A Comparative Study of the Readability of English Reading Materials Generated by Different Large Language Models

Xinghan Li and Shu Zhang*

School of Foreign Languages, Chengdu University of Information Technology, Chengdu, 610225, China
*tonyzhang@cuit.edu.cn

Abstract. Large language models are widely used in English language teaching, yet the yet differences in readability among reading materials generated by different large language models remain unclear. This study compared English reading passages generated by three mainstream models—DeepSeek, ChatGPT, and Gemini—across two genres: expository and narrative. Readability was assessed using Flesch Reading Ease and Flesch-Kincaid Grade Level, two standardized metrics widely adopted in educational research. The results revealed systematic differences in the readability of texts generated by the three models. DeepSeek consistently produced the more readable texts, Gemini produced the least readable, and ChatGPT ranked in between. Across all models, narrative passages were found to be significantly more readable than expository ones, indicating that text genre plays a crucial role in determining difficulty levels for language learners. These findings address a critical gap in model selection criteria for AI-assisted reading instruction. They provide empirical guidance for teachers and instructional designers in choosing appropriate AI tools based on target genre and desired difficulty level, contributing to more effective and informed integration of large language models into English language classrooms.

Keywords: Large language models, English reading instruction, Readability, AI-assisted teaching

1 Introduction

Large language models (LLMs) are increasingly used for content generation, tutoring, and educational support.[1] As they are also applied to produce reading materials for language learning, readability has become an important factor in evaluating their suitability for instruction.[2] Therefore, when different models generate passages for the same instructional purpose, variations in readability may affect their suitability for educational use.

Previous studies have explored AI and LLM-based technologies in English teaching, including writing support, material generation, reading assistance, and instructional design.[3] Recent research also suggests that AI-supported practices, such as

B. Kanoksilapatham et al. (eds.), *Proceedings of the 2026 2nd International Conference on Education Reform, Ideology and Politics (ERIP 2026)*, Advances in Social Science, Education and Humanities Research 1028,
https://doi.org/10.2991/978-2-38476-601-7_45

graded reading and adaptive assistance, may reshape learning experiences.[4] Some studies have compared AI-generated and human-written texts in terms of linguistic quality and pedagogical value.[5] However, most focus on single-model performance or AI-human comparisons, while readability differences among texts generated by different LLMs remain underexplored. Therefore, an important question remains: when mainstream LLMs generate English reading passages for the same learners and topics, how do their readability levels differ?

This study compares English reading materials generated by three mainstream models—DeepSeek, ChatGPT, and Gemini—using readability theory and two established metrics: Flesch Reading Ease and Flesch-Kincaid Grade Level. Reading passages are generated in expository and narrative genres based on topics aligned with secondary school English textbooks published by the People's Education Press (PEP) and Foreign Language Teaching and Research Press (FLTRP), covering Grades 8 and 10. By examining readability, consistency, and genre differences, this study aims to identify model-specific readability patterns and provide practical guidance for teachers using AI tools in material preparation.

2 Theoretical Framework

This study is grounded in readability theory, which examines how linguistic features of texts relate to comprehension difficulty. Originally developed to estimate text difficulty through measurable features such as sentence length, word complexity, and syntactic structure,[6] readability theory has been widely applied in language teaching and textbook evaluation to assess text suitability for learners at different proficiency levels.[7] In second language education, it also supports text selection and graded reading design.[4] Since this study compares passages generated by different large language models for the same target learners, readability provides an appropriate framework for evaluating differences in text difficulty.

Among readability assessment methods, formula-based metrics remain widely used because they provide objective and replicable measurements.[8] This study adopts Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) as the core indicators, as both are established formulas for English texts.[9] FRE measures difficulty through sentence length and syllabic complexity, whereas FKGL converts these features into approximate U.S. grade-level equivalents. Together, they support systematic comparison across model-generated passages.

Although readability is also influenced by cohesion, semantic complexity, and background knowledge,[10] readability formulas cannot capture all dimensions of text difficulty.[7] However, FRE and FKGL remain suitable for comparative analysis because they provide transparent and consistent measurements across texts generated by the three selected models.[7]

3 Research Method

3.1 Model Selection

This study selected three mainstream LLMs for comparison: DeepSeek-V3.2, ChatGPT-5.5 Instant, and Gemini 3.5 Flash. These three models are publicly accessible general-purpose LLMs widely used in English language teaching, offering good representativeness. All models were used with their default settings, with web search and deep reasoning features disabled, ensuring that generated outputs were based solely on the models' parametric knowledge.

