International Journal of Technology in Education and Science

www.ijtes.net

AI vs Human Writers: A Comparative Analysis of Story Rewriting through Readability Metrics

Fahimul Habib [1 *], Md Shafiur Raihan Shafi [2], Muhammad Usama Islam [3]

[1] University of Louisiana at Lafayette, USA, 0009-0001-0166-8024
[2] University of Louisiana at Lafayette, USA, 0000-0003-4479-0617
[3] University of Louisiana at Lafayette, USA, 0000-0003-2080-2484
* Corresponding author: Fahimul Habib (fahimul.habib1@louisiana.edu)

Article Info

Article History

Received:
21 September 2025

Revised:
27 January 2026

Accepted:
19 February 2026

Published:
13 March 2026

Keywords

HCI
Generative AI
User studies
Readability
Verbosity
Lexical density
Lexical diversity

Abstract

This study examines how generative artificial intelligence (GenAI) might be used to rewrite narratives in comparison to texts written by humans, with an emphasis on writing style, readability, and verbosity. GenAI's performance is evaluated by means of quantitative analysis and readability metrics. The results indicate that GenAI often generates writings with higher verbosity and readability scores than stories written by humans. Furthermore, the examination of lexical density and diversity reveals subtle variations in writing styles between human, ChatGPT, and Gemini; GenAI exhibits competitive performance in these metrics. Although the results point to potential applications of GenAI in narrative, more research is needed to determine how effective the technology is when compared to human authors.

Citation: Habib, F., Shafi, M. S. R., & Islam, M. U. (2026). AI vs human writers: A comparative analysis of story rewriting through readability metrics. *International Journal of Technology in Education and Science (IJTES), 10*(2), 360-378. https://doi.org/10.46328/ijtes.5909

Introduction

Significant interest has been garnered by the rise of generative artificial intelligence (GenAI) and its potential to revolutionize the creative industries, especially storytelling (Cheung & Shi, 2025). The intriguing subject of whether GenAI can efficiently rewrite stories to a standard that is on par with or better than that of human specialists is researched in this study and also reported by Habib and El Tarabishi (2024). The research specifically aims on comparing AI-generated texts to texts written by professional authors in order to analyze important metrics like verbosity, readability, and writing skill. The discipline of story writing has been profoundly impacted by the new era of automated content generation which was brought about by the substantial development of generative artificial intelligence (GenAI) (Fang et al., 2023b). This paper dives into investigating the use of GenAI technology to rewrite stories, critically assessing the differences between human-written and AI-generated texts in terms of verbosity, readability, and literary style. The scholarship explores stylistic subtleties, lexical diversity, and the overall communication efficiency of texts created by human writers and sophisticated GenAI models like ChatGPT and Gemini using quantitative analysis and recognized readability criteria. Given the increasing use of AI in the creative and educational sectors, it is critical to comprehend these differences in order to evaluate the usefulness, advantages, and constraints of GenAI in story writing.

The way we write includes different language and style elements that shape stories, like the choice of words, how sentences are put together, the use of figurative language, and the overall voice of the narrative (Al-Alami, 2025). Generative AI systems have been getting better at creating text that fits well with different styles and makes sense in context, changing tone and structure to match various storytelling needs. GenAI works by looking at large sets of existing texts, figuring out the connections between words and phrases so it can create smooth writing that often mimics human styles (Mariani & Dwivedi, 2024). This method based on data allows the AI to mimic stylistic patterns while also sticking to known writing rules, which limits true originality or major creative breakthroughs.

Comparative analyses show that AI-generated narratives are consistent and logically coherent, but they usually don't have the figurative richness and descriptive novelty that human-written texts do (Sardinha, 2023). Human writing frequently incorporates metaphors, idiomatic expressions, and emotional subtleties that are closely tied to personal and cultural backgrounds elements that Generative AI struggles to genuinely replicate because it lacks consciousness and real-life experiences as per the note of Beguš (2024). Additionally, GenAI's reliance on learned patterns might result in outputs that seem formulaic or overly generic, unlike human authors who can bring in their unique voice and creative nuances.

Lexical density and diversity serve as measurable indicators of stylistic complexity. Studies show that earlier GenAI models usually have a smaller vocabulary and a more restricted range of language compared to humans, leading to a tendency for repetitive and formulaic expressions (Durak et al., 2025). Wu and Xu (2025) found that recent GenAI systems, like Gemini, show improved vocabulary and varied phrasing, making them capable of matching human writing in some situations. The progress shows that GenAI is still improving its language skills, but there are still some subtle differences in how deeply it can express ideas.

Readability is a critical factor in narrative writing, as it directly affects reader engagement and comprehension. In order to quantify textual accessibility, readability metrics, including the Flesch-Kincaid Grade Level and Reading Ease scores, assess features such as sentence length, word complexity, and syntactic structure. A consistent finding in quantitative research comparing GenAI-generated narratives and human writing is that AI outputs score higher on readability measures due to their generally clearer sentence structures and more unambiguous diction (Zhao, 2024). The improved readability of AI-generated texts can render them particularly appropriate for audiences who are interested in content that is readily digestible or for preliminary manuscripts that require further enrichment.

