Proceedings of the Human Factors and Ergonomics Society Annual Meeting
1–8

DOI: 10.1177/10711813241261689
journals.sagepub.com/home/pro
Sage

Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content

Md Mamunur Rashid[1], Nilsu Atilgan[1], Jonathan Dobres[1], Stephanie Day[1], Veronika Penkova[1], Mert Küçük[1], Steven R. Clapp[1], and Ben D. Sawyer[1]

Abstract

Generative AI (GenAI), specifically the Large Language Model (LLM), focuses on creating content like text, images, audio, or other data types similar to that created by humans. Advancements in various LLMs have accelerated AI development, potentially impacting various fields, including education. Can LLMs produce educational content with similar readability features as human-generated reading passages? This study compared readability measures such as reading speed, comprehension, and qualitative characteristics (familiarity, interest, and perceived quality) of 300-word 8th-grade level text passages authored by humans and ChatGPT3.5. We found that ChatGPT3.5-generated passages can be read faster with better comprehension than conventionally human-authored passages. ChatGPT3.5 passages were also rated higher-quality passages and had comparable ratings for interest and familiarity compared to human-authored passages. These findings implicate that the LLM enhances educational materials' usability and effectiveness. More importantly, LLM offers scalability and consistency, catering to diverse learning needs for educators and students.

Keywords

large language model, ChatGPT, readability, educational content, generative AI

Introduction

In recent years, progress in Artificial Intelligence (AI) has gained speed with the advancements in GenAI technologies. The latest release of GenAI technology, ChatGPT3, has quickly drawn the interest of millions of users (Dale, 2021). Other LLMs have also been competing, for example, PaLM (Chowdhery et al., 2022), LaMDA (Touvron et al., 2023), Gemini by Google (Saeidnia, 2023), Claude by Anthropic (Abbas et al., 2024), Ernie by Baidu (Wang et al., 2021), LLaMA by Meta (Touvron et al., 2023), and Bing Chat by Microsoft (Motlagh et al., 2023) are popular among the hundreds of LLMs. This groundbreaking growth has the potential to make a drastic impact in various fields, one of which is education.

GenAI technology can significantly impact educational practices, aiding educators in class design, assessments, and policymaking, while also assisting students with tasks like essay writing and math problem-solving (Haque et al., 2022). GenAI is an emerging technology that may not be entirely reliable all the time, as researchers and users are still exploring and understanding their capabilities (Cao et al., 2023; Wagner & Ertl-Wagner, 2023; Walker et al., 2023). Given the current status of these technologies, this study aims to assess whether LLMs can create reading passages and comprehension questions comparable in quality and difficulty to those produced by human experts.

Generating educational content, such as reading passages, is costly in terms of time and resources due to the expertise required in curriculum and assessment development, and the need to consider factors like length, grade level, and coherence, making it a challenging task (Benjamin & Schwanenflugel, 2010). Developing educational content requires careful consideration to ensure it is written at an appropriate grade level, which can be challenging for teachers due to the significant variation in reading levels among students in a classroom (Connor & Morrison, 2016; Rog & Burton, 2001). Tailoring

[1]University of Central Florida, Orlando, FL, USA

Corresponding Author:
Md. Mamunur Rashid, The Readability Consortium, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816-8005, USA.
Email: mdmamunur.rashid2@ucf.edu

reading passages for accessibility is crucial, particularly for individuals with learning disabilities like dyslexia and ADHD, who often face disadvantages in traditional classroom settings due to materials not being tailored to their needs (Spiel et al., 2014; Vickers, 2010). While awareness of the need for tailored teaching materials is growing, their availability remains limited; however, GenAI offers the potential to generate such materials more quickly and efficiently than humans can after algorithm training and facilitating necessary adjustments (Ahmed & Ganapathy, 2021; Attard & Dingli, 2023).

ChatGPT, a leading GenAI technology, is capable of generating human-like conversational dialogues (Chiu, 2023). It has outperformed other models in various evaluations, including questioning care for myopia (Lim et al., 2023), neurosurgery (Ali et al., 2023), and answering calculus and statistical questions (Calonge et al., 2023). For this study, ChatGPT3.5 was used to create 8th-grade-level reading passages for testing readability metrics. Exploring the use of ChatGPT3.5 for creating reading passages and comprehension questions is crucial due to the importance of having well-prepared material for teaching and research, along with the challenges of generating ideal content efficiently. Based on the literature above, our research question is to find out if LLM (ChatGPT3.5) can generate passages similar to human-generated ones. In the current study, we investigated reading performance both in terms of reading speed and comprehension through both human-generated and AI-generated passages.