3.2 Material Generation

The study targeted Chinese students in Grades 8 to 10 as the intended readers. Six topics aligned with mainstream English textbooks (PEP and FLTRP) were selected: three expository topics (environmental protection, internet impact, and balanced diet) and three narrative topics (an unforgettable journey, helping someone, and first day in high school). A standardized prompt was used to generate one passage per topic: "You are an experienced English teacher. Write an English reading passage on [topic]. The target audience is Grade 9 students (CEFR B1 level). The passage should be approximately 350-400 words. Use [genre] style. Do not include any meta-commentary. Just output the passage directly." Each model generated two independent passages per topic, resulting in 36 passages total (3 models × 6 topics × 2 generations).

3.3 Readability Assessment

Two standardized readability metrics were employed: Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). FRE scores range from 0 to 100, with higher scores indicating easier texts. FKGL outputs the corresponding U.S. grade level, with lower values indicating simpler texts. An online readability calculator was used to score each passage, recording both FRE and FKGL values.

3.4 Text Coding

Each passage was assigned a unique identifier following the format "[Model Code]-[Genre Code][Number]". Model codes: D for DeepSeek, C for ChatGPT, G for Gemini. Genre codes: E for expository, N for narrative. Numbers 01–06 indicate the passage index within each condition. For example, G-E02 refers to the second expository passage generated by Gemini.

4 Results and Discussion

4.1 Descriptive Statistics

A total of 36 English reading passages generated by three large language models were analyzed, with each model producing 12 passages (six expository and six narrative). Table 1 presents the descriptive statistics of readability by model and genre.

As shown in Table 1, the three models exhibited clear readability differences. Across both expository and narrative texts, DeepSeek produced the highest readability, Gemini the lowest, and ChatGPT fell in between, as indicated by both FRE and FKGL scores. In addition, narrative passages were consistently more readable than expository ones across all three models.

Table 1. Mean FRE and FKGL Scores by Model and Genre

Model	Genre	n	Mean FRE	SD FRE	Mean FKGL	SD FKGL
DeepSeek	Expository	6	62.11	4.07	7.69	0.72
DeepSeek	Narrative	6	81.05	6.57	4.47	1.23
ChatGPT	Expository	6	51.37	4.19	9.31	0.7
ChatGPT	Narrative	6	74.13	3.06	5.86	0.64
Gemini	Expository	6	47.69	6.62	10.8	0.87
Gemini	Narrative	6	61.66	2.86	8.66	0.62

4.2 Cross-Model Comparison

Figure 1 visualizes the readability comparison among the three models. DeepSeek consistently outperformed the other two models across both genres, while Gemini ranked lowest, and ChatGPT fell in between. This ranking remained stable across expository and narrative texts, suggesting that the readability characteristics of each model are intrinsic and consistent across genres.

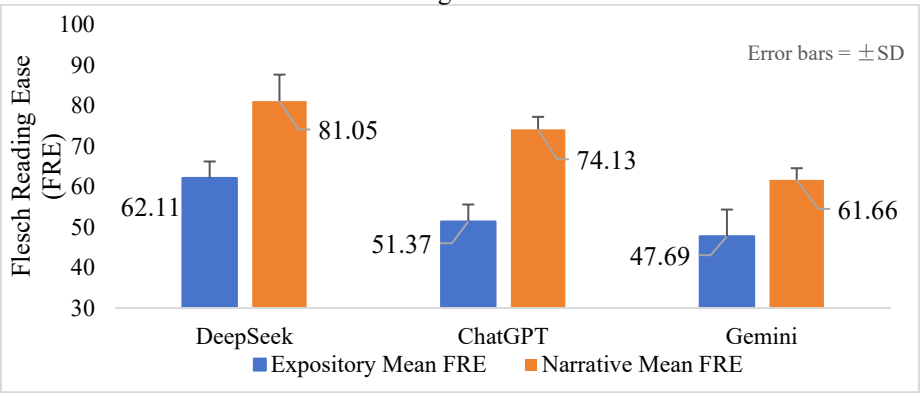

Fig. 1. Bar chart of mean FRE scores by model and genre, with SD error bars

Regarding generation consistency, ChatGPT demonstrated the smallest standard deviations (SD FRE = 3.06 for narrative, 4.19 for expository), indicating the most stable and predictable outputs. Gemini showed the largest standard deviation in expository texts (SD FRE = 6.62), primarily due to one outlier (G-E02) with an exceptionally low FRE score of 38.45 (FKGL = 11.64). DeepSeek exhibited the largest variation in narrative texts (SD FRE = 6.57), with one passage (D-N04) reaching an extremely high FRE of 89.68 (FKGL = 2.85).

4.3 Analysis of Extreme Cases

Three extreme cases were examined qualitatively to understand the textual features underlying the readability scores.