On the other hand, the increased intelligibility is frequently accompanied by verbosity, as GenAI texts are more intricate and wordy than those written by humans. This level of verbosity encompasses longer sentences, a higher word count, and occasionally redundant or repetitive expressions that may impede narrative rhythm and conciseness. Excessive elaboration can overshadow narrative momentum, which is a skill that experienced human writers typically calibrate more adeptly to maintain reader interest and story flow, despite the fact that such verbosity can enhance detail and clarity. The polish of AI-generated text in terms of grammar, spelling, and punctuation is typically impressive, largely due to the AI's algorithmic precision in sentence construction and its training on vast corpora with high linguistic standards. However, Yang et al. (2024) says that AI-generated prose frequently fails to incorporate the nuanced stylistic embellishments or emotionally resonant phrasing that human editors and authors intentionally employ to enhance the reader's experience.

Verbosity is the degree of depth and elaboration shown in a story. GenAI's tendency for verbosity may be related to its design, which creates text token-by-token depending on learning probabilities rather than deliberate intents about narrative economy. Using repeated words and phrases more often than usual human writing, this produces material that may be complex yet occasionally repetitive. For example, in AI-generated sections, repeated theme terms like "mystery" or "fear" stand out, which might compromise the authenticity of the text and the reader's immersion.

Furthermore, the AI's lack of thorough memory across long narrative spans causes plot discrepancies and narrative logic holes, problems that human writers usually solve by deliberate storytelling technique. Highlighting the constraints of modern AI creativity and contextual awareness, AI-generated narratives might include plot gaps or character behavior discrepancies that shatter reader suspension of disbelief. Notwithstanding their flaws, GenAI tools provide great value as helpers in idea generating, writing, and breaking writer's block, hence acting as cooperative partners rather than solo creators. GenAI may increase output by quickly generating coherent material that people can hone and modify when directed by knowledgeable human writers who grasp narrative goals and story structure. This symbiosis fits the developing perspective of GenAI as "intelligent assistance" complementing rather than replacing human creative effort.

Method

Research Design

For our research design, we have hired a professional story writer with a background in English literature to write

a subjective story. We then instructed ChatGPT and Gemini (the freely available version) to rewrite the story with no specific instruction was given as to summarizing within certain words, for certain levels, or for specific genres.

Data Analysis

A combination of Python (Van Rossum, 2007), MS Excel, and SPSS(Hinton et al., 2014) was used to analyze the texts generated by these three parties. Descriptive statistics(George & Mallery, 2018) pertaining to usage or parts of speech, word and sentence count, complex word count, word lengths, and sentence lengths were measured. A readability analysis was performed for several readability matrices based on phrase and word level, such as Linsear Write (LW) (Wang et al., 2022), SMOG Index (SMOG) (McLaughlin, 1969), The Coleman-Liau Index (CLI) (Coleman & Liau, 1975), The Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., n.d.), The Gunning Fog Index (GFI) (Gunning, 1968). Furthermore, we analyzed the writing style of three prompts using the concept of Lexical Density and Lexical Diversity.

Linsear Write

Linsear Write is an approach to determine how hard it is to read English text. The U.S. Air Force originally developed it to evaluate the clarity of their technical manuals (Wang et al., 2022). The Linsear Write formula determines the U.S. grade level of a text by analyzing the length of sentences and counting the words that have three or more syllables.

To apply the formula, we need to count all the words that contain more than two syllables while avoiding the usage of specific names, general technical language, or words that combine multiple concepts. Next, we start to tally the words that have one or two syllables, which are referred to as "easy words," and include this count in the overall total weight. Finally, we divide the total number of words by the number.

The formula can be written as:

$$Score(LW) = (ASL + (2 * HDW)) / SL$$

Here,

Score (LW) = Linsear Write readability score
ASL = Average sentence length (the total number of words divided by the number of sentences)
HDW = Number of "hard words" (words with more than two syllables)
SL = Number of sentences in the text

The Linsear Write formula assists in determining the ease or difficulty of reading an English text. This straightforward scoring system assists writers in determining whether their text is appropriate for the intended audience. It provides a general indication of the grade level at which the writing would be comprehensible to the reader. This formula is particularly beneficial for the evaluation of technical or scientific materials.

The Linsear Write score is determined by the U.S. school grade system. A score of 1 indicates that the text is

comprehensible to a first-grade student, a score of 2 is indicative of a second-grade student, and so forth for subsequent grades. For instance, A score of 8 indicates that the text is comprehensible to an eighth-grade pupil, who is approximately 13–14 years old. A score of 12 indicates that the text is appropriate for a high school senior. Texts that receive a score of 16 are considered to be at the reading level of a university senior.

The SMOG Index

The SMOG Index, an acronym for "Simple Measure of Gobbledygook," is a useful instrument for evaluating the complexity of text, particularly in terms of legibility and comprehension (McLaughlin, 1969). It allows writers and editors to ensure that their content is compatible with the literacy level of their intended audience. To be more precise, the SMOG Index offers an estimate of the educational level necessary to completely understand a specific piece of writing. This index is particularly dependable for the assessment of extended texts, including books, reports, and articles. Its credibility is derived from its emphasis on the frequency of polysyllabic words, which are words that contain three or more syllables. These words are known to exacerbate sentence complexity and impede comprehension.