Methodology

Stimuli

Human-authored and AI-generated passages have been used to conduct this study.

Human-Generated Content. An expert in education developed human-generated passages at the 8th-grade reading level using the Flesch-Kincaid Grade Level (FKGL) metric. Ten 8th-grade human-generated passages were selected, making sure the passage contents were mixed in the areas of science, history, and biography. See Figure 2 for the complete set of passage topics. Each passage, about 300 (+/−10%) words long and was accompanied by five comprehension questions.

AI-Generated Content. From the hundreds of available LLMs, we used ChatGPT3.5 to create passages and related comprehension questions.

(i) The Technology Used: ChatGPT3.5 is chosen for its availability to the general public, training on massive datasets, user-friendly interface, ability to consider entire dialogue history for corrections (Haque et al., 2022), and its suitability for text generation, including answering factual questions, adhering to specific requirements, and producing text summaries (Tate et al., 2023). Additionally, it can identify and correct errors in generated output (Osmanovic-Thunström & Steingrimsson, 2023), resulting in output that closely resembles human natural language and maintains coherence.

(ii) Prompt Engineering: Ten AI-generated passages were created using ChatGPT3.5 on the same topics as human-written ones, tuning prompts to match human criteria without human input. A prompt for generating four multiple-choice comprehension questions was also created for each passage, specifying that the answers must include three incorrect and one correct option, along with a final main idea question.

(iii) Prompts: Two separate prompts were created through trial and error: one for generating passages on selected topics and another for generating comprehension questions. The prompt for passage generation is,

> Generate 4 interactive passages about "topic name." The passages should contain curious and interesting facts. Each passage should be approximately 75 words. The passages should stand alone. The text must be at grade level 8.

The prompt for question generation is,

> Generate 4 difficult multi-choice comprehension questions for each of the passages below, with 3 incorrect and 1 correct answer. Highlight the correct answer. Don't write the correct answer longer than the other answers. Generate 5th question answering the question: "What is the main idea of the text?

All the passages and questions were generated using the same prompt, only changing the "topic name."

(iv) Quality Checking: AI passages were assessed for quality based on criteria such as topic relevance, passage length, grade-level, and language appropriateness, while questions were evaluated for relevance to the passage, correctness of answers and without multiple correct answers. If criteria were not met, passages or questions were regenerated using the same prompt. An example of passage regeneration due to inappropriate or offensive language in the "Amarna Period" passage:

> Akhenaten looked different from other pharaohs. He had long features and looked more like a woman.

For the questions, in some cases, ChatGPT3.5 gave a correct answer choice in place of one of the incorrect answers. For example, this question was generated for the "Ancient City: Sparta" passage and had two correct answers.

> What was the law of Lycurgus?, A) A set of laws that regulated the lives of the Helots; B) A set of laws that regulated the lives

of the Perioeci; C) Lycurgus' law mandated that all Spartans be educated in the same way (this answer is intended to be incorrect but is actually correct); D) A set of laws that regulated the lives of the Spartans (the correct answer).

Both human- and AI-generated passages were scored within a narrow FKGL range, but AI passages were slightly lower in average grade level than human passages, as indicated by formal testing (8.4 vs. 8.7, $t[9] = 2.32$, $p = .046$).

Participants

A total of 30 native-English speaking participants (Age Mean: 45.76, 14 women), ranging in age from 18 to 75, were recruited through the crowdsourcing platform Prolific. All the participants self-reported having normal (20/20) or corrected normal vision. Participants were compensated for their participation.

Experiment Design

This study aimed to compare human-authored and AI-authored text passages in terms of reading speed, comprehension, and qualitative characteristics. Human authors wrote passages on ten topics, while AI versions were generated using ChatGPT, matching length and grade level. Passages were divided into sets A and B, with each set containing five human and five AI passages without topic repetition, randomizing the relationship between topic and author. Each passage was split into four screens for better visualization. The experiment was conducted online using Pavlovia, an online iterative tool of Psychopy (Peirce et al., 2019, 2023). Participants in the online study were restricted to using desktop or laptop computers to minimize variability in device usage, though a potential variation in lighting conditions might present.