Gemini E02 (most difficult). With an FRE of 38.45 and FKGL of 11.64, this text was the most difficult among all 36 passages. Qualitative analysis revealed the following features: the text contained long sentences (one sentence reached 35 words), dense abstract nouns (e.g., "transformed," "distant relatives"), and lacked direct engagement with the reader. In contrast, ChatGPT and DeepSeek generated passages on the same topic using shorter sentences and first-person plural pronouns ("we," "our"), which significantly lowered reading difficulty.

DeepSeek N04 (easiest). This passage achieved an FRE of 89.68 and FKGL of 2.85, indicating readability at approximately second-grade level. The text consisted of very short sentences (8–10 words on average), high-frequency vocabulary (e.g., "rainy," "walking," "slippery"), and included direct speech (e.g., "Excuse me, ma'am," I said). While highly accessible, this level may be too simplistic for Grade 9 students, suggesting that DeepSeek tends to oversimplify narrative texts.

ChatGPT series (no titles). Unlike DeepSeek and Gemini, which generated titles for most expository passages, ChatGPT produced no standalone titles across all 36 passages. This likely reflects ChatGPT's stricter interpretation of the prompt instruction "Just output the passage directly," treating titles as a form of meta-commentary. Teachers using ChatGPT may need to explicitly request titles or add them manually.

5 Conclusion and Insights

This study compared the readability of English reading materials generated by DeepSeek, ChatGPT, and Gemini and found clear differences among the three models. DeepSeek produced the most readable texts, Gemini the least readable, and ChatGPT fell in between. This pattern remained consistent across expository and narrative genres. Narrative passages were generally more readable than expository ones, confirming the influence of genre on text difficulty. ChatGPT produced the most stable outputs, whereas DeepSeek and Gemini showed greater variation.

These findings suggest that readability differences may relate to model-specific training data patterns and that genre may shape model output tendencies.[11] DeepSeek showed a stronger readability advantage in narrative texts, while Gemini generated relatively denser expository passages.

Practically, the findings support a model–genre matching strategy: DeepSeek may be more suitable for lower-difficulty narratives, ChatGPT for medium-level instructional texts, and Gemini for more challenging expository materials. Teachers should therefore treat LLM outputs as adaptable drafts and refine prompts to better control difficulty.

This study is limited by sample size and text scope. Future research could include larger datasets, additional models, and learner-based evaluation for a broader understanding of AI-generated reading materials.

References

1. Chan, L.Y.: Practical Research on the Application of Large Language Models in Language Teaching in Vocational Colleges. Plastics Packaging 35(6), 392–394 (2025). https://link.cnki.net/urlid/11.3722.TQ.20251215.0850.196.
2. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103, 102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274.
3. Lo, C.K.: What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences 13(4), 410 (2023). https://doi.org/10.3390/educsci13040410.
4. Feng, D.P., Zhang, S.: Research on the Application of Foreign Language Teaching Technology in the Era of Digital Intelligence. Seismological Press, Beijing (2025). https://webpac.tongji.edu.cn/opac/item.php?marc_no=414639457575342b4859707054346561447158504d513d3d&list=1.
5. Rashid, M. M., Atilgan, N., Dobres, J., Day, S., Penkova, V., Küçük, M., Clapp, S. R., & Sawyer, B. D. (2024). Humanizing AI in education: A readability comparison of LLM and human-created educational content. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68(1), 1044–1048. https://doi.org/10.1177/10711813241261689.
6. Flesch, R.: A New Readability Yardstick. Journal of Applied Psychology 32(3), 221–233 (1948). https://doi.org/10.1037/h0057532.
7. Crossley, S.A., Greenfield, J., McNamara, D.S.: Assessing text readability using cognitively based indices. TESOL Quarterly 42(3), 475–493 (2008). https://doi.org/10.1002/j.1545-7249.2008.tb00142.x.
8. DuBay, W. H. (2004). The Principles of Readability. Costa Mesa, CA: Impact Information. https://files.eric.ed.gov/fulltext/ED490073.pdf.
9. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of New Readability Formulas for Navy Enlisted Personnel. Research Branch Report 8–75, Naval Technical Training Command, Memphis, TN (1975). https://stars.library.ucf.edu/istlibrary/56/?utm_source=chatgpt.com.
10. Hong, J.F., Song, Y.T., Zeng, H.Q., Zhang, G.E., Chen, R.L.: A Multilevel Linguistic Feature Analysis of Factors Affecting Chinese Text Readability. Taiwan Journal of Chinese as a Second Language 13, 95–126 (2016). https://doi.org/10.29748/TJCSL.201612_(13).0005.
11. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922.