI refers to a category of AI systems designed to create novel content, typically mimicking human creativity, based on patterns and information learned from existing data; these systems analyze and produce linguistic data in diverse ways (Dong et al., 2022). To explicate, a generative AI system can be instructed to create a text on a specific topic, such as a news article, a short story, an email, a medical report, etc., and analyze the linguistic features of such generated output, like word choices, grammatical structures, stylistic features, and discourse patterns (Dong et al., 2022). Furthermore, generative AI, as stated in Devlin et al., 2019, Dong et al., 2022, and Ouyang et al., 2022, can create texts addressing pre-determined linguistic parameters, level of difficulty, writing style, and register, i.e. customized learning materials, since these AI systems, by analyzing enormous amounts of text data, learn the various linguistic rules sentence structure (syntax) and word meanings (semantics). AI, for instance, can quickly process large volumes of text, saving time and effort in tasks like reviewing language patterns or eliciting information from written content. Examples of generative AI models are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and, importantly, LLMs such as BERT, T5, GPT-3, LaMDA, and PaLM, among others (Brown et al., 2020;

Chowdhery et al., 2022; Thoppilan et al., 2022; Devlin et al., 2019; Ouyang et al., 2022; Raffel et al., 2020). As these LLMs are trained on huge amounts of text data, they have displayed remarkable performance in various linguistic tasks, such as text generation, translation, summarization, and question answering.

LLMs can, accordingly, be used for creating language practice material for various learning objectives and proficiency levels as pointed out by Jiang (2022), Kurdi et al. (2020), and Mohamed and Alian (2023). For instance, in their work, Kurdi et al. (2020) observed that LLMs could be tasked with creating reading comprehension texts, vocabulary exercises, and dialogues mimicking real life everyday conversations, among others so as to make the process of learning more meaningful to learners. Furthermore, being experts in assessing language practice materials, LLMs evaluate the readability and complexity of text, identify problematic areas for learners, correct students' written works, which gives them specific tutoring and guidance about their performance (Cavalcanti et al., 2021). For example, Zhao et al. (2022) performed a study on the use of BART summarization model trained on Fairytale-QA data set for question generation. In this case, the model was taught to create summaries containing key events and derive questions aligned with specific learning goals from given facts regarding fairy tales.

To adequately use LLMs, prompt engineering is indispensable. It involves the planned design and shaping of instructions, named as prompts, which guide LLMs in generating the intended response (Liu et al., 2021). Such procedure requires prudently considering various elements of the prompt, such as the wording, context, target audience, desired length, format, tone and style of the output (Atlas, 2023; "Prompt Engineering," 2024; Reynolds & McDonell, 2021). Furthermore, there are several prompt engineering strategies that have demonstrated their efficiency in guiding LLMs. These strategies, as defined in the "Prompt Engineering Guide" (2024), include, but not limited to:

- **Zero-Shot Prompting:** It involves a direct instruction or question without any examples, and the performance of LLMs depends on its pre-trained knowledge and capability to understand natural language instructions. For example, a study by Kung et al. (2022) utilized a Query template, such as "Can you generate a writing about [a specific topic]?", to generate responses from ChatGPT.

- **Few-Shot Prompting:** It means using a few representative examples of task-answer pairs to help LLMs learn from these examples as models, hence improving their performance in generating similar responses.
- **Template-Based Prompting:** This strategy involves using pre-defined templates to guide LLMs in generating particular types of content, e.g., context template incorporating contextual information and example template incorporating high-quality examples to elicit more relevant responses (Kung et al., 2022; Liu et al., 2021; Wang et al., 2022).
- **Chain-of-Thought (CoT) Prompting:** It asks LLMs to provide reasoning steps, i.e. logical thinking before generating a final answer, making the process more transparent and providing a chance for adding modifications before creating the final response (Kojima et al., 2023).
- **Regenerate to Improve:** It aids in prompting LLMs to assess their initial output and accordingly generate an improved version based on discovered areas of improvement. This can be implemented through Reflect on Reflections (RoR) technique, where the LLM is instructed to reflect on its output and suggest modifications (Susnjak, 2022).

Hence, simplifying the prompt engineering process or providing pre-defined models, is vital for non-expert teachers to enable them to make use of generative AI in their classrooms by making prompt engineering more accessible and user-friendly.

4.2 Combining Linguistic Theories with Generative AI

The present study integrates three eminent and distinct linguistic theories that tackle language structure, function, and sociocultural variation into a prompt engineering model of parameters, providing a comprehensive framework for guiding the LLMs' generation of customized ESL reading materials, as described below:

4.2.1 TGG Theory

This theory is developed by Chomsky in 1957 and posits that sentences are generated by a set of rules to an underlying structure (Chomsky, 2024). It emphasizes, as stated in Chomsky (2024), Whong, (2007), and Yadav and Yadav (2020), the distinction between deep structures, representing meaning, and surface structures, representing the actual spoken or written forms of sentences, which could inform the design of AI-driven chatbots for generating ESL reading texts that are not only grammatically correct but also meaningful. Hence, in the context of AI-generated reading materials, TGG can be used to analyze and control the grammatical complexity and variation of such materials, ensuring that they align with the learners' proficiency level. This implies that an effective AI prompt for chatbots should incorporate mechanisms for generating deep structures that can be transformed into a variety of surface structures, reflecting the complexity and variability of natural language. However, the theory focuses on sentence-level generation and ESL learners need understanding of language at the discourse level (Alduais, 2015; Whong, 2007). Hence, combining TGG with other language theories would help LLMs produce reading texts that are both grammatically correct and pedagogically relevant for ESL learners, and the TGG's emphasis on deep and surface structures and adaptability of transformational rules to reflect linguistic diversity are pertinent features for effective AI prompts (see Section 6.1.2).

4.2.2 SFL Theory

As described by Halliday and Matthiessen (2014), SFL regards language as a social semiotic system; that is, language is used to create meaning in different contexts for various purposes. Hence, teachers, as revealed in Alshalan and Alyousef (2020) and Bartlett and O'Grady (2017), can use SFL to ensure that the utilized ESL reading materials are functionally relevant and contextually appropriate for diverse communicative situations, which, as proposed by the researcher, works in the same vein with texts generated by AI. Moreover, the SFL theory's focus on the metafunctions of language, namely, ideational, interpersonal, and textual, provides a prosperous framework for devising prompts for AI chatbots to generate ESL reading texts. In Alshalan and Alyousef's (2020) and Bartlett and O'Grady's (2017) description of such functions, the ideational function revolves around the content and experience conveyed by the text, which can help ensure that the generated texts are thematically relevant and engaging for learners. The interpersonal function deals with social relations and roles, which can be used to create texts that mimic real-life communicative situations, hence fostering interactive learning. The textual function refers to the organization of information in the text, which is crucial for coherence and cohesion, ensuring comprehension. As AI chatbots have been remarkably shown to simulate human collective behavior, forming communities around common language use (Wu et al., 2024), this suggests that such AI bots could possibly employ SFL features (see Section 6.1.2) to generate ESL texts that satisfy not only the learners' linguistic level but also their cultural and social contexts.

4.2.3 GEs Theory

According to Alasmari (2018) and Jenkins (2000), GEs theory recognizes that English usage is diverse and pluralistic around the world, so it legitimates different forms of English and the sociolinguistic realities of English as an international language. The theory has been popularized through Jenkins' (2000) work which underlies the comprehension of English as a global language. GEs endorses an inclusive approach to language teaching and learning and embraces the diversity of English in various cultural contexts. However, there is often resistance to teaching with local varieties of English because they are regarded as unequal to native speakers' varieties, such as those found in Saudi Arabia or Hong Kong besides other localized versions of English (Alasmari, 2018). Hence, they can help ensure that AI-generated ESL materials are more culturally responsive by reflecting how English is used globally and by meeting learners' needs and interests from different cultures. This can be achieved by incorporating aspects such as linguistic diversity, non-native norms acceptance, and intercultural communication competence into generative AI prompts (see Section 6.1.2).

As explained above, by combining TGG's focus on grammatical accuracy with SFL's emphasis on functional relevance and GEs' advocacy of diversity, the current research seeks to integrate these three language theories with generative AI to offer a comprehensive approach to producing ESL reading materials that are linguistically accurate, communicatively meaningful, and culturally sensitive. Such reading materials can be further enriched by AI's ability to adjust the texts' level of difficulty based on individual needs (See Section 6.2.1.3 on CEFR). By personalizing the learning experience, AI can help ESL learners focus on areas where they need additional support and save teachers much time and effort in searching for reading materials that address the various needs of their learners.

5. Review of Literature

5.1 AI's (LLMs) Role in Education & (ESL) Language Learning

Using AI in education, particularly LLMs to perform several educational tasks has recently gained significant momentum (Li et al., 2023). Past research focused on simple tasks, such as generating educational materials assessment questions of short-answer and multiple-choice types, producing adaptive feedback systems, and creating chatbots for engaging in human-like conversations with learners (Cavalcanti et al., 2021; Kurdi et al., 2020; Wollny et al., 2021). The use of more advanced LLMs like GPT-3, as seen in Brown et al. (2020), has helped perform more complex educational tasks. More recent studies examining the capabilities of ChatGPT, which is a chatbot based on an advanced LLM, have spotlighted its proficiency in successfully completing standardized medical examinations (Gilson et al., 2023; Huh, 2023), solving complex reasoning questions in pathology and microbiology (Das et al., 2023; Sinha et al., 2023), and performing comparably to students on law school exams (Choi et al., 2023). These findings along with Li's et al. (2023), which also focused on ChatGPT's ability to generate reflective writing and its implications for education, have underscored the ability of LLMs to handle complex cognitive educational tasks, knowledge retrieval, and text generation. This, in turn, stresses LLMs' capability to revolutionize educational assessment and personalized learning experiences.

Since chatbots have been recognized as AI tools capable of providing personalized language learning practices, instant feedback, and a low-pressure educational environment, particularly in ESL education, they have recently been the focus of research in this field (Fryer et al., 2020; Jeon, 2022; Kim et al., 2022). In fact, text-based chatbots, which are basically rule-based systems with limited conversational abilities, were the central theme of earlier studies. For instance, Google Assistant was investigated by Kim (2018), as an example of chatbots, for vocabulary learning among Korean English as a Foreign Language (EFL) learners, and Kim highlighted the benefits of such chatbots in interaction, and accordingly in augmenting language acquisition. Furthermore, Fryer et al. (2019) emphasized the importance of building chatbots that align with educational principles; by focusing on chatbot learning partners, they found a positive correlation between learning experiences, learner's interest, and competence. Subsequent research took into consideration advancements in speech recognition and synthesis technologies and explored the use of voice-enabled chatbots in ESL education. One of these studies was conducted by Han (2020) which demonstrated the positive effects of such AI voice chatbots, like Alexa, on Korean EFL students' conversational competence and affective domains, signifying that such chatbots could improve pronunciation, fluency, and learner’s motivation. Task-oriented chatbots designed for language learning tasks, moreover, displayed promising results in facilitating extended conversations and promoting problem-solving skills (Fryer et al., 2020; Kim et al., 2022).

Advanced LLMs like ChatGPT, in more recent studies, have been explored in the field of language learning, regarding their use in generating dialogue. Young and Shishido (2023), for example, focused on creating reference dialogues for a voice-based chatbot designed for ESL learners. Their study advocated that ChatGPT could produce dialogues for different levels of language proficiency, which, in turn, underpins its ability for creating engaging and relevant language learning materials. Similarly, Ochieng (2023) found out that LLMs could aid in guided reading with their efficiency in question-generation, hence working as supporting tools in language education. In the same vein, Woo et al. (2023) examined the use of LLMs in writing tasks via SOTA chatbots and the EFL students' interaction with them through prompts and underlined their success in such practical use in language learning. However, further research is needed to explore specific applications and pedagogical implications related to utilizing such LLMs in ESL/EFL teaching practices.

5.2 Prompt Engineering in Language Education & ESL

Prompt engineering is fundamental for effectively using LLMs, but it has not received enough attention in language education research, particularly in the ESL education. Clarisó and Cabot (2023) proposed a model-driven approach to prompt engineering in an attempt to simplify such process. Their study is not on language education; however, it underscores the challenges of prompt engineering and the need for tools and methodologies to facilitate the process. Studies on prompt engineering in language education have centered on developing strategies to enhance LLMs' ability to produce learning materials of high-quality. To exemplify, Zhao et al. (2022) investigated the use of BART summarization model to generate educational questions for children's storybooks. The research revealed the capability of prompting LLMs for question generation despite focusing on a restricted dataset and a narrow task. Ochieng's (2023), similarly, assessed the quality and variety of questions generated by LLMs and emphasized how important prompting strategies for achieving exact learning outcomes.

Another research that examined prompt engineering for educational purposes looked at the generation of content in general domains. For example, Wang et al. (2022) concentrated on generating human-like educational questions using different prompting techniques for LLMs. However, such studies did not address the specific challenges in prompting LLMs to produce tailored ESL reading materials. In another recent study by Woo et al. (2023), it specifically sought to find out how EFL students used prompt engineering when they interacted with chatbots during a writing task. The findings of the research displayed that the students used diverse methods while drafting their prompts, which indicated their need to improve their understanding and knowledge of prompt engineering so as to use LLMs more effectively in language learning. Thus, in fact, the lack of research on prompt engineering in ESL reading text generation implies that a more systematic and linguistically informed approach is needed, considering the various linguistic needs of learners and the pedagogical objectives of instructors alike.