Procedure

Participants in the online experiment calibrated their screens by adjusting a credit card-sized image to match its size on their display to an actual credit card using the arrow button on the keyboard. They were instructed to sit at a 51 cm (approximately arm's length) viewing distance and then presented with an informed consent form. After giving consent, they read the passages silently, aiming for speed without repetition. The passages, presented in a counterbalanced sequence, alternated between human-generated and AI-generated passages until each participant had read ten passages.

Data Preprocessing

The study calculated reading speeds in words per minute (WPM) for each screen based on the number of words and time spent. A small percentage of AI screens (2.0%) were read slower than 100 WPM compared to human passages (2.7%), while a higher percentage of AI screens (11.5%) were read faster than 650 WPM compared to human passages (10.0%). An interquartile range (IQR) analysis suggested upper cutoffs of 789 WPM for AI passages and 635 WPM for human passages, with no applicable lower cutoff. Individual screens were excluded if they were read slower than 100 WPM or faster than 650 WPM based on previous recommendations of WPM measures (Carver, 1990, 1992; Wallace, Treitman, Huang, et al., 2020). Comprehension scores were tightly clustered, and no valid cutoffs were identified for excluding outlier scores. The final sample included 25 participants, with some screens removed based on the specified criteria. The sample was 52% male, with an age range of 26 to 71 and a mean age of 47.9. 80% of the sample used vision correction.

Results

Using the same prompt, the study regenerated passages with ChatGPT3.5 until they reached the desired 8th-grade reading level with expected quality. It took 1 to 7 regenerations per passage, averaging 3.2 times, and 5 to 45 min, averaging 19.5 min, to achieve this. The passages' readability, measured by FKGL, improved from an initial mean of 11.25 to 8.4 after regeneration.

Quantitative Metrics

Reading speed data were analyzed in a linear mixed-effect model that specified reading speed (WPM per screen) as the dependent measure. Passage author (human or AI), passage order, passage topic, screen number, and participant age were specified as fixed effects. The participant was specified as a random effect with a constant slope. The main effect of the author was significant, with AI-authored passages reading faster than human passages on average (304 WPM vs. 249 WPM, $\chi^2 = 124.88$, $p < .001$, Type II Wald chi-square test). Reading speeds for human passages are in line with speeds observed in previous studies (Wallace et al., 2021; Wallace, Treitman, Huang, et al., 2020; Wallace, Treitman, Kumawat, et al., 2020a, b), while speeds for AI passages are faster. Consistent with previous studies on interlude reading (Wallace, Treitman, Kumawat, et al., 2020b), the effect of the screen was significant (Figure 1), with later screens reading faster than earlier screens ($\chi^2 = 87.56$, $p < .001$). Reading speeds for human passages increase as the session continues, while reading speeds for AI passages begin fast and remain so (Figure 1). Passage topics significantly ($\chi^2 = 57.02$, $p < .001$) affected reading speed (Figure 2). There were also significant interactions between author and screen number ($\chi^2 = 8.84$, $p = .003$) and author and topic ($\chi^2 = 32.58$, $p < .001$).