In order to determine the SMOG Index Readability Score of a text, it is necessary to select a sample that includes a minimum of 30 sentences from the document. Next, we must determine the total number of polysyllabic words in the sentences. The square root of the total number of polysyllabic words must be calculated after this measure has been obtained. Finally, the SMOG Index score is obtained by adding 3 to the square root value:

The formula for the SMOG Index: $\sqrt{Number\ of\ Polysyllabic\ Words} + 3$

The final number called the SMOG Index Readability Score, tells how easy or difficult the text is to understand. This score is based on two main things; the first one being the average number of syllables in the words and the second factor is the total number of long words in the sample we have chosen. The SMOG Index serves as a way to measure how complex a text is and gives a rough idea of how many years of schooling someone usually needs to understand a certain piece of writing easily. This is really useful for making sure that written materials match the reading levels of different audiences.

A SMOG Index score ranging from 6 to 7 shows that the text is pretty straightforward and easy to understand, typically for those who have gone through about six to seven years of school. This relates to the higher grades of elementary school, where students have basic reading skills that allow them to understand simple sentences and common words. If the score is between 8 and 9, it means the text is kind of easy to read. This is usually appropriate for individuals who have completed around eight to nine years of schooling, which is about the middle school level. At this point, readers can handle a bit more complex sentences and a wider variety of words, but they still need a text that is clear and straightforward. A score ranging from 10 to 12 indicates that the text has a moderate level of difficulty, making it most appropriate for those who have finished ten to twelve years of formal education, which aligns with the high school level. These texts usually have more complex vocabulary and sentence structures, which require a greater level of reading skills to understand completely. Texts that score 13 or higher on the SMOG Index are seen as difficult and are usually meant for readers who have a college-level education or

more. These texts have a lot of complicated academic language, specific terms, and intricate sentence structures. Because of this, they need strong reading and writing skills, along with an understanding of complex ideas and the specific language used in different fields to be easily understood. The SMOG Index offers a structured way for writers and editors to assess and improve the readability of their writing, making sure it aligns well with the educational level of the audience they are targeting.

The Coleman-Liau Index

The Coleman-Liau Index is a well-known readability formula that measures how easy or hard it is to understand and read a piece of English writing (Coleman & Liau, 1975). It gives a rough idea of the grade level in the U.S. school system that a typical reader would need to reach in order to understand the text well. The Coleman-Liau Index is different from many other standard reading measures because it doesn't count syllables or look for words with more than one syllable. Instead, it is based on looking at two things that can be measured: the number of letters and the number of words in the text. Based on the idea that longer words (which usually have more letters) and longer sentences make reading harder, this design was made.

There are two easy averages that are used to figure out the Coleman-Liau Index readability number. The first is the average amount of letters in 100 words, which shows how long things are and how hard they are to read. The second number shows the average number of sentences per 100 words and shows how long and complicated the sentences are. Since the Coleman-Liau Index looks at characters and phrases instead of sounds, it is a simpler and more computer-friendly way to judge how easy something is to read. This makes it perfect for automatically analyzing digital texts.

By using numbers to show how hard a book is, the index helps writers, teachers, and editors make sure their work is appropriate for the reading level of the people they want to read it. It is especially useful in school settings, where matching the level of difficulty of the text to the reading level of the students is important for helping them understand and learn.

The formula is:

$$Score = 0.0588 * L - 0.296 * S - 15.8$$

Here,

L = Average number of characters per 100 words

S = Average number of sentences per 100 words

The Coleman-Liau Index is a well-known readability formula that measures how easy or hard it is to understand and read a bit of English writing. It gives a rough idea of the grade level in the U.S. school system that a typical reader would need to reach in order to understand the book well. The Coleman-Liau Index is different from many other standard reading measures because it doesn't count syllables or look for words with more than one syllable. Instead, it is based on looking at two things that can be measured: the number of letters and the number of words in the text. Based on the idea that longer words (which usually have more letters) and longer lines make reading

harder, this design was made. There are two easy averages that are used to figure out the Coleman-Liau Index readability number. The first is the average amount of letters in 100 words, which shows how long things are and how hard they are to read. The second number shows the average number of sentences per 100 words and shows how long and complicated the sentences are. Since the Coleman-Liau Index looks at characters and phrases instead of sounds, it is a simpler and more computer-friendly way to judge how easy something is to read. This makes it perfect for automatically analyzing digital texts.

By using numbers to show how hard a book is, the index helps writers, teachers, and editors make sure their work is appropriate for the reading level of the people they want to read it. It is especially useful in school settings, where matching the level of difficulty of the text to the reading level of the students is important for helping them understand and learn.

The Flesch-Kincaid Grade Level

The Flesch-Kincaid Grade Level is a widely utilized readability formula designed to assess the relative difficulty of English-language texts (Kincaid et al., n.d.). It provides an estimate of the U.S. school grade level necessary for a reader to comprehend the material comfortably, making it a valuable tool for educators, publishers, and writers who seek to align their content with the reading abilities of specific audiences.