5.3 Application of TGG, SFL, & GEs in Generative AI (LLMs) Systems

The integration of linguistic theories with generative AI is a significant area of research. However, the studies in this concern are limited but promising, underlining both benefits and challenges. Ding et al. (2023) examined the use of GPT-3 (an LLM) for data annotation in NLP tasks and highlighted the LLM's capability of language data analysis, which represents a vital aspect of integrating linguistic theories into AI applications. Labruna et al. (2023), while not employing language theories, used dialogue annotation schemas which relied on linguistic principles to evaluate ChatGPT's annotation capabilities. They unveiled, accordingly, the possibility of incorporating linguistic knowledge into AI systems for performing language tasks. Despite this potential, challenges remain in effectively integrating language theories into practical AI applications. The complexity of the linguistic frameworks requires careful adaptation for AI system design. Further research is needed to develop methodologies for incorporating linguistic principles into prompt engineering and evaluating the impact of such integration on the quality and effectiveness of AI-generated language learning materials.

To the researcher's best knowledge, no previous studies have tackled the use or integration of the TGG, SFL, and GEs theories in engineering prompts that enhance the LLMs' capability in generating customized ESL reading texts, and since the aforementioned theories help understand and analyze language in various contexts, the current research attempts to examine such innovative area. To point out, the current research, focusing on developing a linguistically informed, predefined prompt engineering parameters model for generating customized ESL reading texts, addresses gaps identified in the existing research. First, none of the past studies explicitly focused on the specific challenges and requirements of generating customized ESL reading materials. While some explored chatbot applications in language learning, like Young and Shishido (2023), they did not tackle the intricacies of adapting generative AI for creating diverse and pedagogically sound reading texts. The present study handles this gap by developing a model to enhance the AI's generation of ESL reading materials addressing the linguistic needs of learners and the pedagogical goals of educators. Second, whereas some studies mentioned the importance of prompt engineering, they often treated it as a purely technical process, without explicitly incorporating linguistic principles. For instance, Clarisó and Cabot (2023) introduced a domain-specific language for platform-independent prompts but did not explore the linguistic nuances of the prompts themselves. In the present research, three prominent linguistic theories, TGG, SFL, and GEs, are systematically integrated into the prompt engineering developed model. This integration ensures, to a considerable extent, that AI-generated texts are not only grammatically accurate but also functionally relevant, contextually appropriate, and culturally sensitive, addressing the multilayered nature of language. Besides, many studies highlighted the trial-and-error process involved in prompt engineering, emphasizing the challenges faced by non-experts (Woo et al., 2023). None of the studies proposed a predefined model to simplify this process, particularly for educators who may not have extensive expertise in AI or prompt engineering. The current study develops a user-friendly model of prompt parameters, specifically designed for educators to easily craft effective AI prompts for ESL texts generation. This model reduces the need for iterative experimentation with prompting, making AI more accessible for ESL teaching practices. By addressing these gaps, the present study aims to contribute to a deeper understanding of how AI can be effectively leveraged to enhance ESL education.

6. Methodology

6.1 Developing a Linguistically Informed Predefined Model of Prompting Parameters

The developed model incorporates two types of parameters, as explained below:

6.1.1 Basic Generative AI Prompting Parameters

The quality of the LLM's responses is based on the quantity of information you provide in the prompt and the operative construction of the prompt, as stated in Atlas (2023), "Prompt Engineering Guide" (2024), Reynolds and McDonell (2021), and Woo et al. (2023). Hence, the present study suggests, as a fundamental part of the developed model, that all or most of the following basic elements of prompt engineering, based on the required task, as illustrated in the aforementioned studies**,** be available in crafting effective AI prompts of one-shot instruction/zero-shot prompting technique for generating customized ESL reading texts that satisfy the teacher(s)' learning objectives and suit the language level(s) of their learners:

a. **Action/Task:** Specify the desired action using clear instructions or questions, e.g., write, explain, research, plan, can you translate, would you compare, etc.
b. **Input Data:** Define the topic or question the LLM should address and find a response for, e.g., highlight in digits *how far Mars is from Earth*.
c. **Context:** Describe clearly and in detail the context (i.e. external or additional information and relevant background) related to the topic, as well as the goal or intent, e.g. scenario, task details, date, etc. This helps the LLM understand the current status and generate a relevant response.
d. **Output Format:** Specify the desired format or structure for the generated response, e.g. a list, essay, bullet points, dialogue, table, I accomplished X by measure Y that resulted in Z, etc.
e. **Tone:** Specify the desired tone for the generated response, e.g., formal, casual, serious, friendly, optimistic, humorous, professional, scientific, persuasive, informative, etc.

f. **Persona:** Define a specific persona for the LLM to adopt, i.e. role and/or (writing) style. Instruct the LLM on how to behave, its intent and identity. Examples of assumed roles for the LLM are, among others, a lawyer, a teacher of English, literary editor, senior product marketing manager at Apple, or an experienced physical therapist with over 20 years of experience. Define a (writing) style for the generated response, e.g. speak in President Barack Obama's voice, write in the style of master storyteller Ernest Hemingway, act like billionaire entrepreneur Elon Musk, etc.

g. **Audience:** Specify the target audience and their level of expertise (e.g. educational, language, etc.) and tailor the devised prompt accordingly. For instance, write for 5th graders, provide an answer suitable for a university-level economics class, or explain it as if you're talking to a 12-year-old.

h. **Output Length:** Set a desired length limit for the generated response. This helps avoid long or irrelevant responses. For example, you might request a 500-word text, a concise paragraph, etc.

i. **Source/Reference:** Recommend reference materials or relevant sources to guide the LLM's output so that the response contains targeted information, e.g., a book, a URL, pdf file, etc.

j. **Example/Exemplar:** Provide specific examples for the language model to review before generating the response. Queries coupled with exemplars (i.e. demonstrations) help the LLMs generate similar ones, e.g., a Twitter post, an article, an email, a story, a report, etc.

6.1.2 Language Theories-Based Prompting Parameters

As for the three employed language theories in the present paper, their identified parameters in the developed model are not to be all present in the engineered prompt; rather, the parameters are to be selected by the teacher/user as required. However, it is recommended in the present study that parameters from each theory be integrated into the devised prompt so that the generated ESL reading text represents and meets a comprehensive approach to language teaching, i.e. contextually appropriate, functionally relevant, and culturally inclusive, adhering to the principles of the three linguistic theories. Some of these theories-based prompt parameters may interfere with the aforementioned prompt engineering basic parameters, but from linguistic and/or language teaching perspectives since generative AI plays a crucial role in linguistic analysis and production, which, in turn, contributes to ensuring the effectiveness of the formulated prompt, and thus, the generated result(s)

a. TGG-Based Prompting Parameters

1. **Structure Identification:** Specify the syntactic structure of the expected output, such as the word order, the number and type of clauses, the tense and aspect, etc. This helps create reading texts that focus on particular grammatical topics.
2. **Variation in Syntactic Complexity:** Adjust the syntactic complexity of the required output based on the learners' language level, e.g. from simple sentences for beginners to intricate sentences for advanced learners, aiding in the scaffolding of reading materials.
3. **Rule Application:** Develop prompts that request from the LLM to apply certain transformational rules and syntactic changes, like, for example, moving auxiliary verbs for question formation, which can be used to create reading texts that reinforce these rules.

b. SFL-Based Prompting Parameters

4. **Communicative Function Identification:** Define the communicative intent of the output text as represented in its genre (purpose, structure, style, and audience), e.g. narrative, expository, descriptive, etc. This can help produce texts that are appropriate and effective for different contexts and goals, as well as for the learners' ability to evaluate them.
5. **Contextual Relevance & Register Focus:** Specify the social context of the generated text, such as the field (subject matter), the tenor (social roles and relationships between the participants), and the mode (channel of communication). This makes it easy for the AI to generate an output that is suitable and relevant for the communicative situation and relationship between speakers since registers vary based on such factors, e.g. formal, informal, technical, etc.
6. **Language Metafunctions Inclusion:** Specify the ideational (conveying experiences), interpersonal (enacting social interactions), and/or textual (organizing language) metafunctions of

language so that the AI-generated reading content is aligned accordingly.

c. GEs-Based Prompting Parameters

7. **Diversity Emphasis & Cultural Sensitivity:** Define the sociocultural context features of the generated text such as the variety of English (dialects & accents) and the culture (e.g. cultural references & idioms from different communities). This can help AI produce texts that reflect the diversity and dynamism of English as a global language and respond to learners' needs to appreciate different cultures and perspectives.
8. **Authenticity in Language Use:** Request the generation of reading material showcasing language use from various English-speaking communities, moving beyond standard forms often found in textbooks, like incorporating colloquialisms and uncommon regional variations. This would prepare learners for real-world language usage.
9. **Interactive Scenarios Inclusion:** Request the creation of situations in the AI generated reading content that simulate everyday interactions across different English-speaking cultures, which helps enhance learners' communicative competence and displays the richness of English, providing exposure to a wide range of linguistic expressions.

6.2 Data Collection & Analysis

The process of data collection and analysis is divided into two main stages, as elucidate below, and a mixed method approach combining quantitative and qualitative data analysis is employed:

6.2.1 Generating ESL Reading Texts

To start with, the process of generating the reading texts via the developed model undergoes the following steps:

6.2.1.1 Selecting the LLM: Microsoft Copilot Chatbot

Microsoft Copilot (formerly Bing Chat) is an AI-powered search assistant designed to revolutionize how users interact with the web (Microsoft, 2024b). Going beyond traditional search engines, Copilot engages in a conversational manner, providing comprehensive answers, generating text and images, and assisting with various tasks. Copilot is accessible through multiple platforms: web-based Copilot accessed directly at Copilot, Microsoft Edge browser sidebar for quick access, and mobile application available on iOS and android devices for on-the-go use. Using Copilot involves inputting a question or command in the "Ask me anything..." box and engaging in a conversational back-and-forth to refine results (Microsoft, 2024b). It can answer questions, summarize articles, provide product comparisons, generate workout plans, create itineraries, write stories, find cheap airline tickets, and much more. Copilot also supports visual search using images and can generate unique images using its built-in AI-powered Designer tool.

Copilot is selected in the present study since, according to (Microsoft, 2024a), it leverages a powerful combination of AI technologies: OpenAI's ChatGPT-4, Microsoft's Prometheus Model, and DALL-E 3. The first is customized for search, optimized for speed, and adept at generating creative text formats, thus contributing to creating ESL reading texts of various types benefiting from increasing the temperature value for creative tasks since it increases the weight of all possible tokens and makes use of the creative properties of the paid GPT-4 but for free. The second works in conjunction with ChatGPT-4 to enhance the timeliness, relevance, and safety of search responses, addressing a limitation of relying solely on pre-trained data in free ChatGPT and other LLMs. The integration of DALL-E 3 also empowers Copilot to generate images, expanding its capabilities beyond text-based content creation. The third is OpenAI's text-to-image neural network, enabling Copilot to generate unique images based on user prompts, which could help in generating images that suit the generated reading texts upon the teacher’s request, however, this is not the interest of the present study. Last but not least, Copilot is also accessible via multi-platforms, as highlighted above, which ensures a seamless and convenient user experience.

6.2.1.2 Selecting the Prompt Engineering Technique/Strategy

To simplify the prompt design process for language teachers, the predefined model of prompt parameters is aimed at lessening the iterative process of designing prompts by language teachers who are not expert enough in prompt engineering. Hence, each prompt is to be of a one-shot approach. The prompting technique used is the zero-shot strategy/setting (direct instruction or question without exemplars), as described in Kojima et.al. (2023), Kung, et.al. (2022), and the “Prompt Engineering Guide” (2024); the zero-shot setting is of the Generate by Template type of prompting where the Query Template is the one utilized since it accords with the suggested model in the present paper. In the Query Template, the prompt, according to Kung, et.al. (2022), contains just enough information to generate an appropriate reading text, based on the developed model parameters in the current study. However, the source/reference and example/exemplar parameters in my model are not used in the present study since the zero-shot prompting technique is employed. Both parameters require other prompting techniques, on top of which is the Example Template which includes high-quality examples to elicit more focused and relevant responses (Kung et al., 2022; Liu et al., 2021; Wang et al., 2022), and the present study opts for the simplest and most direct prompting strategy for teachers who are not, by necessity, experienced enough in such area. It also seeks to make use of the employed LLM's capability to create various reading materials based on its reliance of pre-trained data and its relevance in retrieving information from the web. However, both parameters are included in the developed model since they are among the basic effective prompting elements.

6.2.1.3 Employing the CEFR Rubric for Defining Language Levels

The CEFR is utilized in the study to determine the language level of the AI-generated reading material according to the target ESL learners so that it best suits their linguistic needs. The researcher attempts to highlight how LLMs are advanced enough to identify such an international rubric and create reading texts accordingly by only feeding its name and the required language level in the prompt. In contrast, Young and Shishido (2023) used such rubric outside the prompting process; they used it to determine the target audience best suited for the AI-generated dialogue materials after the generation process, and based on the readability scores, they concluded that the ChatGPT-generated dialogues were most appropriate for students at the CEFR A2 (elementary) proficiency level.