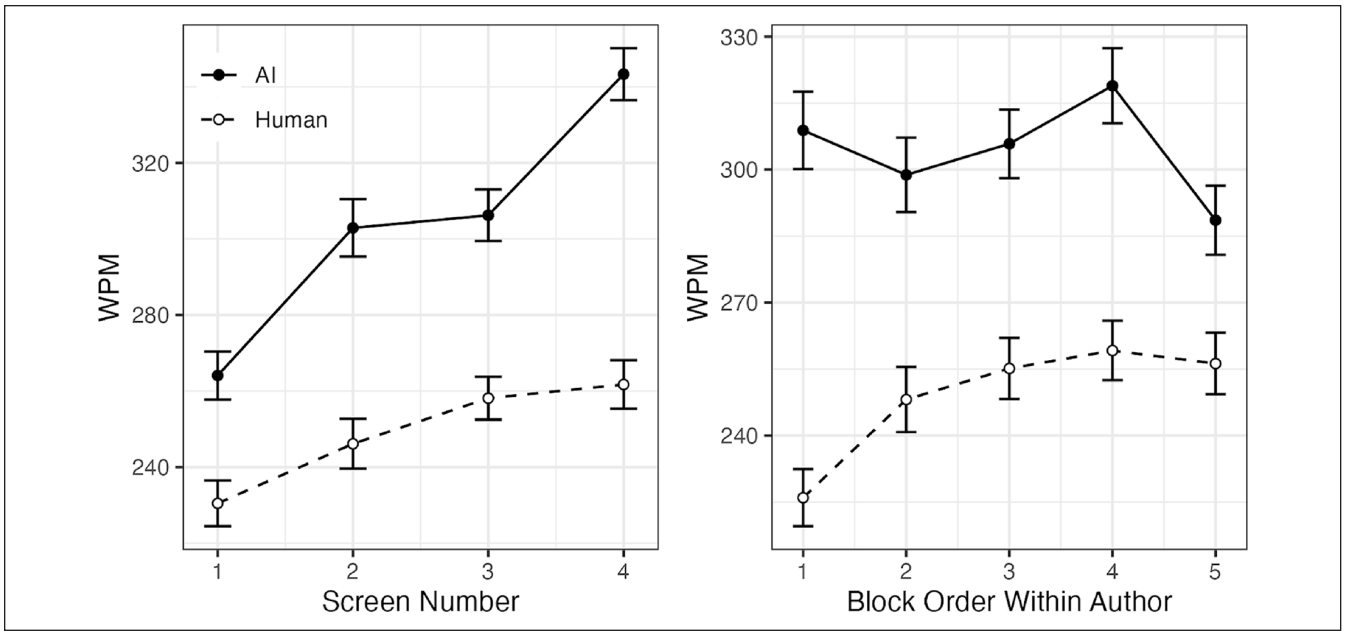

Figure 1. Mean reading speed (WPM) per screen number and author (left) and Mean reading speed per chronological passage (right). Error bars represent ±1 within-participant standard error.