This formula evaluates text complexity by analyzing two primary linguistic features: sentence length and word difficulty. Specifically, it considers the average number of words per sentence as an indicator of syntactic complexity and the average number of syllables per word as a measure of lexical difficulty. The underlying assumption is that longer sentences often require greater cognitive effort to process, while words containing more syllables tend to be more advanced and less familiar to readers, thereby increasing the overall reading challenge.

Originally developed as part of efforts to improve the accessibility of U.S. Navy technical manuals in the 1970s, the Flesch-Kincaid Grade Level has since become a standard tool in educational settings, government communications, and the publishing industry. It is particularly useful in evaluating whether textbooks, instructional materials, and general publications are appropriately tailored to the intended reader's educational level. By providing a score that directly corresponds to a U.S. grade level, the Flesch-Kincaid Grade Level offers an intuitive and practical means of readability assessment. A lower score indicates simpler, more accessible text suitable for younger readers, while a higher score reflects increased complexity, requiring more advanced reading skills. This capacity to translate linguistic features into educational benchmarks has made the formula an enduring and essential component of readability evaluation across a wide range of fields.

The specific mathematical formula looks like this:

$$FKRA = (0.39 * ASL) + (11.8 * ASW) - 15.59$$

Here,

FKRA is the Flesch-Kincaid Reading Age\newline

ASL is the Average Sentence Length (calculated by dividing the total number of words by the number of sentences)

ASW is the Average number of Syllable per Word (calculated by dividing the total number of syllables by the number of words)

The flowchart and step by step process for calculating FKGL is given in Figure 1.

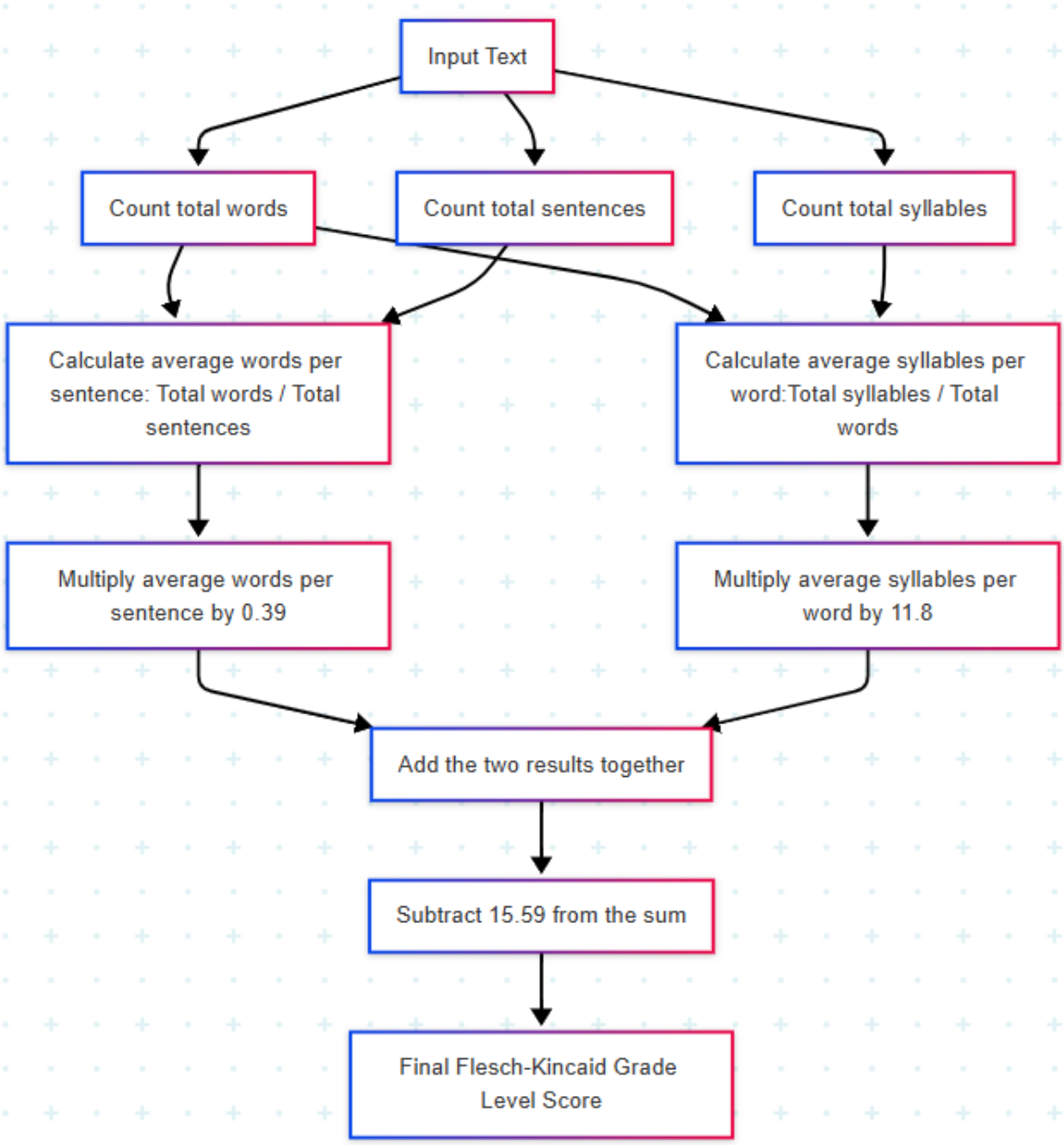

Figure 1. Flowchart of Working Principle of FKGL

The Flesch-Kincaid Grade Level gives a number score that is in line with standards for schooling. For instance, an 8.2 score means that the text is appropriate for users at or above the eighth-grade level, while a 12.5 score means that the material is appropriate for people at the twelfth-grade level or in their first year of college. The method works well to make sure that written materials are easy to read, understand, and right for the level of understanding of the people they are meant for by turning linguistic features into clear, grade-level measurements.