The choice of CEFR stems from its widespread adoption as a standardized framework for describing language proficiency levels, allowing for consistent assessment and comparison across different learners and educational contexts. By aligning the AI-generated materials with specific CEFR levels, the researcher aims to ensure that the content is appropriate for the target learners, maximizing its effectiveness in supporting language learning. Here are the CEFR six language levels, as summarized by Council of Europe (2001) in Table 1 below:

Table 1: CEFR Language Levels

Level	Description
A1	**Beginner:** Can understand and use very basic phrases, introduce themselves, and ask and answer simple questions about personal details.
A2	**Elementary:** Can understand and use sentences and frequently used expressions related to areas of most immediate relevance (e.g., personal information, shopping, local geography, employment). Can communicate in simple and routine tasks.
B1	**Intermediate:** Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise while traveling in an area where the language is spoken.
B2	**Upper Intermediate:** Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in their field of specialization. Can interact with a degree of fluency and spontaneity.
C1	**Advanced:** Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express ideas fluently and spontaneously without much obvious searching for expressions.
C2	**Proficient:** Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation.

Since the CEFR grades each language skill on a six-level scale of language proficiency that are grouped into three broader levels, as underscored by Council of Europe (2001, p. 24): A1-A2 (Basic User), B1-B2 (Independent User), and C1-C2 (Proficient User), the attempted generated reading texts are six, one representing each language level; that is, the developed model is tested in creating six different reading texts, a text for each language level in order to see how far the model succeeds in generating a customized reading material that satisfies a particular language level, achieves readability and incorporates the linguistic elements required by the teacher to address the language needs of their ESL learners. Furthermore, the aimed reading texts are of different types since the researcher aims to evaluate the model's versatility. The six reading texts are described in Tables 2 and 3, where the former displays the basic elements of the targeted ESL reading texts and the latter describes the main linguistic elements incorporated in such texts from the three language theories employed in the study, as expounded below:

Table 2: Text Basic Elements Description

Text Type & Format	Topic & Title	Length	Tone	Audience	(Writing) Style
Text 1 conversational lyric.	Three ancient myths: Greek, Roman, and Egyptian where a few characters are conversing with each other. "Mythic Dreams: From Olympus to the Nile"	about 200 words	gothic and conversational	A1 ESL learners	A trio
Text 2 narrative description	description of Abu Simbel temple and its story including comparisons between the temple and other similar sightseeing from different countries. "Guardians of Time: The Wonders of Abu Simbel"	350 words	exciting and adventurous	A2 ESL learners of various countries on a tour to Abu Simbel temple in Aswan, Egypt/language level similar to 6th graders'	A tourist guide
Text 3 informative article	The impact of excessive mobile phone usage on children "Breaking News: The Mobile Dilemma – How Phones Shape Our Kids' World"	300 words	Friendly	B1 Learners/ 7th graders	A TV announcer
Text 4 opinion article	Why the American society should unite to fight against covid 19. "Uniting Against COVID-19: A Call to Action for ESL Learners in the USA"	around 600 words	enthusiastic & persuasive	ESL learners living in the USA and belonging to diverse cultural backgrounds/B2 (10th graders)	President Obama's style in his speeches
Text 5 poetic dialogue in a play scene	A medical diagnosis "The Dermatologist's Chamber"	Twenty-four turns between two characters: dermatologist & patient	empathetic	C1 learners/12th graders	Shakespearean style in his tragedies
Text 6 critical film review	The Last Samurai" film released in 2003, starring Tom Cruise and directed by Edward Zwick "The Last Samurai: A Profound Cinematic Odyssey"	about 1000 words	praise & constructive criticism & professional language in the field	C2 ESL learners/ senior college students registered in a writing course on film reviews	The complex style of T.S. Eliot

Table 3: Text Linguistic Elements Description

Text	TGG	SFL	GE
Text 1	**Structure Identification:** specifying the syntactic simplicity required for the output; learners can understand familiar names, words, and very simple sentences. **Rule Application:** asking to manipulate sentence structure, demonstrating different syntactic constructions; play with themes and rhemes changing their slots in the sentences.	**Communicative Function Identification:** specifying the interactive nature of the text, which involves multiple speakers; a few characters are conversing with each other. **Language Metafunctions Inclusion:** emphasizing the textual function, making structural elements of language explicit; such themes and rhemes must be written in bold. **Contextual Relevance & Register Focus:** specifying the social context (the gothic genre) and register (the conversational exchange between characters from different myths), ensuring the content is suitable for the communicative situation and the relationships between the characters.	**Authenticity in Language Use:** referencing real-world cultural myths, the lyric reflects authentic use of language that acknowledges the diversity of English influences; 3 ancient myths: Greek, Roman, and Egyptian. **Interactive Scenarios Inclusion:** the dialogue between mythological characters creates an interactive scenario that can help learners navigate different cultural contexts; characters are conversing with each other.
Text 2	**Rule Application:** requiring direct and indirect speech involves applying transformational rules for sentence structure.	**Communicative Function Identification:** specifying the communicative function (engaging visitors) and context (comparisons with other sightseeing), including comparisons between Abu Simbel temple and other similar sightseeing from different countries to engage the visitors more into the tour by exposing them to various experiences. **Language Metafunctions Inclusion:** direct and indirect speech convey experiences and interactions and incorporating comparisons with other sightseeing from different countries represents various experiences, modifiers (adjectives, adverbs) enhance interpersonal meaning, and comparative and superlative forms organize language.	**Diversity Emphasis & Cultural Sensitivity:** considering the language level and cultural diversity of ESL learners; adapting your writing to be comprehensible to ESL learners of various countries.

Text	TGG	SFL	GE
Text 3	**Structure Identification:** specifying the tenses to be used (present simple, present continuous, and present perfect). **Variation in Syntactic Complexity:** changing tenses complexity (simple, continuous, perfect).	**Communicative Function Identification:** specifying the communicative intent (informative article). **Contextual Relevance & Register Focus:** reflecting on social implications in different communities & the friendly tone throughout the article.	**Diversity Emphasis & Cultural Sensitivity:** emphasizing relatability to diverse readership and cultural contexts.
Text 4	**Rule Application:** instructing the AI to apply transformational rules (conditional clauses) in the generated speech; the speech must showcase examples of rhetorical questions and conditional clauses in English: first, second, and third conditional sentences.	**Communicative Function Identification, Contextual Relevance & Register Focus:** identifying the communicative function (persuasion), emphasizing the need for relevant examples (contextual relevance), and specifying the audience (ESL learners from diverse cultural backgrounds), supporting your argument with examples from different regions around the world to be able to properly persuade such learners.	**Diversity Emphasis & Cultural Sensitivity, Authenticity in Language Use:** emphasizing cultural diversity by requesting cultural references from different regions & encouraging authenticity in language use by incorporating different English varieties.
Text 5	**Rule Application:** requiring active-to-passive voice transformations involves applying specific transformational rules, reinforcing grammar concepts.	**Contextual Relevance & Register Focus:** specifying the characters (dermatologist and patient), which informs the AI about the social roles and relationship between participants. This falls under the register focus parameter, ensuring the generated dialogue suits the context of a medical consultation. Moreover, the requirement to use technical medical terms involves the adaptation of language based on subject matter, social roles, and communication channel. **Communicative Function Identification:** specifying an	**Diversity Emphasis & Cultural Sensitivity:** incorporating British and Indian English varieties reflects the diversity of English, acknowledging different linguistic features

Text	TGG	SFL	GE
		empathetic tone and the purpose (explaining a diagnosis), indicating the intent behind the dialogue. **Language Metafunctions Inclusion:** using modality and evaluative language (expressing probability and obligation)	
Text 6	**Structure Identification:** requiring complex syntactic structures (relative and noun Clauses)	**Contextual Relevance & Register Focus:** specifying the social context (senior college students) and the register (professional language) **Language Metafunctions Inclusion**: specifying the textual organization (covering specific aspects: plot summary, direction, screenplay, acting, cinematography, editing, soundtrack and sound design, production design, themes and messages, and overall impact)	**Diversity Emphasis & Cultural Sensitivity:** incorporating language varieties (American and Japanese English) reflects global English diversity and cultural context.

6.2.1.4 The Developed Model Processing Steps

a. **Start:** The process begins with defining the teacher's learning objectives and the learners' language level. The CEFR, only by name, along with the learners' language level are incorporated into the prompt engineered. The teacher's learning objectives are fed into the prompt via the parameters of the developed model.
b. **Basic Effective AI Prompting Parameters:** These parameters form the foundation of the prompt and are largely universal for effective output (See Section 6.1.1).
c. **Language Theories-Based Prompt Parameters:** Based on the learning objectives and learners' language level, the teacher selects one or more parameters from the three language theories to integrate into the prompts, which allows for a tailored comprehensive approach to language instruction (See Section 6.1.2).
d. **Combine Elements:** The basic prompts elements are combined with the selected theory-based elements to create a comprehensive and effective prompt.
e. **Generate Reading Material:** The final devised prompt is fed into the LLM to generate the required response.

f. **Evaluate Output:** The generated reading material is evaluated based on its relevance to the teacher's requirements, i.e. its inclusion of the prompt elements, hence effectiveness in meeting the learning objectives and learners' language level.
g. **Adjust Prompt (If Needed):** If the output does not meet the requirements, the prompt can be adjusted and the process repeats.
h. **End:** The process concludes with a satisfactory piece of ESL reading material generated by the LLM.

Figure 1 below illustrates the model processing:

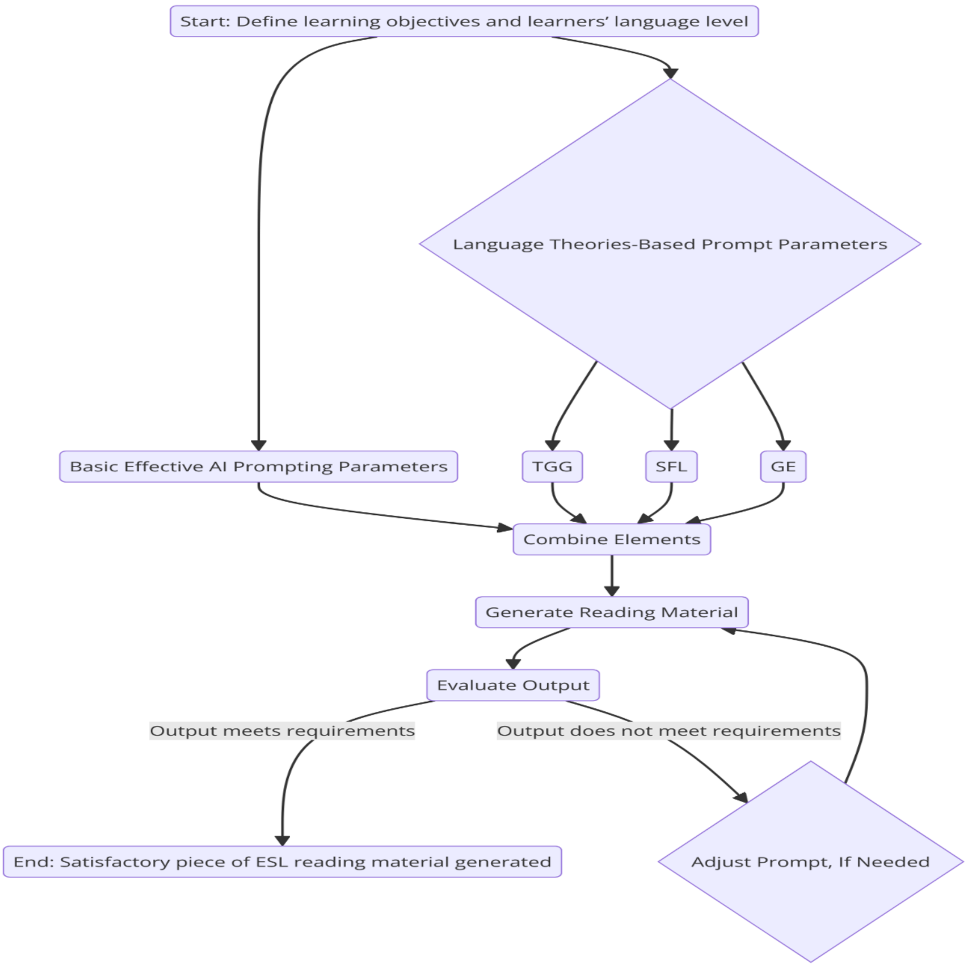


Figure 1: Developed Model Processing Steps

6.2.2 Evaluating the Effectiveness of the Developed Model

6.2.2.1 Evaluating the Generated Texts via Readability Assessment

The researcher uses three readability metrics to evaluate the language level of each generated reading text to see if it aligns with the CEFR language level intended by the teacher for the text and fed into the engineered prompt. Such quantitative analysis, in turn, evaluates the LLM's ability to understand the instructions in the prompt in this respect and the CEFR to create a text accordingly. The readability metrics employed are Flesch-Kincaid Reading Ease (FKRE), Flesch-Kincaid Grade Level (FKGL), and Gunning Fog Index (GFI), and the calculation process is done automatically via a free online tool named charactercalculator.com