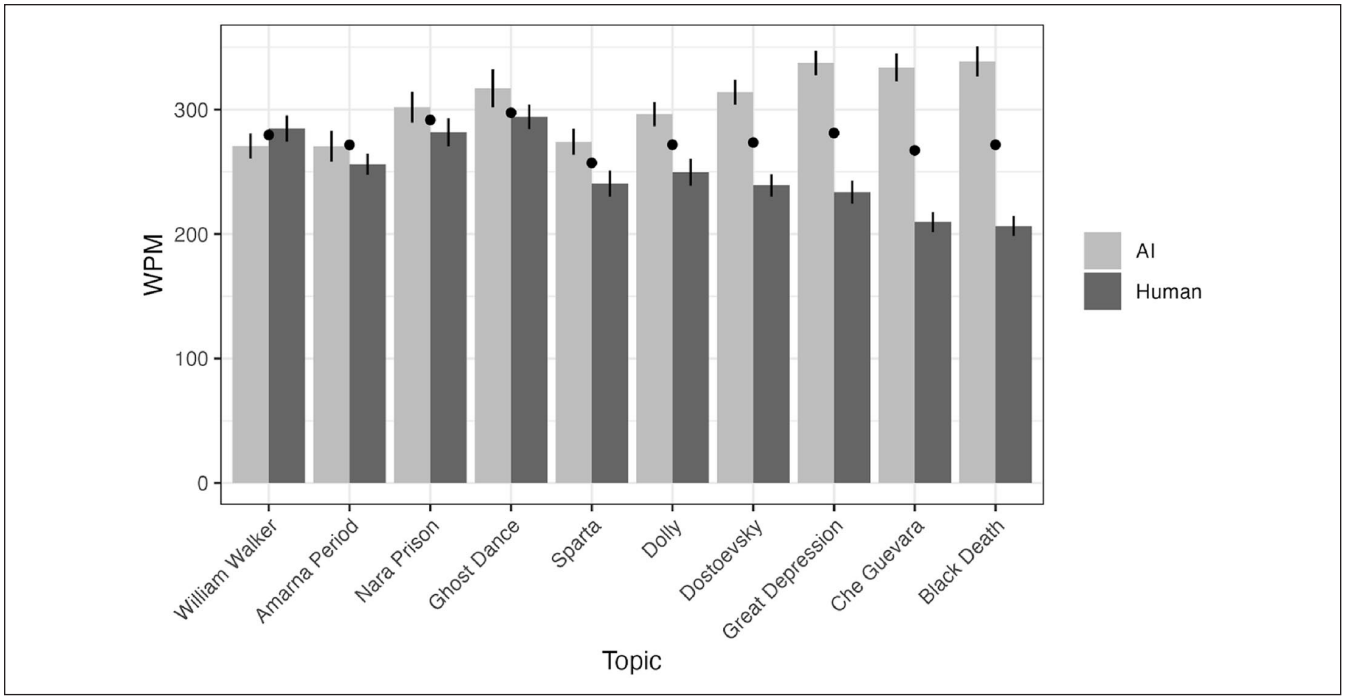

Figure 2. Mean reading speed by passage topic and author. The mean difference between AI and human reading speeds for each passage determines the plot's order. While some human passages have much more pronounced effects on reading speeds, AI passages consistently have fast reading speeds.

Reading comprehension scores were calculated per participant and passage based on response accuracy to the set of five comprehension questions. Scores ranged between 0 and 1 in increments of 0.2. Comprehension scores were first analyzed under a linear mixed-effect model with the following fixed effects: passage author (AI or human), passage topic, block order, participant age, and participant was specified as random effect.