The formula evaluates text difficulty by analyzing two key linguistic features: the length of sentences and the complexity of words. Specifically, it calculates the average number of words per sentence as an indicator of syntactic complexity, alongside the average number of syllables per word as a measure of lexical difficulty. The underlying principle is that longer sentences tend to impose greater cognitive demands on readers, while words with multiple syllables are often more advanced and less familiar, thereby raising the overall reading challenge.
This formula can be used in many real-world situations, not just in school. For example, the United States Army uses the Flesch-Kincaid Grade Level to check how easy it is to read its technical manuals, making sure that important operational and instructional texts are understandable for people with different education backgrounds. In Pennsylvania, the law says that car insurance policies have to be written using this formula, making sure they are easy enough for someone at a ninth-grade reading level to understand. This policy is designed to help people understand financial matters better by making sure that key financial documents are available to everyone. Some other states have also started using the formula to assess how easy legal documents are to read, like business policies, financial disclosures, and consumer contracts, showing a wider trend in public communication.

The Gunning Fog Index

The Gunning Fog Index is a recognized readability formula intended to assess the difficulty of English texts and to estimate the educational level necessary for a reader to easily understand the content(Gunning, 1968). Created by Robert Gunning in the 1970s, this index serves as a crucial instrument for writers, educators, and publishers seeking to match their writings with the reading capabilities of their target audience. The Gunning Fog Index, similar to other readability metrics, produces a numerical score indicative of a U.S. school grade level, providing a definitive and practical standard for evaluating text complexity.

The Gunning Fog Index is calculated using two main linguistic factors: the average sentence length and the ratio of complicated terms in the text. The algorithm specifically evaluates the average number of words per sentence as a measure of syntactic difficulty and designates words with three or more syllables as challenging. The index posits that lengthier sentences and polysyllabic words elevate cognitive load, rendering a text more difficult to read and comprehend.

The Gunning Fog Index algorithm integrates several characteristics into a singular score. It calculates a grade-level estimate by multiplying the sum of the average sentence length and the percentage of complicated words by a fixed factor. This technique facilitates a direct and efficient assessment of readability, rendering it especially valuable in circumstances where clear and accessible information is paramount.

The Gunning Fog Index is frequently utilized for various written products, including digital content, news stories, corporate reports, and instructional publications. In the age of digital media, it has emerged as a mechanism for guaranteeing that online material is available to a diverse audience, including those with differing reading skills. Furthermore, in academic and professional contexts, the index functions as a reference for creating writings that harmonize precision with clarity, thereby improving understanding and engagement.

The Gunning Fog Index, emphasizing sentence structure and lexical complexity, provides a realistic and dependable metric for readability, facilitating successful communication across all domains and audiences:

$$Fog\ Index\ =\ 0.4\ *\ ((words/sentences)\ +\ (percentage\ of\ (complex\ words/\ words))$$

Calculating the Gunning Fog Index requires two critical components of information. The average sentence length is determined by dividing the total number of words by the total number of sentences, providing a measure of syntactic complexity. The proportion of complex or polysyllabic words is assessed. In this index, complex words are defined as those with three or more syllables, excluding common suffixes like -ed, -es, and -ing, as well as proper nouns and well-known compound words. Furthermore, the use of highly technical or specialized terminology can elevate the perceived complexity of the text, as these terms may create comprehension challenges for general audiences.

The Gunning Fog Index formula quantitatively assesses a text's readability by incorporating sentence length and word complexity to generate an overall score. This score reflects the years of formal education generally necessary for a reader to understand the material upon initial reading. It is generally advised that public communication texts target a Fog Index score of 7 to 8 to maintain accessibility for most adult readers. A score above 12 indicates that the material likely poses considerable difficulties for numerous readers and may necessitate advanced literacy skills for effective comprehension.

Classic children's literature, exemplified by Charlotte's Web by E.B. White, typically attains a Fog Index of approximately 6, signifying its accessibility to middle-grade readers. In contrast, prominent publications like The Economist or Scientific American often produce Fog Index scores between 12 and 14, indicating a more complex and intellectually rigorous style aimed at educated readers. A Fog Index of 12 indicates that the text is appropriate for readers with a proficiency level comparable to that of a U.S. high school senior, generally aged 17 to 18 years. The Gunning Fog Index provides a standardized metric that emphasizes both sentence structure and lexical complexity. This tool is beneficial for writers, educators, and policymakers aiming to achieve a balance between precision and clarity, ensuring effective and accessible written communication for the intended audience.