The FKRE, developed by Rudolf Flesch in 1948, measures the ease of understanding a text. It primarily considers sentence length (average number of words per sentence) and syllable count (average number of syllables per word) to determine readability ("Flesch Kincaid," 2024). The score is interpreted as seen in Table 4 below (Taken from "Online Readability," 2024):

Table 4: FKRE Score Interpretation

Score	School Level	Comprehension/Description
90.0-100.0	5th Grade	Very easy to read
80.0-90.0	6th Grade	Easy to read
70.0-80.0	7th Grade	Fairly easy to read
60.0-70.0	8th & 9th Grades	Plain/ Conversational English
50.0-60.0	10th, 11th & 12th Grades	Fairly difficult to read
30.0-50.0	College	Difficult to read
10.0-30.0	College Graduate	Very difficult to read
0.0-10.0	Professional	Extremely difficult to read

The text is fairly easy to read and understand for an average adult if the FKRE score is 70 or above. This metric is valuable for assessing the general readability of a text and useful for determining some material is suitable for specific audiences. The formula for calculating the Flesch Reading Ease score is:

206.835 - 1.015 × (total words ÷ total sentences) - 84.6 × (total syllables ÷ total words)

Rudolf Flesch also developed the FKGL to measure the readability of a text via estimating the U.S. school grade level required to understand it ("Flesch Kincaid," 2024). Similar to the FKRE, it depends on sentence length and syllable count to evaluate readability. Unlike the FKRE, the FKGL lower scores indicate easier readability. The score is interpreted as shown in Table 5 (Taken from "Online Readability," 2024):

Table 5: FKGL Score Interpretation

Score	School Level	Comprehension/Description
5.0-5.9	5th Grade	Very easy to read
6.0-6.9	6th Grade	Easy to read
7.0-7.9	7th Grade	Fairly easy to read
8.0-9.9	8th & 9th Grades	Plain/ Conversational English
10.0-12.9	10th, 11th & 12th Grades	Fairly difficult to read
13.0-15.9	College	Difficult to read
16.0-17.9	College Graduate	Very difficult to read
18.0+	Professional	Extremely difficult to read

The FKGL is particularly useful in educational contexts, allowing educators to match reading materials to students' reading abilities. It can also be helpful for writers aiming to target their content to specific audiences based on their estimated education levels. The formula for calculating the FKGL is:

0.39 × (total words ÷ total sentences) + 11.8 × (total syllables ÷ total words) - 15.59

The GFI, introduced by American businessman Robert Gunning in 1952, measures readability by considering sentence length and the percentage of complex words (words with three or more syllables; Lower scores indicate easier readability ("Gunning Fog," 2024). The score is interpreted as seen in Table 6 (Taken from "Online Readability," 2024):

Table 6: GFI Score Interpretation

FOG Score	School Level	Comprehension/Description
0-5	5th Grade and below	Very easy to read
6	6th Grade	Easy to read
7	7th Grade	Fairly easy to read
8	8th Grade	Plain/Conversational English
9-12	9th & 12th Grade	Fairly difficult to read
13-16	College	Difficult to read
17	College Graduate	Very difficult to read
18-20	Professional	Extremely difficult to read

The GFI is particularly useful for assessing the readability of technical and academic writing, which often contains longer sentences and more complex vocabulary. The formula for calculating the GFI for a passage of approximately 100 words is:

0.4 × ((total words ÷ total sentences) + 100 × total complex words ÷ total words)

Using multiple readability metrics, particularly those that rely on different criteria like sentence length, syllable count, and word complexity, provides a more comprehensive and accurate evaluation of a text's readability. In the present research, using this multi-metric approach enables a thorough assessment of the language level of each AI-generated ESL reading text and enhances the accuracy of determining whether the generated texts align with the intended CEFR levels and meet the specific linguistic needs of the target learners.

In the present research, the use of the charactercalculator.com tool for automatically calculating the FKRE, FKGL, and GFI scores is due to various merits. First, it is a user-friendly online tool for easily evaluating text readability; since calculations are conducted mechanically, they are more accurate and less tedious, saving researchers time and effort. Second, the tool calculates the scores of both FKRE and FKGL in an integrated manner, offering a comprehensive representation of a text readability. Third, it includes a GFI calculator, enabling the assessment of multiple readability metrics within the same platform. Besides, the tool also provides character, word, sentence, and syllable counts, and these additional features, in turn, can be valuable for further analyzing the length of the generated ESL reading texts, thus ensuring that the LLM has aligned with the length instruction incorporated by the researcher in the devised prompts. In summary, the charactercalculator.com tool's accessibility, additional text analysis features, and focus on readability assessment make it an ideal choice for evaluating the language level of the AI-generated ESL reading texts.

6.2.2.2 Evaluating the LLM's Understanding & Implementation of the Devised Prompts

The researcher aims to qualitatively analyze the generated texts to examine if Copilot has effectively incorporated the specified prompts elements in the created texts, hence, exploring the Chatbot's full understanding of the instructions in the prompts and production of the texts accordingly as well as the adaptability of the developed model via generating texts across different genres and registers, demonstrating its compliance with various pedagogical needs. Such qualitative analysis is conducted for each generated text in a table format of three columns: Prompt Element, Reading Text Content which displays where each prompt element exists in the text, and Example which highlights examples from the text that represent each prompt element. Every table of analysis is further supported by another table that contains an analysis of each component of the prompt and its alignment with the developed model parameters along with an explanation of such configuration in three columns named as follows: Prompt Element, Model Parameter, and Explanation. In Section 7.3 below, an example of the generated texts using the developed model processing in Section 6.2.1.4 is provided along with full quantitative and qualitative analyses of such example after a detailed explanation of the (quantitative) readability assessment of the six generated texts and the (qualitative) evaluation of Copilot's understanding and implementation of the engineered prompts for the six reading texts. The detailed example is meant to increase the credibility of the entire evaluation procedure of the model's effectiveness, and other examples of the generated reading texts along with their prompts, analyses, and raw readability metric data are provided in the Appendix as well to save space in Section 7 below.

7. Results

7.1 Readability Scores Analysis

The researcher analyzes the readability scores obtained for the six AI-generated ESL reading texts and compares them to the targeted CEFR levels defined in the prompts to see whether the texts are aligned accordingly (see Table 7 below):

Table 7: Multi-Metric Readability Scores Description of All Texts Via Charactercalculator.co

Text	CEFR Level	FKRES	FKGLS	GFIS	Reading Level	Reading Note	Chr.	Word	Sent.	Syll.
Text 1	A1 (Beginner)	85.88	4.29	6.31	6th grade	Easy to read	1688	218	18	280
Text 2	A2 (Elementary)	73.80	6.50	8.62	7-8th grades	Fairly easy to read/Plain English	1988	356	25	499
Text 3	B1 (Intermediate)	73.56	5.53	8.55	7-8th grades	Fairly easy to read/Plain English	3322	510	50	741
Text 4	B2 (Upper Intermediate)	58.61	8.62	11.09	10th to 12th grade	Fairly difficult to read	4411	726	51	1148
Text 5	C1 (Advanced)	44.40	8.72	10.87	College/10th grade	Difficult to read/ Fairly difficult to read	2597	400	60	736
Text 6	C2 (Proficient)	28.59	13.06	16.06	College graduate/College Senior	Very difficult to read/ Difficult to read	6927	975	64	1876

As seen in Table 7 above, the readability scores generally align with the targeted CEFR levels, exhibiting a trend of decreasing readability scores as the CEFR levels progress from A1 to C2. This suggests that the LLM effectively incorporated the CEFR framework into its text generation process. The number of words, sentences, and syllables also tends to increase with higher CEFR levels, reflecting greater linguistic complexity. In the three bar charts below, which show the FKRE, FKGL, and GFI scores distribution across all texts and where the x-axis shows the text number along with its CEFR level, while the y-axis represents the score values, as the CEFR level increases from A1 to C2, the FKRE scores generally decrease, indicating increasing difficulty, and the FKGL and GFI scores generally increase, also indicating increasing difficulty (see Figures 2, 3, & 4):

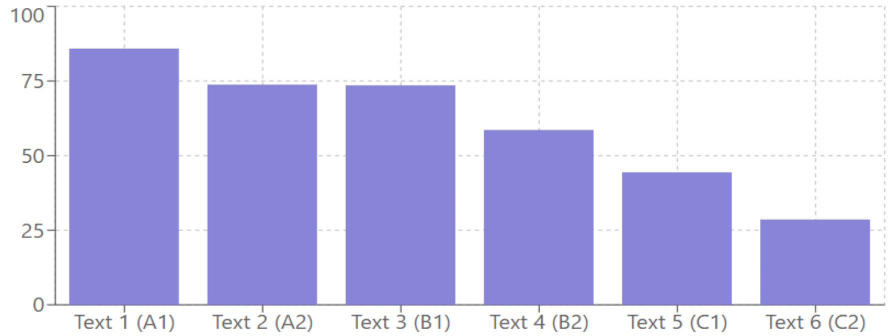


Figure 2: FKRE Scores Distribution Across All Texts

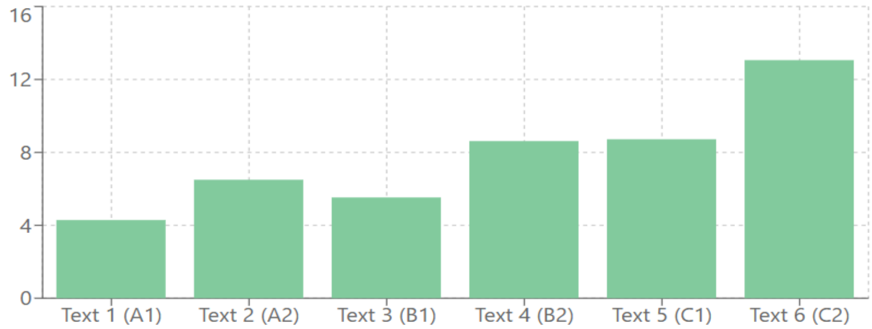


Figure 3: FKGL Scores Distribution Across All Text

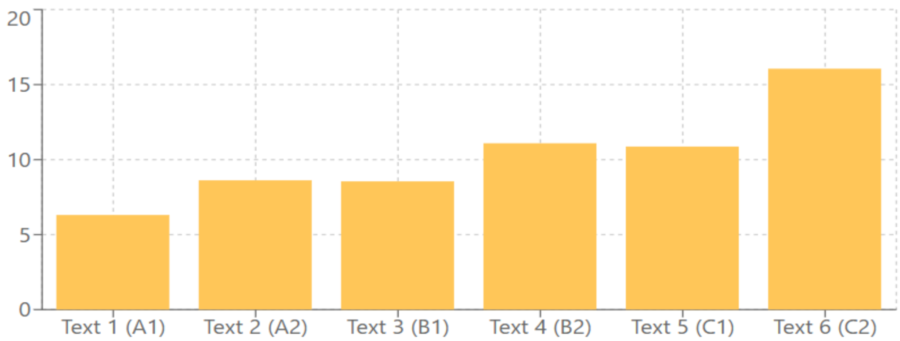


Figure 4: GFI Scores Distribution Across All Texts

The following notes are observed in Table 7:

- **Text 1 (A1):** The scores indicate that this text is easy to read, appropriate for the A1 (Beginner) level, representing the initial stage of English language learning.
- **Text 2 (A2):** The scores indicate that this text is fairly easy to read (plain English), suitable for the A2 (Elementary) level.
- **Text 3 (B1):** The readability scores indicate that this text is fairly easy to read (plain English), aligning with the B1 (Intermediate) level.
- **Text 4 (B2):** The scores suggest that this text is fairly difficult to read, which aligns well with the B2 (Upper Intermediate) level.
- **Text 5 (C1):** The readability scores indicate that this text is (fairly) difficult to read, consistent with the C1 (Advanced) level.
- **Text 6 (C2):** The scores indicate that this text is (very) difficult to read, aligning with the C2 (Proficient) level, which represents the highest level of English proficiency.

A note to be highlighted is that the parenthetical description of the readability scores in the above observations indicate that the CEFR levels are broader than the educational US school grades, i.e. each CEFR language level covers more than a US school grade since the former are six and the latter are twelve apart from college students and graduates, as underlined in Tables 4, 5, and 6 above. This ensures that the very slightly different readability scores descriptions of the same text, as provided by the utilized online calculator tool, is by no means significant.

As illustrated in Table 7 and Figures 2, 3, and 4, there is a strong agreement between the readability scores of each generated text and the intended CEFR level of each; however, there are some minor discrepancies that are worth noting. First, even though Text 2 targets the A2 (Elementary) level and Text 3 targets the B1 (Intermediate) level, Text 2 exhibits slightly higher readability scores in the FKGL and GFI, suggesting that Text 2 is slightly easier to read despite targeting a lower CEFR level, which is counterintuitive. This discrepancy could be attributed to factors like sentence structure, vocabulary choices, and the specific content of the texts, which readability metrics may not fully capture. To explicate, upon examining the data and text content as seen in Table 8 below, it is revealed that despite having simpler vocabulary, Text 2 has a significantly higher average sentence length (14.24 words) than Text 3 (10.2 words). This longer sentence length likely contributes to its marginally higher FKGL and GFI scores. While the overall vocabulary in Text 2 is simpler since it contains only 21 complex words compared to Text 3 (55 complex words), the higher proportion of complex words per sentence in Text 2 compared to Text 3 may also contribute to the discrepancy. Moreover, Text 3 is significantly longer than Text 2 (510 words vs. 356 words). The sheer length of Text 3 could contribute to its lower readability FKRE score, even if the language itself is not significantly more complex since longer texts can be more demanding for readers, regardless of their proficiency level.