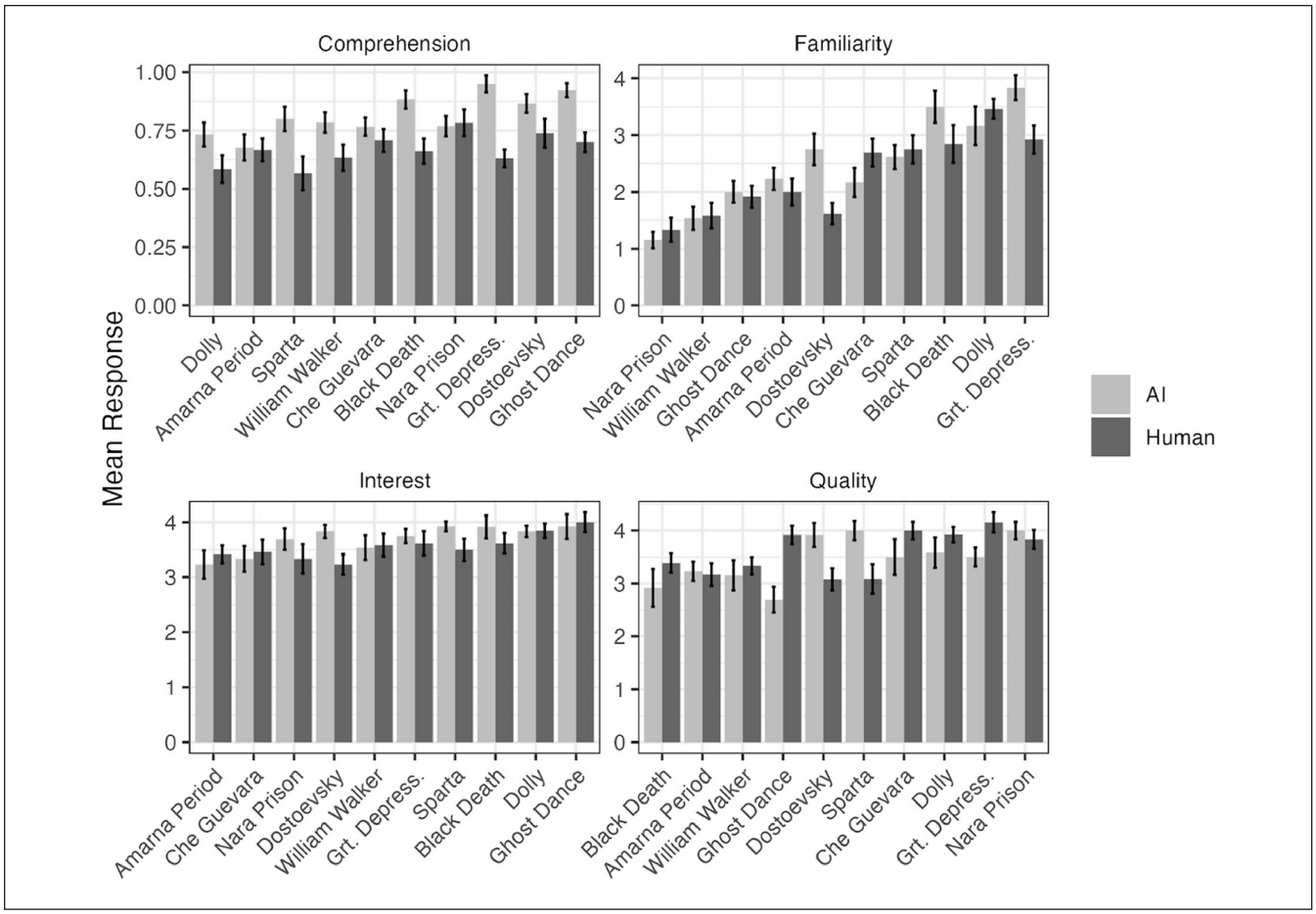

Figure 3. Mean metrics (comprehension, familiarity, interest, and quality) for each passage and author type.

Reading comprehension differed significantly by author, with AI passages having higher scores than human-generated passages (81% vs. 67% accuracy, $\chi^2 = 40.53, < 0.001$). Comprehension also varied significantly by topic ($\chi^2 = 21.11$, $p = .012$; Figure 3). The author and topic also interacted significantly ($\chi^2 = 19.96$, $p = .018$); comprehension scores for human passages were more uniform than for AI passages. The author and age interaction is possibly an artifact ($\chi^2 = 4.31$, $p = .038$); age did not affect the comprehension of human passages, while two participants in their 70s had lower comprehension scores for some AI passages, which appears to drive most of this effect.

Qualitative Metrics

Familiarity, Interest, Quality ratings were analyzed under the similar linear mixed-effect model with the following fixed effects: passage author (AI or human), passage topic, block order, participant age, and participant was specified as random effect.

There was a borderline significant effect of the author on familiarity ($\chi^2 = 3.20$, $p = .07$). Familiarity varies significantly by passage topic ($\chi^2 = 131.61$, $p < .001$). There were significant author and topic interactions ($\chi^2 = 19.90$, $p = .019$); topic and age interactions ($\chi^2 = 21.04$, $p = .012$).

Interest varied significantly by passage topic ($\chi^2 = 17.27$, $p = .045$). There was a significant block order by age interaction ($\chi^2 = 4.99$, $p = .026$). Participants younger than age 48 showed a slightly steeper decline in interest as the session progressed, though the difference is minor.

There was a significant author-by-topic interaction ($\chi^2 = 32.62$, $p < .001$), suggesting that the quality of AI-generated passages varied depending on the topic. There was also a main effect of the passage topic overall ($\chi^2 = 23.87$, $p = .005$). From Figure 3, Ancient Greek City: Sparta, Fyodor Dostoevsky, and Nara Juvenile Prison were rated as high-quality AI-generated passages.

Turing Test

Participants guessed the authorship of passages after reading them, with human authorship as the positive outcome. Hits were defined as correctly identifying a human passage, while misses were deemed as incorrectly labeling a human passage as AI. Correctly identifying an AI passage was a correct rejection, and incorrectly labeling it as human was a false positive.

Sensitivity (d-prime) and bias (c) metrics were calculated for each participant. Wilcoxon signed-rank tests showed that neither d-prime (mean=0.15, p=.12) nor c (mean=−0.10, p=.399) were significantly different from zero. This suggests that participants were unable to distinguish between human and AI passages and were not significantly biased toward either choice, though nominally toward AI.

Discussion

Initially, AI-generated passages varied widely in word count and grade level but were adjusted through multiple regenerations to meet desired specifications. Participants found AI passages comparable in quality to human-authored ones, unable to differentiate between them. The study found that AI-authored passages were read faster and more thoroughly than human-authored passages. However, expert quality assessments suggest that this might be because AI passages had lower information density.

Metrics and Future Directions

The study found that AI-generated passages were read much faster than human-authored passages. This difference could be due to AI passages containing more filler text and repetitive sentences, making them less informationally dense and easier to read. These findings align with earlier studies (Spencer et al., 2019) showing quicker reading of easier texts. Another possible explanation for the higher speed is that AI-authored passages had instances of repetitive sentences, which may increase skimming. For example, both sentences were generated within the same passage.

> The ancient city of Sparta was renowned for its rigid social structure. Sparta was known for its strict laws and customs.