The Flesch Reading Ease Formula

The Flesch Reading Ease Formula provides a simple method for assessing the reading level of a text. This tool is highly reliable, providing accurate results with minimal need for detailed verification. Initially developed for educational purposes, it has evolved into a widely utilized resource among various U.S. government agencies, organizations, and businesses. This formula assesses the readability of a piece of English writing (Flesch, 1948). To apply the formula, one must first quantify the number of words, sentences, and syllables present in the text. Subsequently, the following formula can be utilized to compute the Flesch Reading Ease score:

$$Reading\ Ease\ =\ (206.835\ -\ (1.015\ *\ (words/sentences))\ -\ (84.6\ *\ (syllables/words))$$

The Flesch Reading Ease score is a popular tool used to evaluate how easy it is to read English texts. Created in 1948 by Rudolf Flesch, who was a key figure in readability research, this formula offers a structured method to

assess how easily a text can be comprehended by its target readers (Flesch, 1948). The Flesch Reading Ease score is determined by looking at two key elements of language: the average syllables in each word and the average number of words in each sentence. By concentrating on these aspects, the formula reflects both the vocabulary and sentence structure of a text, providing a useful way to gauge how easy it is to understand.

The score is shown as a number between 0 and 100, where higher numbers mean it's easier to read. Texts that get better scores usually use shorter words and simpler sentences, which makes them easier for everyone to read. A score between 60 and 70 is usually seen as acceptable for most types of public communication because it balances clarity and sophistication pretty well.

Scores that fall between 90 and 100 are considered very easy to read and are usually appropriate for materials aimed at younger readers or those with basic literacy skills. Texts that score between 80 and 89 are seen as easy and suitable for a wide general audience. A score between 70 and 79 shows that the text is pretty easy to read, while scores from 60 to 69 are considered standard and suitable for high school readers.

As scores go down, the difficulty goes up: texts that score between 50 and 59 are pretty tough, those between 30 and 49 are seen as difficult, and materials scoring between 0 and 29 are considered really confusing and hard for most readers to understand without some extra knowledge or reading them multiple times.So, the Flesch Reading Ease score is not just a number; it’s also a helpful guide for writers, editors, and teachers who want to adjust their texts to fit the reading skills of their audience. By choosing the right words and organizing sentences well, writers can make their work easier to read and communicate their ideas clearly.

The Automated Readability Index

A well-known way to figure out how easy or hard it is to read English writing is to use the Automated Readability Index (ARI) (Smith & Senter, 1967). Researchers from the U.S. Air Force made it in the 1960s to help figure out how easy it is to read scientific papers and other kinds of learning materials. The ARI uses two factors to determine the level of difficulty of a piece of writing: the average number of characters in each word and the average number of words in each sentence. This grade level score tells you what grade level the work is at.

The ARI looks at two things to see how easy a text is to read: the average number of characters in a word and the average number of words in a sentence. To find the ARI, use this formula:

$$ARI = (4.71 * (characters/words)) + (0.5 * (words/sentences)) - 21.43$$

To calculate the ARI, initially tally the total amount of characters, words, and sentences in the text. Subsequently, insert these values into the formula to compute the ARI score. This score indicates the grade level required to comprehend the material. The ARI quantifies text complexity and indicates its readability level. Dr. Rudolf Flesch, who previously devised the Flesch Reading Ease formula, established the Automated Readability Index (ARI) in the 1960s as an alternative. Currently, the ARI is widely employed by educators, academics, and professionals to assess diverse written content.

The ARI score indicates the corresponding grade level of the text, ranging from grade 1 to grade 14 or above. If the ARI score is 6.0, the text corresponds to a 6th-grade reading level. The scale aligns with the U.S. education system, wherein grade 1 represents the inaugural year of elementary school and grade 14 signifies the concluding year of college.

Writing Style Analysis

Two ways to look at how people write are lexical density and lexical diversity. These ways help us figure out how hard someone's writing is and how many different words they use. This test checks how much information is packed into each word, and this test checks how many different words are used in the text. You can use these tools to learn more about how people write or to compare different types of writing. But they don't say how good the writing style is or how easy the text is to read (Halliday, 1995; Johnson, 1944).

The number of words in a piece of writing is called its lexical density. Among the words in the text, it figures out how many "content words" (like names, verbs, adjectives, and adverbs) are used. Words that are used a lot in a writing generally mean that it is formal or full of facts and ideas. Most of the time, writing that is less dense in words means it is more relaxed and chatty. Most of the time, writing that is less dense in words means it is more relaxed and chatty (Halliday, 1995).

Lexical variety is the number of unique words a piece of writing uses. To find it, divide the number of unique words by the total number of words. A lot of different words used shows that the writer has a big language and doesn't use the same words over and over. A low vocabulary variety, on the other hand, could mean that the text repeats words a lot, only uses certain terms for a certain topic, or doesn't have many words to choose from. The overall research design process for readability is illustrated in Figure 2 and Figure 3 illustrates the design process for writing style.