Table 8: Discrepancy One

Text	Word	Sent.	Syll.	Complex Word	Avg. Sent. Length
Text 2 (A2)	356	25	499	21	14.24
Text 3 (B1)	510	50	741	55	10.2

The content and style of the texts might also play a role. Text 3, despite targeting a higher CEFR level, might address a more familiar topic or use a more straightforward style, making it easier to comprehend despite having more challenging vocabulary overall. To point out, Text 3's topic of excessive mobile phone usage among children is likely to be familiar to a wide range of readers, regardless of their English proficiency level. Most individuals have some experience with mobile phones and their impact on society, making the topic easily relatable. In contrast, the topic of Text 2, which is on the Abu Simbel temple, while interesting, may be less familiar to readers, particularly those at the A2 level. Understanding the historical and cultural context of the temple might require some background knowledge that elementary learners may not possess. Furthermore, the style of Text 3 is deliberately engaging, interactive and conversational, mimicking a friendly TV announcer addressing a young audience. The use of rhetorical questions, direct addresses to the reader, e.g., "Remember, young viewers...", and vivid imagery, e.g., "Phones, don't steal our sunshine!" creates a lively and accessible tone captivating for readers. Conversely, although the style of Text 2 is also narrative and engaging, it relies more on descriptive language and factual information. While the sentences are generally shorter, the density of information and the need for background knowledge could make it less immediately comprehensible for A2 learners. This comparison highlights that readability scores, while helpful, cannot fully account for the impact of content and style on comprehension. A text on a familiar topic presented in an engaging and straightforward style might be easier to understand than a text on a less familiar topic, even if the latter has a slightly lower readability score. Thus, upon evaluating the suitability of ESL reading materials, considering both the quantifiable aspects measured by readability metrics and the qualitative aspects of content and style is crucial.

Second, the GFI for Text 4 (B2) is slightly higher than that of Text 5 (C1), potentially indicating a higher proportion of complex words in Text 4. By analyzing data and content as seen in Table 9, it can be highlighted that Text 4 contains more complex words than Text 5 (101 vs. 65). Besides, Text 4 contains more domain-specific vocabulary related to the topic of COVID-19 and public health, even if those words are not considered more advanced in general English proficiency. Examples include "pandemic," "social distancing," and "vaccinated." Text 4 has longer sentences on average (14.24 words) than Text 5 (6.67 words). While longer sentences generally make a text more challenging, the higher FKRE score for Text 4 suggests that this factor is not significantly influencing its overall readability.

Table 9: Discrepancy Two

Text	Word	Sent.	Syll.	Complex Word	Avg. Sent. Length
Text 4 (B2)	726	51	1148	101	14.24
Text 5 (C1)	400	60	736	65	6.67

Third, Text 5's FKRE score (44.40) indicates it is "Difficult to read" and suits a college student's language level. However, the FKGL (8.72) and GFI (10.87) scores suggest a "Fairly difficult" level suitable for a 10th-grade student, which demonstrates inconsistency across metrics where the three metrics do not perfectly align in their assessment of a text. This signifies that each metric works on different aspects of readability. To explicate, the Shakespearean style employed in Text 5 includes archaic vocabulary and complex sentence structures, which suggestively impacts the FKRE score. However, the GFI focuses solely on complex words, so it may not capture the complex syntax. This strengthens the need for using multiple metrics to obtain a more accurate understanding of a text's readability.

In conclusion, the above discrepancies could be attributed to the fundamental limitations of readability metrics since they rely on simplified formulas and may not fully consider all angles of language complexity. They mainly focus on sentence length and syllable counts, but do not explain factors like idiomatic expressions, cultural references, or the overall coherence and flow of the text. Accordingly, careful interpretation is necessary upon utilizing readability metrics, and they should be considered alongside other factors, such as sentence structure, vocabulary choice, text length, and text coherence despite the fact that they represent an important starting point for assessing language complexity. In addition, it is crucial to remember that readability formulas are just tools; they cannot perfectly capture the multifaceted nature of language proficiency. In fact, the analysis of readability scores provides initial evidence that the LLM has effectively understood and incorporated the targeted CEFR levels into its text generation process, demonstrating the LLM's ability to fit language complexity to specific learner proficiencies.

7.2 Analysis of the AI’s Performance in Understanding & Executing Prompts

In this section, the researcher analyzes the extent to which the generated texts incorporate the prompt parameters derived from the three language theories (TGG, SFL, and GEs) and basic prompt engineering elements as specified in the devised prompt for each text. Across the six generated texts, the AI demonstrates a strong ability to understand and execute the prompts, integrating the stated prompt components into the generated content. This alignment is evident in the consistent presence of the targeted linguistic features, stylistic elements, and text types requested in the prompts, as evident in the following brief qualitative analysis of the core elements of each text content (a full analysis of an example text is provided in Section 7.3 & and others in the Appendix for space considerations):

- **Text 1 (A1):** This lyric, targeting the A1 level, displays the LLM's ability to generate material that caters to beginners in English language learning. The use of familiar mythological names, Zeus, Venus, and Osiris, provides a thematic framework that is likely to be engaging for learners. Simple sentence structures, using common verbs and basic sentence patterns, ensure that the text is accessible for those with limited English proficiency (TGG). The Gothic and conversational tone, achieved through word choices like "shadows," "whispers," and the use of direct address like "So listen well...", creates an engaging and somewhat dramatic atmosphere, aligning with SFL's concept of register. The incorporation of prepositions of time, place, and direction such as "Upon Olympus," "In Rome's heart," and "Through the Nile" provides grammatical exposure suitable for this proficiency level, reinforcing the understanding of spatial and temporal relations. The emphasis on theme and rheme manipulation "Zeus speaks" vs. "mighty and tall, Zeus speaks", highlighting sentence structure, further underpins the model's understanding of TGG principles and its ability to introduce grammatical concepts in a simplified manner.

- **Text 2 (A2):** This narrative text effectively encompasses elements from all three linguistic theories. The use of simple language and straightforward sentences aligns with the A2 level, ensuring accessibility for learners with basic English proficiency. The inclusion of direct and indirect speech "One visitor might say, 'I've never seen anything so grand!'" vs. "In a similar adventure a year ago, a visitor wondered how genius the pharaohs had been to build such an amazing construction" showcases the model's ability to represent different modes of speaking and reporting, which aligns with SFL's focus on the interpersonal function, and various transformational structures, which accords with the key aspect of the TGG. The use of modifiers as in "huge temples and "giant statues", comparative forms such as "even older", and superlative forms like "one of the grandest" demonstrates the model's understanding of how to enhance descriptions and create emphasis, again aligning with SFL's emphasis on the ideational function. Besides, including all these types of grammatical structures reflects the model's understanding of such basic structures. The exciting and adventurous tone aligns with SFL principles by enhancing engagement and creating a specific register appropriate for a travelogue-style narrative. The text also highlights cultural sensitivity (GEs) by focusing on the Abu Simbel temple and comparing it to other significant structures from different cultures, promoting an appreciation for global diversity.

- **Text 3 (B1):** The LLM successfully adopted the role of a friendly TV announcer, using a conversational tone and addressing the audience directly, as requested in the prompt. This aligns with the SFL principle of considering the tenor, which encompasses the social roles and relationships within a text. The text incorporates the specified tenses (Present Simple, Present Continuous, Present Perfect) to discuss the impact of excessive mobile phone usage on children, demonstrating the AI's understanding of how to manipulate tense and aspect, key elements of TGG. The inclusion of concrete examples from diverse cultural contexts (Maria from Spain, Raj from India, etc.) and the reflection on social implications in different communities further demonstrate the model's ability to execute the prompt's instructions regarding GEs principles, showcasing its capacity for cultural sensitivity and awareness of diverse English varieties.
- **Text 4 (B2):** This text successfully incorporates various prompt elements related to SFL. The use of rhetorical questions ("Why should we unite?") aims to engage the reader and provoke reflection, a key aspect of the interpersonal meaning in SFL. The inclusion of conditional sentences (first, second, and third) reveals the model's ability to express hypothetical situations and their consequences, aligning with the ideational function of language in SFL, which focuses on representing experiences and logical relations. The persuasive tone throughout the text aims to influence the reader's attitudes and actions and convey the message of uniting against COVID-19, again aligning with SFL's emphasis on the interpersonal function. The model also effectively employed Barack Obama’s speech style features, as requested, further exhibiting its ability to adopt a specific persona and tailor its language to a particular audience.
- **Text 5 (C1):** The model effectively generated a dialogue in the Shakespearean style, incorporating elevated language, archaic expressions ("verily," "doth," etc.), and technical medical terms, as requested. Such stylistic features agree with SFL's focus on register, i.e. the language variety appropriate for a specific situation. Employing both British and Indian English varieties ("verily" vs. "badmash") further highlights the model's understanding of the

GEs' parameters, exposing its ability to incorporate various linguistic and cultural contexts. In addition, the delicate integration of passive voice transformations, such as "The sebum, trapped within, festers and provokes inflammation" pinpoints the model's ability to manipulate syntactic structures and apply the principles of TGG, as requested in the prompt.

- **Text 6 (C2):** This text signals the LLM's capability to perform complex writing tasks and generate sophisticated language structures. The prompt requested a comprehensive film review of "The Last Samurai" targeting senior college students and containing specific stylistic and linguistic features. The model successfully generated a text that met these requirements, using a complex style copying T.S. Eliot's writing characterized by intricate sentence structures and literary language. Using American and Japanese English varieties, such as the use of the Japanese term "bushido," further exposes the model's understanding of the GEs' features through presenting linguistic diversity within a single text. The use of subordinate relative and noun clauses, such as "Within the hallowed celluloid sanctum, **where** narratives entwine like Möbius strips, Edward Zwick's 'The Last Samurai' stands as an enigmatic tapestry" and "His odyssey mirrors Japan's own tempestuous struggle to reconcile its ancient heritage with the inexorable march toward industrialization **that** reverberates through mist-shrouded forests," respectively reflects the model's ability to generate various complex syntactic structures (TGG), as required.

As elucidated in the above analysis, the linguistically developed prompt engineering model has efficiently guided the AI in producing ESL reading materials that contain the specified parameters from the language theories and basic prompt engineering. The model has successfully addressed different CEFR levels, used diverse linguistic features, and adopted various styles and tones, as requested in the prompts. While minor discrepancies in readability scores and metric alignment highlight the need for careful interpretation, the overall effectiveness of the model provides strong evidence for its potential to assist teachers in creating customized ESL reading materials. In fact, using quantitative readability score analysis and qualitative analysis of the AI's performance in understanding and executing the prompts has been instrumental in evaluating the effectiveness of the developed prompt engineering model. This approach provides a more comprehensive understanding of the model's capabilities, limitations, and areas for improvement. The quantitative analysis, employing FKRE, FKGL, and GFI, provides an objective evaluation of the language complexity of the generated texts. It allows for a direct comparison between the targeted CEFR levels and the assessed readability levels, which, in turn, points out the AI's ability to adapt language complexity to intended learner(s)' proficiencies. The qualitative analysis deeply tackles the specific linguistic features present in the generated texts and thus examines how well the AI has incorporated the prompt parameters derived from the language theories (TGG, SFL, and GEs) and basic prompt engineering elements. This analysis goes beyond the surface-level assessment provided by readability scores, considering aspects such as sentence structure, vocabulary choices, stylistic elements, tone, and cultural sensitivity (see Section 8 for further discussion).

7.3 A Full Analysis of an Example Generated Text

This section provides a detailed analysis of a generated text, Text 3, following the step-by-step processing of the present study's developed prompt engineering model. This analysis underscores the AI model's ability to understand and execute the prompts, incorporating the specified parameters to generate customized ESL reading material, reflecting the efficacy and applicability of the developed model, as seen below:

A) Start: The process begins by defining the teacher’s learning objectives and the learners' language level. In this case, the targeted learners are 7th graders with a B1 (Intermediate) level of English proficiency, as defined by the CEFR. The teacher has multiple language, content, and skill learning objectives for this reading material by the end of which students will be able to:

a. Accurately use the present simple, present continuous, and present perfect tenses, demonstrating an understanding of their different functions and forms.
b. Expand their vocabulary related to technology, social issues, and global cultures, using new words in context.
c. Identify the impact of excessive mobile phone usage on children.
d. Demonstrate an understanding of how mobile phone usage and its social implications vary across different cultures around the world.

e. Read and comprehend an informative text on mobile phone usage, identifying key ideas and supporting details.
f. Critically evaluate the information presented in the text on social and personal implications of technology use, forming their own opinions and supporting them with evidence.

By defining these specific learning objectives, the teacher can then select appropriate prompt parameters from the developed model to guide the AI in generating a text that effectively meets these objectives.

B) Basic Effective AI Prompting Parameters: These elements form the foundation of the prompt and are largely universal for effective output. The prompt contains the basic elements described in Table 10 below.