Though the reading speed varied by topic, comprehension accuracy was greater for AI-generated passages. The significant difference between topics implies that the accuracy or coherence of an AI passage depends on the topic; some may be more accurate or coherent than others. After the experiment, our human experts assessed AI passages and questions, finding that in many cases, the correct answer was more distinguishable from the incorrect choices. For example: *"What is the bust of Nefertiti? (A) A new city built during the Amarna period; (B) A famous temple for the god Aten; (C) A type of religious ceremony; (D) A sculpture made from limestone."*

The study found that ChatGPT generated passages outperformed human-generated ones, suggesting the need for further research to explore metrics like enjoyment, immersion, and fatigue to understand AI text quality better. Future studies should also investigate how participants process and evaluate AI-generated passages over longer exposures, focusing on accuracy, coherence, and overall linguistic quality. More research is needed to understand how humans perceive and assess AI-generated language compared to human-generated literature. Comparison between ChatGPT4's, other LLMs and human generated content would be interesting in this domain. Using reinforcement learning to train on diverse passages and questions could further improve AI quality, and studies should involve various reader groups and content lengths for comprehensive insights.

Limitations

ChatGPT3.5 is proficient in generating coherent text but faces challenges in crafting complex comprehension questions and answer choices, which could affect assessment quality. It tends to fabricate information when lacking knowledge (Wagner & Ertl-Wagner, 2023), uses repetitive sentences, and may produce insensitive content. Future research could compare human and AI-generated passages on the same topic and improve randomization of passage topics. Addressing the lower word count and grade level of AI-generated passages compared to human-authored ones could enhance the capabilities of ChatGPT3.5 and other LLMs.

Conclusion

The results suggest that AI can be a valuable tool for developing educational content. While more research is needed to fully utilize AI's potential in content creation, human writers could use AI to support them, saving time and resources. Editing AI-generated content may be quicker than creating new content from scratch, especially when re-leveling material. This technology could also benefit researchers needing stimuli for reading experiments. With further research, AI has the potential to help educators deliver personalized content more efficiently, leading to better student outcomes.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: None of the authors has a financial interest in Adobe.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The funding support for this project was provided by Adobe.

ORCID iD

Md. Mamunur Rashid https://orcid.org/0000-0001-8210-939X

References

Abbas, A., Rehman, M. S., Rehman, S. S. (2024) Comparing the performance of popular large language models on the National Board of Medical Examiners Sample Questions. *Cureus*, *16*(3), e55991. https://doi.org/10.7759/cureus.55991

Ahmed, A. A. A., & Ganapathy, A. (2021). Creation of automated content with embedded artificial intelligence: A study on learning management system for educational entrepreneurship. *Academy of Entrepreneurship Journal*, *27*(3), 1–10.

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Zadnik Sullivan, P. L., Cielo, D., Oyelese, A. A., Doberstein, C. E., Telfeian, A. E., Gokaslan, Z. L., & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery, Publish Ahead of Print*. https://doi.org/10.1227/neu.0000000000002551

Attard, A. E., & Dingli, A. (2023). Automated content generation for intelligent tutoring systems. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds.), *HCI international 2023 posters* (Vol. 1834, pp. 194–201). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35998-9_27

Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly*, *45*(4), 388–404. https://doi.org/10.1598/RRQ.45.4.2

Calonge, D. S., Smail, L., & Kamalov, F. (2023). Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. *Journal of Applied Learning & Teaching*, *6*(2), 346–357 https://doi.org/10.37074/jalt.2023.6.2.22

Cao, J. J., Kwon, D. H., Ghaziani, T. T., Kwo, P., Tse, G., Kesselman, A., Kamaya, A., & Tse, J. R. (2023). Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. *American Journal of Roentgenology*, *221*(4), 556–559 https://doi.org/10.2214/AJR.23.29493

Carver, R. P. (1990). *Reading rate: A review of research and theory* (pp. x, 514). Academic Press.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84–95.