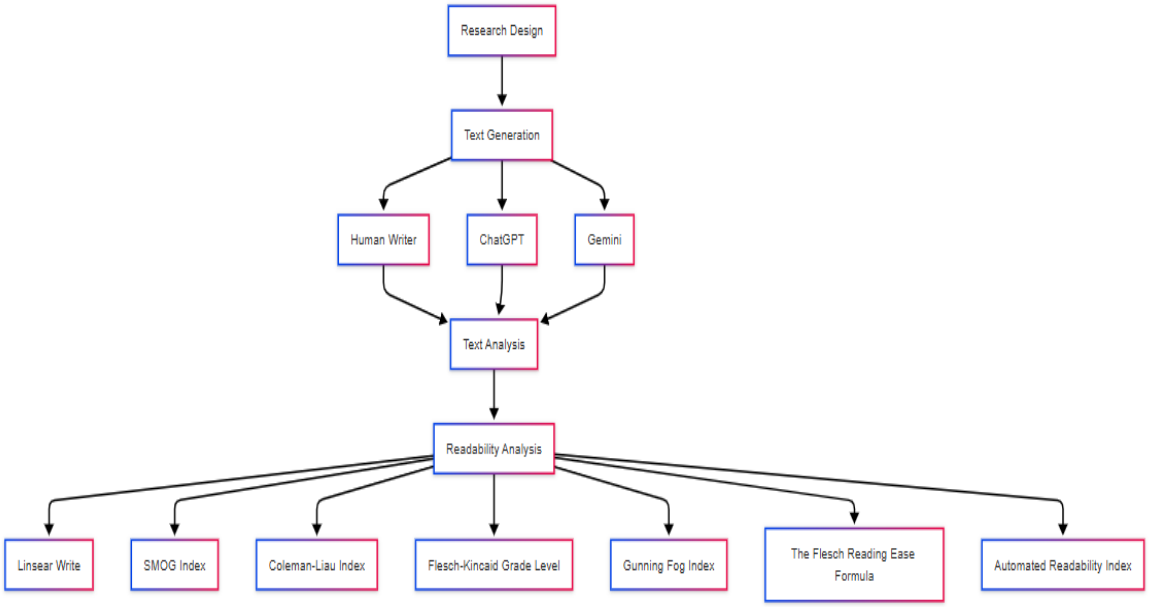

Figure 2. Research Design for Measuring Readability

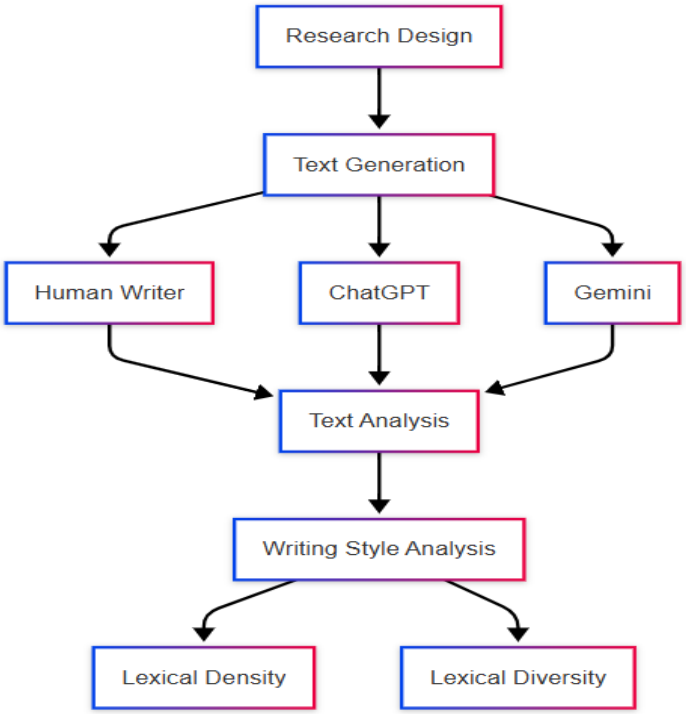

Figure 3. Research Design for Writing Style Analysis

Results

We observe a decreasing word count while analyzing the text's descriptions while shifting from human to GenAI. This instance points towards the enrichment of the verboseness of the texts as the text shifts from human to GenAI. We note for human, at the peak of the graph, 25.85% of words in the text comprised of at least 3 letters while 0.62% of words in the text had 12 letters as the highest lettered word length. As per sentence length, 9 sentences in the text had 16 words on one spectrum and on the other spectrum, 1 sentence had 54 words. For ChatGPT, 23.39% of words in the text comprised of at least 3 letters while 0.4% of words in the text had 13 letters as the highest lettered word length. As far as sentence length is concerned, 8 sentences in the text had 16 words on one spectrum, and on the other spectrum, 1 sentence had 30 words. The details are reported in Table 1.

Table 1. Descriptive Statistics of Story Texts Generated by Human, ChatGPT and Gemini

Descriptive Analysis	Human	ChatGPT	Gemini
Word Count	975	868	795
Sentence Count	69	59	52
Complex Word Count	88	106	137
Syllable Count	1358	1297	1305
Avg. Word Length	4.41	4.72	5.08
SD of Word Length	2.38	2.6	2.9
Avg. Sentence Length	14.13	14.68	15.1
SD of Sentence Length	7.47	5.28	6.91

We have reported the data on several readability measures along with scores, difficulty level, grade level, and age range in Table 2.