C) Language Theories-Based Prompt Parameters: Based on the learning objectives and learners' language level, the teacher selects one or more elements from the three language theories to integrate into the prompts, which allows for a tailored comprehensive approach to language instruction. In this example text, the prompt comprises the described elements from TGG, SFL, and GEs in Table 10 below.

D) Combine Elements: The basic prompt elements are combined with the selected theories-based elements to create a comprehensive and effective AI prompt, as follows:

Assume the role of a TV announcer and compose an informative article of about 300 words that examines the impact of excessive mobile phone usage on children. Employ the present simple tense, the present continuous tense, and the present perfect tense. Use the present continuous tense 4 times in the text and the present perfect tense 4 times in the text and write the rest of the text in the present simple tense. Adapt your writing to be comprehensible and engaging for ESL learners of B1 language level according to the Common European Framework of Reference (CEFR), i.e., 7th graders of an intermediate language level, where such learners can understand texts that consist mainly of high-frequency everyday or job-related language and can also understand the description of events, feelings, and wishes in personal letters. Maintain a friendly tone throughout the article. Illustrate each key point with concrete examples drawn from a variety of cultural contexts, ensuring the content is relatable to a diverse readership and resonates with the young audience. Reflect on the social implications of mobile phone usage in different communities around the world.

Table 10: Prompt Elements Analysis According to the Developed Model Parameters

Prompt Element	Model Parameter	Explanation
Role of a TV Announcer	Persona	The prompt explicitly instructs the AI to assume the role of a TV announcer, which sets the tone and style for the generated article.
Informative article that examines the impact of excessive mobile phone usage on children	Action/Task, Input Data & Output Format	The prompt clearly states that the AI should compose an informative article, providing a specific task and topic.
about 300 words	Output Length	The prompt clearly sets the maximum length of the article to avoid unnecessarily long and irrelevant responses Copilot Chatbot might provide.
Tenses Used: Present Simple (throughout the article), Present Continuous (4 times), Present Perfect (4 times)	TGG: Structure Identification	The prompt specifies the type of tense to be used (present simple, present continuous, and present perfect) and how many times each is used.
Tenses Used: Present Simple, Present Continuous, Present Perfect	TGG: Variation in Syntactic Complexity	The variation in tenses complexity (simple, continuous, perfect) aligns with TGG's focus on syntactic complexity.
Informative Article	SFL: Communicative Function Identification	The prompt specifies the communicative intent (informative article: genre), aligning with SFL's focus on language functions.
ESL learners (B1 level/ 7th graders) & friendly tone	SFL: Contextual Relevance & Register Focus, Audience & Tone	The adaptation of the intended text for ESL learners (B1 level) according to the CEFR (target audience) and the friendly tone throughout the article align with SFL's consideration of context and register.
Reflection on Social Implications in Different Communities	SFL: Language Metafunctions Inclusion	Such reflection on social implications in different communities aligns with SFL's focus on the ideational language metafunction (conveying experiences).
Concrete Examples from Cultural Contexts & Relatability to Diverse Readership	GEs: Diversity Emphasis & Cultural Sensitivity	The prompt emphasizes relatability to diverse readership and cultural background, reflecting GEs' focus on global diversity.

E) Generate Reading Material: The final devised prompt is fed into the LLM (Microsoft Copilot in this context) to generate the required response. Here is the AI Chatbot's generated Text 3:

Breaking News: The Mobile Dilemma – How Phones Shape Our Kids' World

By Alex "Curious" Copilot

Introduction: The Tiny Screens That Hold Big Secrets

Hello, dear viewers! I'm Alex, your friendly TV announcer, and today we delve into a topic that affects kids worldwide: mobile phones. These pocket-sized gadgets have become our digital companions, but are they a blessing or a curse? Let's explore the impact of excessive mobile phone usage on our young generation.

A Snapshot of Habits

Kids love their phones. It's a fact. Whether they're texting friends, scrolling through memes, or capturing TikTok dances, these little screens are glued to their hands. But what does this mean for their well-being?

Sleep Disruption *Excessive phone use disrupts sleep patterns. Imagine young Maria, a soccer enthusiast from Spain. She stays up late chatting with her virtual teammates, missing out on precious rest. The next day, she's yawning through math class. Poor Maria!*

Social Disconnect *Phones connect us globally, yet sometimes they disconnect us locally. Take Raj, a cricket fan from India. He's so busy posting selfies that he forgets to chat with his grandma. She's right there, sipping chai, waiting for a real conversation. Phones, don't steal our moments!*

Caught in the Act

Right now, kids are swiping. They're liking, sharing, and LOL-ing. But what's happening behind those screens?

Cyberbullying *Meet Aisha, a tech-savvy girl from Egypt. She's online, but so are the bullies. Hurtful messages pop up like digital ghosts. Aisha feels trapped, her heart racing. We must protect our young warriors!*

FOMO (Fear of Missing Out) *Across the globe, kids are feeling FOMO. In Brazil, Pedro checks Instagram while his friends surf at Copacabana Beach. He sighs, missing the salty breeze. Phones, don't steal our sunshine!*

Echoes of Choices Made

Kids have unlocked secrets. They've snapped, shared, and swiped. But what's the aftermath?

Lost Adventures *Remember Li Wei from China? She climbed the Great Wall but missed the view. Her eyes were on the screen, capturing memories for followers. Now she wonders, "Did I really see the Wall?" Phones, don't steal our wonders!*

Digital Footprints *In Nigeria, Chidi's phone holds his life story. Every photo, every emoji – it's all there. But what about privacy? His digital footprints follow him like invisible ink. Phones, don't steal our secrets!*

Conclusion: The Global Symphony of Beeps and Buzzes

Phones are bridges and barriers. They connect us to distant lands but sometimes disconnect us from our own backyards. Let's teach our kids balance – the art of swiping and savoring sunsets. Because in this digital dance, they hold the remote control.

Remember, young viewers, phones are tools, not rulers. Use them wisely, and let's create a world where screens enhance our lives, not steal our moments. Signing off, this is Alex, your curious companion, urging you to find the right frequency in this mobile symphony.

F) Evaluate Output: The generated reading material is evaluated based on its relevance to the teacher's requirements, i.e., its inclusion of the prompt elements, hence effectiveness in meeting the learning objectives and learners' language level, as illustrated in the Table 11 below:

Table 11: Analysis of Prompt Elements Incorporated in Example Text 3

Prompt Element	Reading Text Content	Example
Role of a TV Announcer	The entire article assumes the role of a friendly TV announcer, addressing the audience directly.	"Hello, dear viewers! I'm Alex, your friendly TV announcer..."
Informative Article (about 300 words)	The article is informative, discussing the impact of excessive mobile phone usage on children, but its length is 510 words.	"Breaking News: The Mobile Dilemma – How Phones Shape Our Kids' World". Length is more than 300 words since the word "about" was used, which means that the length must be around this range, maybe less and maybe more, and the researcher also set a definite number of present continuous and present perfect tenses examples to be available in the text (4 times each) along with concrete examples from various cultural contexts, and this was met by the Chatbot, which definitely influenced the Chatbot's abiding by the required number of words.
Tenses Used: Present Simple, Present Continuous, Present Perfect	- Present Simple: Used throughout the article for general statements and facts. - Present Continuous: Used four times to describe ongoing actions related to phone usage. - Present Perfect: Used four times to discuss past actions with present relevance.	"Kids love their phones.", "Phones connect us globally..." "Right now, kids are swiping.", "Kids are feeling FOMO." "Kids have unlocked secrets.", "They've snapped.."
Adaptation for ESL Learners (B1 Level)	The language is clear, straightforward, and suitable for intermediate learners (B1 level).	"She stays up late chatting," Kids love their phones."
Friendly Tone	The tone is friendly and engaging, addressing the young audience directly.	"Remember, young viewers...", "Signing off, this is Alex..."
Concrete Examples from Cultural Contexts	Examples include Maria (Spain), Raj (India), Aisha (Egypt), Pedro (Brazil), Li Wei (China), and Chidi (Nigeria).	"Imagine young Maria, a soccer enthusiast from Spain.", "Take Raj, a cricket fan from India."
Reflection on Social Implications in Different Communities	The article touches on sleep disruption, social disconnect, cyberbullying, FOMO, lost adventures, and digital footprints.	"Excessive phone use disrupts sleep patterns," "Phones connect us globally, yet sometimes they disconnect us locally."
Relatability to Diverse Readership	The content resonates with a diverse audience, emphasizing universal experiences related to phone usage.	"Phones are bridges and barriers," "They connect us to distant lands but sometimes disconnect us from our own backyards"

G) Adjust Prompt (If Needed): If the output does not meet the requirements, the prompt can be adjusted and the process repeats. In this case, the generated Text 3 generally aligns with the prompt requirements, but the length exceeds the requested 300 words. If a stricter word limit is desired, the prompt could be adjusted to specify a more precise word count range. However, the researcher did not modify the length since the linguistic and stylistic features required were successfully and accurately incorporated in the prompt.

H) End: The process concludes with a satisfactory piece of ESL reading material generated by the LLM. In this example, the AI-generated Text 3 effectively encompasses the specified prompt elements, demonstrating the LLM's ability to generate customized ESL materials that accords with the teacher's learning objectives and the learners' language level, thus highlighting the success of the developed model in addressing the AI via devising effective prompts.

8. Discussion

The results of the present study reveal that integrating prompt engineering and language theories to design a linguistically informed model of prompt parameters can tremendously augments LLMs' ability to generate customized ESL reading materials, which answers the central research question. The analysis of the generated texts' readability scores as well as the qualitative investigation of their linguistic features indicate that the current developed model successfully guides the AI in producing texts that satisfy both targeted CEFR proficiency levels and specific pedagogical goals.

With regard to the sub-questions on the model's development, the present study has suggested elements from TGG, SFL, and GEs that are essential for addressing LLMs to produce customized ESL reading materials. These elements encompass specifying syntactic structures (TGG), identifying communicative functions and contextual relevance (SFL), and endorsing diversity and cultural sensitivity (GEs). The study has further pointed out how such linguistic elements can be integrated with basic prompt engineering principles to create an effective model of prompt parameters. The researcher sought to design a model that is accessible for educators regardless of their different levels of expertise in prompt engineering and utilized a simple prompting strategy, one-shot/zero-shot query template, (see Sections 6.2.1.2, 6.2.1.4, & Figure 1), to enable teachers to leverage AI without requiring extensive technical knowledge. This addresses a gap highlighted by Woo et al. (2023) who found that students often struggle with prompt engineering due to lack of experience and understanding.

Concerning the model's evaluation, the study has found out that the developed model effectively guides the employed LLM, Microsoft Copilot, in generating ESL reading texts that address the different intended CEFR language levels, answering the first sub-question in this respect. In fact, the analysis of readability scores has revealed a consistent decreasing readability as CEFR levels advanced, which pinpoints that the AI successfully incorporated the CEFR framework into its text generation process. This finding echoes previous research on LLMs' ability to adapt language complexity for different tasks. For instance, studies examining ChatGPT's performance on standardized tests (Choi et al., 2023; Das et al., 2023; Gilson et al., 2023; Huh, 2023; Sinha et al., 2023) demonstrate its capacity to handle complex reasoning and knowledge retrieval across varying levels of difficulty. This is similar to Labruna et al.'s (2023) study which highlights that ChatGPT can generate dialogues enormously resemble human-generated ones in task-oriented scenarios.

Regarding the correspondence between readability scores and intended CEFR levels, the study, responding to the second sub-question, has observed a general alignment between the two, which supports the model's effectiveness. However, minor discrepancies in readability scores for certain texts indicate the inherent limitations of readability metrics, as discussed in Section 7.1, which underscores the importance of using multiple readability metrics and interpreting their scores carefully combined with qualitative analysis. Unlike Young and Shishido (2023), who also used readability metrics to evaluate ChatGPT-generated dialogues, this study has explicitly fed the AI with the intended CEFR language level of each text, demonstrating the LLM’s ability to comprehend and respond to these specific parameters. To explicate, the CEFR was utilized by Young and Shishido (2023) to determine the target audience best suited for the AI-generated dialogue materials. Based on the readability scores, the researchers concluded that the ChatGPT-generated dialogues were most appropriate for students at the CEFR A2 (elementary) proficiency level. That is, the researchers did not feed the AI chatbot, ChatGPT, by the intended language level of the generated texts since they did not assume in advance that the AI can understand the CEFR language levels and create texts accordingly. Conversely, the current study has used three readability metrics to evaluate the suitability of the produced texts for different EFL proficiency levels and fed the LLM via the devised prompts with the intended CEFR language level for every targeted generated text to be produced accordingly. This, in turn, has helped the AI to generate each reading text in accordance with the meant language level. Furthermore, this study, different from Young and Shishido (2023), has complemented such readability evaluation with an analysis of the extent to which the AI-generated texts incorporate the specified prompt parameters derived from language theories and basic prompt engineering elements, ensuring that the materials comply with the intended linguistic features and learning objectives, as patent in Section 7.2 and further discussed below.