Chiu, T. K. F. (2023). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2253861

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:2204.02311). arXiv. http://arxiv.org/abs/2204.02311

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 54–61. https://doi.org/10.1177/2372732215624931

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118. https://doi.org/10.1017/S1351324920000601

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). "I think this is the most disruptive technology": Exploring sentiments of ChatGPT early adopters using Twitter Data (arXiv:2212.05856). *arXiv*. http://arxiv.org/abs/2212.05856

Lim, Z. W., Pushpanathan, K., Yew, S. M. E., Lai, Y., Sun, C.-H., Lam, J. S. H., Chen, D. Z., Goh, J. H. L., Tan, M. C. J., Sheng, B., Cheng, C.-Y., Koh, V. T. C., & Tham, Y.-C. (2023). Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *eBioMedicine*, *95*, 104770. https://doi.org/10.1016/j.ebiom.2023.104770

Motlagh, N. Y., Khajavi, M., Sharifi, A., & Ahmadi, M. (2023). The impact of artificial intelligence on the evolution of digital education: A comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. *arXiv preprint arXiv:2309.02029*. https://doi.org/10.48550/ARXIV.2309.02029

Osmanovic-Thunström, A., & Steingrimsson, S. (2023). Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals according to the ICMJE criteria? A case study. *Discover Artificial Intelligence*, *3*(1), 12. https://doi.org/10.1007/s44163-023-00055-7

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peirce, J., Pronk, T., & Hirst, R. (2023). *Launch your study on Pavlovia.org*. https://www.psychopy.org/online/usingPavlovia.html

Rog, L. J., & Burton, W. (2001). Matching texts and readers: Leveling early reading materials for assessment and instruction. *The Reading Teacher*, *55*(4), 348–356.

Saeidnia, H. R. (2023). Welcome to the Gemini era: Google DeepMind and the information industry. *Library Hi Tech News*. https://doi.org/10.1108/LHTN-12-2023-0214

Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing*, *32*(3), 603–637. https://doi.org/10.1007/s11145-018-9883-0

Spiel, C. F., Evans, S. W., & Langberg, J. M. (2014). Evaluating the content of individualized education programs and 504 plans of young adolescents with attention deficit/hyperactivity disorder. *School Psychology Quarterly*, *29*(4), 452–468. https://doi.org/10.1037/spq0000101

Tate, T. P., Doroudi, S., Ritchie, D., Xu, Y., & Uci, M. W. (2023). Educational research and AI-generated writing: Confronting the coming Tsunami [Preprint]. *EdArXiv*. https://doi.org/10.35542/osf.io/4mec3

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models (arXiv:2302.13971). *arXiv*. http://arxiv.org/abs/2302.13971

Vickers, M. Z. (2010). *Accommodating college students with learning disabilities: ADD, ADHD, and Dyslexia*. John William Pope Center for Higher Education Policy. https://eric.ed.gov/?id=ED535458

Wagner, M. W., & Ertl-Wagner, B. B. (2023). Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, *75*(1), 69–73 https://doi.org/10.1177/08465371231171125

Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, B. P., Raptis, D. A., & Staubli, S. M. (2023). Reliability of medical information provided by ChatGPT: Assessment against

clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*, *25*, e47479. https://doi.org/10.2196/47479

Wallace, S., Dobres, J., & Sawyer, B. D. (2021). Considering the speed and comprehension trade-off in reading mediated by typography. *Journal of Vision*, *21*(9), 2249. https://doi.org/10.1167/jov.21.9.2249

Wallace, S., Treitman, R., Huang, J., Sawyer, B. D., & Bylinskii, Z. (2020). Accelerating adult readers with typeface: A study of individual preferences and effectiveness. *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–9). https://doi.org/10.1145/3334480.3382985

Wallace, S., Treitman, R., Kumawat, N., Arpin, K., Huang, J., Sawyer, B., & Bylinskii, Z. (2020a). Individual differences in font preference & effectiveness as applied to interlude reading in the digital age. *Journal of Vision*, *20*(11), 412–412. https://doi.org/10.1167/jov.20.11.412

Wallace, S., Treitman, R., Kumawat, N., Arpin, K., Huang, J., Sawyer, B., & Bylinskii, Z. (2020b). Towards readability individuation: The right changes to text format make large impacts on reading speed. *Journal of Vision*, *20*(4), 10.

Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., Feng, S., Shang, J., Zhao, Y., Pang, C., Liu, J., Chen, X., Lu, Y., Liu, W., Wang, X., Bai, Y., Chen, Q., Zhao, L., Li, S., ... Wang, H. (2021). ERNIE 3.0 Titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation (arXiv:2112.12731). *arXiv*. http://arxiv.org/abs/2112.12731