Addressing the third sub-question regarding the AI's incorporation of the specified prompt parameters derived from TGG, SFL, and GEs and basic prompt engineering elements into the generated texts, the qualitative analysis has emphasized that the employed LLM effectively integrated such parameters. The consistent presence of targeted linguistic features, such as, but not limited to, specific tenses and structures (TGG), rhetorical questions and various tones (SFL), and culturally diverse examples (GEs), across the six texts demonstrates the LLM’s ability to understand and perform complex prompts that integrate linguistic features along with the stated prompt engineering elements. In fact, the study's focus on incorporating specific linguistic features tailored to different CEFR levels goes beyond relying solely on general-purpose metrics like ROUGE-L and BERT Score, which may not fully capture the nuances of language learning, as highlighted by Ochieng (2023) in his evaluation of LLM-generated questions.

Finally, with respect to the developed model's versatility, the explored LLM in the present study, in response to the fourth sub-question, has successfully generated texts across different text genres and registers, including informative articles, poetic dialogues, opinion articles, film reviews, narrative descriptions, and lyrics. These various text types exhibit the designed model's adaptability to various pedagogical needs, which emphasizes its capability of creating diverse and engaging ESL reading materials. This accords with research on the use of AI in language learning that highlights the importance of exposing learners to a wide range of authentic language use (Fryer et al., 2020; Kim et al., 2022). Furthermore, the model has succeeded in generating several text types using the one-shot prompting approach/zero-shot technique, which contrasts with, but not opposing, the findings of Young and Shishido (2023), who suggested exploring different prompting techniques to increase the variety and quality of AI-generated dialogues. That is, their suggestion is valid; however, it is not a must since using simple prompt strategies and techniques would also help increase the diversity and value of the AI-created texts, as evident in the present study.

The current study has significant implications for the integration of AI in ESL education, particularly with relation to personalized learning. To explain, the built model has enabled AI to generate customized reading materials responsive to targeted CEFR levels and pedagogical goals, which allows educators to create learning experiences that cater to the specific needs of their learners. Creating texts that are grammatically accurate, functionally relevant, contextually appropriate, and culturally sensitive by adopting stylistic elements and linguistic features from the TGG, SFL, and GEs theories causes such texts to be responsive to current trends in ESL education. Such trends emphasize communicative competence, cultural awareness, and real-world language use. This approach mirrors Ochieng's (2023) call for further exploration of LLMs' social influence as reading guides in relation to their production of materials that can address particular educational goals and endorse meaningful learning practices.

Moreover, the present study contributes to the growing field of employing AI in language education via chatbots and LLMs as tools that can help provide personalized practice opportunities and feedback (Fryer et al., 2020; Jeon, 2022; Kim et al., 2022) via presenting a more theoretically based approach to developing AI generated material. The designed model, by integrating different linguistic theories into the prompt engineering process, overcomes a crucial limitation in several AI tools for language learning; such tools give much consideration to grammatical accuracy over communicative competence and cultural relevance. Furthermore, by providing an intelligible model, the study aims to make generative AI more accessible for educators without necessitating expertise in prompt engineering.

Besides, implementing the developed model in real-world ESL classrooms has potential benefits. Teachers can save much time and effort by using LLMs to create customized reading materials that comply with their intended learning objectives and students' language proficiencies. Accordingly, they will be able to devote more time and effort to personalized instruction and student support. Moreover, the adaptability of reading materials can enhance the effectiveness of reading comprehension activities. By generating texts across different text types, the model provides an opportunity for learners to be exposed to a wider range of authentic language use in contrast to traditional textbook-based learning that deprives them of real-world communication contexts. Such approach agrees with Li et al.'s (2023) highlighting ChatGPT's ability to generate high-quality reflective writing, which advocates that AI is a valuable tool for promoting higher-order thinking skills in language learning.

In the present study, there are some limitations that need further investigation. First, the study has employed only one LLM, Microsoft Copilot, in its examination of the developed model, which requires further exploration with other LLMs to evaluate the model's performance across different generative AI systems of this type. Second, the reliance on readability metrics, while providing a quantifiable measure of language complexity, presents inherent limitations since these metrics may not fully capture the nuanced aspects of language proficiency, leading to minor discrepancies between readability scores and perceived difficulty, as explained in Section 7.1. However, it is important to note that the detailed qualitative analysis provided by the researcher for each of the six generated texts complements the quantitative analysis via the metrics and ensures the validity of the model's evaluation (see Section 7.2). This detailed analysis, in addition, compensates largely for the limited sample size of the diverse generated texts, one for each CEFR level, especially since each text has met the intended CEFR language level without the need for any modifications in the devised prompt, which, in turn, lessens the restriction on the generalizability of findings to a wider range of text topics and types. This also supports the prompt technique used in the current study: the one-shot technique/zero-shot prompting setting, meant to ease the prompting process for teachers who are not necessarily experts in prompt engineering.

9. Conclusion

This study has investigated the potential of integrating prompt engineering and linguistic theories to enhance the generation of customized ESL reading materials by AI LLMs. The findings reveal that the developed prompt engineering model, incorporating elements from TGG, SFL, GEs, and basic prompt engineering, has successfully guided the AI in producing texts that address both targeted CEFR proficiency levels and specific learning objectives. The analysis of readability scores unveils a general correspondence between the generated texts and the intended CEFR levels, indicating the AI's ability to modify language complexity. Furthermore, the qualitative analysis reveals the LLM’s ability to understand and perform complex, linguistically informed prompts through highlighting the incorporation of the intended linguistic features, stylistic elements, and culturally relevant content in the LLM's generated texts.

The findings shed light on the way AI can be enhanced to create customized ESL reading materials that respond to the different needs of learners. The AI's success in generating texts across several genres using the developed model fosters the latter's versatility and potential for enriching ESL education. In fact, the model developed in the present study addresses a gap in existing research which views prompt engineering as a purely technical process without taking into account linguistic features. The model represents a pragmatic and theoretically thorough framework for integrating linguistic theories into the generating prompts process, thus contributing to guaranteeing that AI-generated content is grammatically correct, functionally pertinent, contextually suitable, and culturally aware. The study's development of such manageable model of prompt parameters simplifies the creation of prompts on educators' part, which leads AI to be more accessible for non-expert teachers in this area, hence broadening AI adoption in educational settings. This contribution addresses a key challenge highlighted in previous research, which emphasizes the difficulties faced by non-experts in effectively utilizing AI tools in language education.

Building upon the current research, further studies can be conducted in several key areas. Integrating linguistic features, other than these utilized in the present study, such as discourse markers or pragmatic elements, into the developed model needs to be investigated as to their effectiveness in enhancing generative LLMs in producing customized ESL (reading) materials. some studies could examine the developed model's performance with other LLMs, including ChatGPT, Gemini, Claude, Mistral, Llama, and others to determine its generalizability and identify possible variations in text generation quality across several AI systems. Furthermore, empirical studies could be conducted with ESL learners to assess the effectiveness of AI-generated materials in improving their reading comprehension, vocabulary acquisition, and overall language development. Such studies would provide valuable insights into the pedagogical impact of the model and contribute to its refinement. In addition, involving ESL teachers in user studies is essential to evaluate the model's usability and practicality in real classroom settings, which would also contribute to improving the model's design. This aligns with Clarisó and Cabot's (2023) call for research on the practical applications and challenges of model-driven prompt engineering, highlighting the importance of empirical evaluation in real-world scenarios. These recommended future studies can lead to deeper integration of AI in ESL education, thus enhancing language learning experiences and creating personalized, engaging, and culturally sensitive materials for learners of all backgrounds and proficiency levels.

References

Alasmari, A. (2018). The attitudes of Saudi EFL teachers towards the inclusion of Saudi culture in EFL textbooks. *Journal of Language Teaching and Research*, *9*(4), 760–768. https://doi.org/10.17507/jltr.0904.14

Alduais, A. M. S. (2015). Role of the transformational generative grammar and other linguistic theories in teaching English as a foreign language. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2611383

Alshalan, A. A., & Alyousef, H. S. (2020). Evaluating EFL textbooks from a systemic functional perspective: A focus on grammatical metaphor. *Indonesian Journal of Applied Linguistics*, *9*(3), 694–704. https://doi.org/10.17509/ijal.v9i3.23781

Atlas, S. (2023). ChatGPT for higher education and professional development: A guide to conversational AI. *DigitalCommons@URI*. https://digitalcommons.uri.edu/cba_facpubs/548

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education, 36*. https://doi.org/10.1186/s41239-023-00427-0

Bartlett, T., & O'Grady, G. (Eds.). (2017). *The Routledge handbook of systemic functional linguistics*. Routledge.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 1-17. https://doi.org/10.1016/j.caeai.2021.100027

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT goes to law school. *SSRN*, *71*(387), 1-16. http://dx.doi.org/10.2139/ssrn.4335905

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. *arXiv*. https://doi.org/10.48550/arXiv.2204.02311

Clarisó, R., & Cabot, J. (2023). Model-driven prompt engineering. *Proceedings of 2023 ACM/IEEE 26th international conference on model driven engineering languages and systems (MODELS)*, 47-54. IEEE. 10.1109/MODELS58315.2023.00020

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Das, D., Kumar, N., Longjam, L. A., Sinha, R., Roy, A. D., Mondal, H., & Gupta, P. (2023). Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus, 15*(3). e36034. doi:10.7759/cureus.36034

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv.* https://doi.org/10.48550/arXiv.1810.04805

Ding, B., Qin, C., Liu, L., Bing, L., Joty, S., & Li, B. (2023). Is GPT-3 a good data annotator? *arXiv.* https://doi.org/10.48550/arXiv.2212.10450

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A survey of natural language generation. *ACM Computing Surveys, 55*(8), 1–38. https://doi.org/10.1145/3554727

Flesch Kincaid calculator. (2024). Character calculator. Retrieved March 8, 2024, from https://charactercalculator.com/flesch-reading-ease

Online readability test calculator. (2024). Readabilit. Retrieved March 8, 2024, from https://textcompare.org/readability

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology, 24*(2), 8–22. Retrieved March 3, 2024, from http://hdl.handle.net/10125/44719

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior, 93*, 279–289. https://doi.org/10.1016/j.chb.2018.12.023

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education, 9*. 10.2196/45312

Gunning Fog index. (2024). Character calculator. Retrieved March 8, 2024, from https://charactercalculator.com/gunning-fog-index

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar*. Routledge.

Han, D.-E. (2020). The effects of voice-based AI chatbots on Korean EFL middle school students' speaking competence and affective domains. *Asia-pacific Journal of Convergence Research Interchange, 6*(7), 71–80. https://doi.org/10.47116/apjcri.2020.07.07

Huh, S. (2023). Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. *Journal of Educational Evaluation for Health Professions, 20*, (1). 10.3352/jeehp.2023.20.1

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jeon, J. (2022). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, *37*(1), 1–26. https://doi.org/10.1080/09588221.2021.2021241

Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Front. Psychol*, *13*, 1–20. https://doi.org/10.3389/fpsyg.2022.1049401

Kim, H., Yang, H., Shin, D., & Lee, J. H. (2022). Design principles and architecture of a second language learning chatbot. *Language Learning & Technology,* *26*(1), 1–18*. http://hdl.handle.net/10125/73463

Kim, N.-Y. (2018). Chatbots and Korean EFL students' English vocabulary learning. *Journal of Digital Convergence, 16*(2), 1–7. https://doi.org/10.14400/JDC.2018.16.2.001

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *arXiv*. https://doi.org/10.48550/arXiv.2205.11916

Kung, T. H., Cheatham, M., Medinilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *medRxiv*. https://doi.org/10.1101/2022.12.19.22283643

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y

Labruna, T., Brenna, S., Zaninello, A., & Magnini, B. (2023). Unraveling ChatGPT: A critical analysis of AI-generated goal-oriented dialogues and annotations. *arXiv.* https://doi.org/10.48550/arXiv.2305.14556

Leong, K., Sung, A., & Jones, L. (2023). The core technology behind and beyond ChatGPT: A comprehensive review of language models in education research. *IJERI: International Journal of Educational Research and Innovation,* (20), 1-21. https://doi.org/10.46661/ijeri.8449

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively? *Computers and Education: Artificial Intelligence, 4*. https://doi.org/10.1016/j.caeai.2023.100140

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (arXiv:2107.13586). *arXiv*. https://doi.org/10.48550/arXiv.2107.13586

Microsoft. (2024a, Jan. 23). How Copilot works, technically speaking. Microsoft. Retrieved, March 2, 2024, from https://www.microsoft.com/en-us/bing/do-more-with-ai/how-bing-chat-works

Microsoft. (2024b, Jan. 23). What is Copilot, and how can you use it? Microsoft. Retrieved, March 2, 2024, from https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it

Mohamed, S. S., & Alian, E. M. (2023). Students' attitudes toward using chatbot in EFL learning. *Arab World English Journal*, *14*(3), 15-27. https://dx.doi.org/10.24093/awej/vol14no3.2

Ochieng, P. (2023). Are large language models fit for guided reading? *arXiv.* https://doi.org/10.48550/arXiv.2305.10645

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., … Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Petrovic, J., & Jovanovic, M. (2020). Conversational agents for learning foreign languages: A survey. *arXiv.* https://doi.org/10.48550/arXiv.2011.07901

Prompt engineering guide. (2024). DAIR.AI. Retrieved March 3, 2024, from https://www.promptingguide.ai/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. Retrieved March 1, 2024, from https://paperswithcode.com/paper/language-models-are-unsupervised-multitask

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*, 1–67. Retrieved Feb 22, 2024, from https://jmlr.org/papers/volume21/20-074/20-074.pdf

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–7. Association for Computing Machinery. https://doi.org/10.1145/3411763.3451760

Russe, M. F., Reisert, M., Bamberg, F., & Rau, A. (2024). Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *Rofo*. 10.1055/a-2264-5631

Sinha, R. K., Roy, A. D., Kumar, N., & Mondal, H. (2023). Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus, 15*(2). 10.7759/cureus.35237

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv.* https://doi.org/10.48550/arXiv.2212.09292

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv.* https://doi.org/10.48550/arXiv.2201.08239

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. *Proceedings of the 23rd international conference on artificial intelligence in education*, 153–166. Springer, Cham. https://doi.org/10.1007/978-3-031-11644-5_13

Whong, M. (2007). Seeking consensus: Generative linguistics and language teaching. *University of Leeds*, (12), 143–155. Retrieved March 1, 2024, from Whong_2007.pdf (leeds.ac.uk)

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?- A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence, 4*. 10.3389/frai.2021.654924

Woo, D. J., Guo, K., & Susanto, H. (2023). Cases of EFL secondary students' prompt engineering pathways to complete a writing task with ChatGPT. *arXiv.* https://doi.org/10.48550/arXiv.2307.05493

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., & Haffari, G. (2024). Continual learning for large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2402.01364

Yadav, M. S., & Yadav, M. K. (2020). Role of the transformational generative grammar and other language learning theories in English language teaching. *SSRN*. 10.2139/ssrn.3665190

Young, J. C., & Shishido, M. (2023). Investigating OpenAI's ChatGPT potentials in generating chatbot's dialogue for English as a foreign language learning. *International Journal of Advanced Computer Science & Applications, 14*(6). 10.14569/IJACSA.2023.0140607

Zhao, Z., Hou, Y., Wang, D., Yu, M., Liu, C., & Ma, X. (2022). Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv.* https://doi.org/10.48550/arXiv.2203.14187

Appendix

Example Prompts, Generated ESL Reading Texts, & Analyses

Example 1:

Prompt: Please, act as a lyric writer and create a lyric of about 200 words for ESL learners of A1 (Beginners) language level according to the Common European Framework of Reference (CEFR) where learners can understand familiar names, words, and very simple sentences, for example on notices and posters or in catalogues. The lyric is on 3 ancient myths: Greek, Roman, and Egyptian and written in a gothic and conversational tone where a few characters are conversing with each other. The lyric must involve various prepositions of time, place, and direction. It must also play with themes and rhemes changing their slots in the sentences. Such themes and rhemes must be written in bold.

Reading Text 1 (A1)

Microsoft Copilot's Response:

In the **shadows** of the past, **whispers** rise and fall,
Zeus speaks with thunder, **mighty** and tall.
"**Upon** Olympus, we gods **dwell**,
In the sky's embrace, where stars **tell**."

Venus glides in, beauty **unseen**,
"**In** Rome's heart, I reign as queen.
Through the ages, my love **spreads**,
Over hills and time, it **treads**."

Osiris wakes, **wrapped** in night,
"**Beneath** Egypt's sands, I **fight**.
For life's cycle, death's **door**,
Through the Nile, my myths **soar**."

Chorus:
"**From** myths **old**, **to** lands **far**,
Through time's gate, we **travel** **by** star.
In tales **told**, **by** fire's **light**,
We live **on**, **in** dreams **tonight**."

Zeus to **Venus**, lightning **bold**,
"**At** dawn's break, your stories **told**.
Through Rome's streets, whispers **carry**,
Of love's triumphs, moments **airy**."

Venus to **Osiris**, with a **smile**,
"**By** the pyramids, stay **awhile**.
In your realm, secrets **keep**,
Under stars, **where** pharaohs **sleep**."

Osiris to **Zeus**, **ancient** and **wise**,
"**On** the horizon, the sun **rises**.
With each day, hope **renews**,
Among the gods, **in** different **hues**."

Chorus:
"**From** myths **old**, **to** lands **far**,
Through time's gate, we **travel** **by** star.
In tales **told**, **by** fire's **light**,
We live **on**, **in** dreams **tonight**."

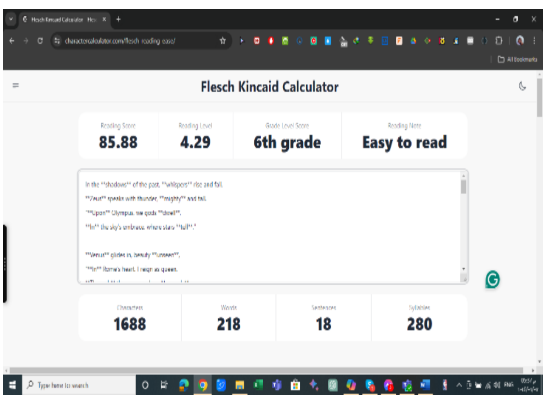
End:
"**So** listen **well**, **to** these words **woven**,
Of gods and myths, **eternally** **proven**.
In stories **shared**, **from** heart **to** heart,
Our legends **live**, **never** to **depart**."

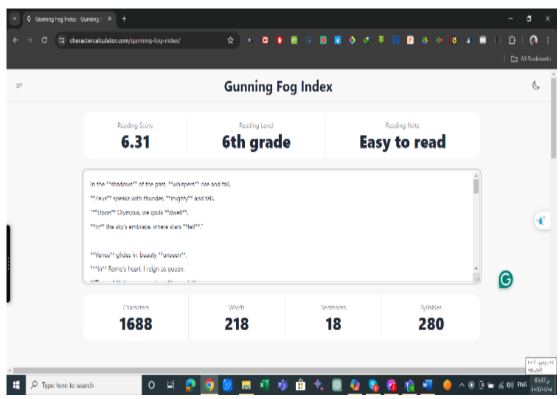
Table A1: Analysis of Prompt Elements Incorporated in Text 1

Prompt Element	Lyric Content	Examples from the Lyric
200 words for A1 ESL learners	The lyric is concise and uses simple language, with a total word count close to 200.	The entire lyric is crafted with simplicity in mind, avoiding complex vocabulary.
Understand familiar names, words, and very simple sentences	Familiar mythological names and simple sentence structures are used.	“Zeus speaks with thunder,” “Venus glides in,” “Osiris wakes.”
3 ancient myths: Greek, Roman, and Egyptian	Each stanza introduces elements from Greek, Roman, and Egyptian mythology.	Zeus (Greek), Venus (Roman), Osiris (Egyptian).

Prompt Element	Lyric Content	Examples from the Lyric
Gothic and conversational tone	The language has a dark, mysterious quality and characters speak to each other.	"In the shadows of the past," "Osiris to Zeus, ancient and wise."
Prepositions of time, place, and direction	Prepositions are used throughout to indicate time, place, and movement.	"Upon Olympus," "In Rome's heart," "Beneath Egypt's sands," "Through the Nile, my myths soar."
Play with themes and rhemes changing their slots	The positions of subjects and predicates are varied for grammatical exposure.	"Zeus speaks," becomes "mighty and tall, Zeus speaks."
Themes and rhemes in bold	Key thematic elements and rhemes are bolded to draw attention to sentence structure.	"**Zeus** speaks with thunder," "**mighty** and tall."

<u>Raw Readability Metric Data of Text 1</u>





Example 2:

Prompt: Please use your creative conversation style to respond to the below prompt: Assume you are a playwright and adopt the Shakespearean style in his tragedies to create a scene between two characters, a male dermatologist with 20 years of experience and a female patient. Generate a dialogue of twenty-four turns between both characters where the dermatologist explains a diagnosis in an empathetic tone. The dialogue must demonstrate the use of technical medical terms, at least six terms, and employ modality and evaluative language to express probability and obligation. Adapt the language to only suit ESL learners in grade 12 and of the C1 (Advanced) language level according to the Common European Framework of Reference (CEFR) where such learners can understand long and complex factual and literary texts, appreciating distinctions of style and can also understand specialized articles and longer technical instructions, even when they do not relate to their field. Incorporate the British and Indian English varieties, the former used by the dermatologist and the latter used by the patient, focusing on their unique lexical and grammatical features. The dialogue must showcase four transformations of active to passive voice in the turns of both the dermatologist and the patient.

Reading Text 5 (C1)

Microsoft Copilot's Response:

Act I, Scene I: The Dermatologist's Chamber

Enter Dr. Reginald Fitzwilliam, a seasoned dermatologist, and Lady Ananya Patel, his anxious patient.

Dr. Fitzwilliam: *Good morrow, fair Lady Ananya. Pray, be seated upon yonder cushioned stool. What ails thee?*

Lady Ananya: *Good morrow, Dr. Fitzwilliam. My skin doth vex me sorely. These crimson patches, they torment my visage.*

Dr. Fitzwilliam: *Fear not, dear Lady. I shall unravel this enigma. Let us delve into the depths of epidermal mysteries.*

Lady Ananya: *Verily, I await your wisdom, noble sir.*

Dr. Fitzwilliam: *First, let us scrutinize these patches. Ah, behold! A cluster of papules, inflamed and obstinate.*

Lady Ananya: *Papules? Pray, what manner of creatures are these?*

Dr. Fitzwilliam: *They are but small elevations, akin to hillocks upon the skin. Their origin lies in blocked sebaceous glands.*

Lady Ananya: *Blocked, you say? Like a clogged conduit?*

Dr. Fitzwilliam: *Indeed, fair Lady. The sebum, trapped within, festers and provokes inflammation.*

Lady Ananya: *And what of these pustules, like miniature Vesuvii?*

Dr. Fitzwilliam: *Ah, pustules—the fiery offspring of papules! Filled with pus, they erupt forth, seeking release.*

Lady Ananya: *Release? Like a tempest unleashed?*

Dr. Fitzwilliam: *Precisely. Now, attend closely. The diagnosis stands thus: Acne Vulgaris.*

Lady Ananya: *Acne? A common malady, is it not?*
Dr. Fitzwilliam: *Indeed, but each case hath its own tale. Yours, fair Lady, is moderate, verging on severe.*
Lady Ananya: *Severe? Pray, what course of action dost thou recommend?*
Dr. Fitzwilliam: *We shall embark upon a regimen most diligent. A portion of benzoyl peroxide, nightly, to quell the rebellion.*
Lady Ananya: *Benzoyl peroxide? Sounds potent.*
Dr. Fitzwilliam: *Indeed, it shall wage war upon the malevolent microbes. And lo, a salve—an antibiotic ointment—to soothe.*
Lady Ananya: *Antibiotics? Like knights in shining armor?*
Dr. Fitzwilliam: *Verily, they shall vanquish the bacterial hordes. And lastly, a gentle cleanser, thrice daily, to purify.*
Lady Ananya: *Purify? Like a sacred ritual?*
Dr. Fitzwilliam: *Precisely. Now, heed this: compliance is thy duty. Obligation rests upon thy fair shoulders.*
Lady Ananya: *Obligation? As solemn as a vow?*
Dr. Fitzwilliam: *Indeed, for the path to clear skin is fraught with trials. And fear not, for I shall guide thee.*
Lady Ananya: *Thou art my beacon, Dr. Fitzwilliam.*
Dr. Fitzwilliam: *Then let us commence this odyssey, fair Lady Ananya. May thy complexion bloom anew.*
Exeunt Dr. Fitzwilliam and Lady Ananya, their fates entwined in the tapestry of dermatological destiny.
Finis.

Tabel A2: Analysis of Prompt Elements Incorporated in Text 5

Prompt Element	Reading Text Content	Example
Shakespearean Style	The text adopts a Shakespearean style with elevated language, archaic expressions, and poetic flair. For instance: "Verily," "Pray," "Thou art," and "Exeunt."	"Verily, I await your wisdom, noble sir."
Technical	The dialogue includes	"They are but small

Prompt Element	Reading Text Content	Example
Medical Terms (at least six)	several technical terms related to dermatology: "papules," "sebaceous glands," "pustules," "benzoyl peroxide," "antibiotic ointment," and "cleanser."	elevations, akin to hillocks upon the skin. Their origin lies in blocked sebaceous glands."
Modality and Evaluative Language	The text employs modality and evaluative language to express probability and obligation. Examples include: "shall," "verily," "precisely," "indeed," and "thy duty."	"Indeed, it shall wage war upon the malevolent microbes."
British and Indian English Varieties	British English leans toward formality and traditional structures, while Indian English incorporates local vocabulary and unique grammatical patterns.	**"Verily"** is an archaic adverb meaning "truly" or "indeed." It's rarely used in modern English but appears in British literature and historical contexts. (- Dr. Fitzwilliam uses "verily" to emphasize his wisdom and formal speech.) - **"Badmash"** is an Indian English word for a hooligan or troublemaker. It's not commonly used outside of India. (- Lady Ananya uses "badmash" to describe someone mischievous or unruly.)

Prompt Element	Reading Text Content	Example
Passive Voice Transformations	The text subtly incorporates passive voice transformations.	*"… The sebum, trapped within, festers and provokes inflammation.", "…pustules—the fiery offspring of papules! Filled with pus, they erupt forth, seeking release."*

Raw Readability Metric Data of Text 5