Table 2. Readability Analysis of Story Texts Generated by Human, ChatGPT and Gemini

Readability	Writer	Score	Reading Difficulty	Grade Level	Age Range
LW	Human	7.43	Average	7th	12 – 13
	ChatGPT	8.4	Average-Slightly Difficult	8th	13 – 14
	Gemini	10.5	Fairly Difficult	11th	16 – 17
SMOG	Human	6.99	Average	7th	12 – 13
	ChatGPT	8.42	Average-Slightly Difficult	8th	13 – 14
	Gemini	9.81	Somewhat Difficult	10th	15 – 16
CLI	Human	7.95	Average-Slightly Difficult	8th	13 – 14
	ChatGPT	9.86	Somewhat Difficult	10th	15 – 16
	Gemini	12	Difficult	12th	17 – 18
FKGL	Human	6.36	Fairly Easy	6th	11 – 12
	ChatGPT	8.04	Average-Slightly Difficult	8th	13 – 14
	Gemini	10.02	Somewhat Difficult	10th	15 – 16
GFI	Human	9.3	Slightly Difficult	9th	14 – 15
	ChatGPT	11	Fairly Difficult	11th	16 – 17
	Gemini	12.8	Very Difficult	College	18 – 20
FREF	Human	75	Fairly Easy	7th	12 – 13
	ChatGPT	64	Standard	8th – 9th	13 – 15
	Gemini	51	Fairly Difficult	10th – 12th	15 – 18
ARI	Human	7.04	Average	7th	12 – 13
	ChatGPT	8.79	Slightly Difficult	9th	14 – 15
	Gemini	10.93	Fairly Difficult	11th	16 – 17

It's interesting to note that, in all formulae across the spectrum discussed in our research, GenAI's score on readability has a higher performance than that of a story written by Human, thus making the writeup a bit difficult from the preceding human-based writeup.

Table 3. Writing Style Analysis of Story Texts Generated by Human, ChatGPT and Gemini

Writing Style	Style Analysis Metrics	Human	ChatGPT	Gemini
Lexical Density	Test Score	58.2%	60%	64.3%
	Density Range	50% - 59%	60% - 69%	60% - 69%
	Scale (Density)	Above Average	Moderately High	Moderately High
Lexical Diversity	Test Score	41.1%	46.6%	53.2%
	Density Range	40% - 49%	40% - 49%	50% - 59%
	Scale (Density)	Average	Average	Above Average

We observe the story written by Human had a density range within 50%-59% which has a balance on accessibility and detail. The write-up caters to wider audience, and it provides a substantial content but skips overwhelming readers with unknown and specialized type of language. As per diversity, this writing style uses a wide range of words. This type of writing often finds a way to balance the introduction of new terms with the reinforcement of already-established ones. This can suggest that the writer is avoiding repetition by utilizing a variety of words and idioms. The results are reported in Table 3.

Discussion

The comparison between stories written by humans and those generated by GenAI shows clear differences in how wordy they are, how easy they are to read, and their writing styles. This helps us understand how AI writing models are improving and where they still have limitations. This part goes over the results in detail, focusing on the main metrics like word and sentence analysis, readability scores, and writing style indices.

Text Length and Verbosity

We noticed that the word count consistently decreases when comparing human-written text, which has 975 words, to the text generated by ChatGPT at 868 words, and then to the text generated by Gemini, which has 795 words. This pattern shows that while GenAI models are good at summarizing stories, they also tend to use more complicated sentences and words. The average sentence length goes up across the three versions: 14.13 words in human writing, 14.68 in ChatGPT, and 15.1 words in Gemini. This gradual increase shows that even though GenAI lowers the overall word count, it creates longer sentences, which might influence the flow and tempo of the story. The standard deviation of sentence lengths shows some interesting differences, with human-written text having more variety (7.47) compared to ChatGPT (5.28) and Gemini (6.91). This suggests that human writers have a more flexible narrative style, while AI tends to stick to a more regular structure.

The length of words tends to follow a similar pattern. The average word length increases from 4.41 characters in human text to 4.72 in ChatGPT, reaching a high of 5.08 in Gemini. Also, the standard deviation in word length goes up, which suggests that GenAI systems, especially Gemini, tend to use longer and more diverse words. The way words are stretched out adds to the complexity of AI outputs, which is shown in the readability metrics.

Readability Metrics

The readability assessment, which uses several well-known formulas like Linsear Write, SMOG Index, Coleman-Liau Index, Flesch-Kincaid Grade Level, and Gunning Fog Index, consistently indicates that AI-generated texts have higher scores, suggesting they are more difficult to read. This might seem a bit surprising at first, especially since AI focuses on being clear. However, it highlights how these models tend to create text that, even though it's grammatically correct, ends up being more wordy and complex in structure. For instance, when applying the Linsear Write formula, text written by humans scores 7.43, which is appropriate for 7th-grade readers aged 12 to 13. In comparison, ChatGPT scores 8.4, and Gemini scores 10.5, indicating that the latter is suitable for 11th-

grade students aged 16 to 17. This upward trend is regularly shown in other equations. The SMOG Index goes from 6.99 for humans to 9.81 for Gemini, and the Coleman-Liau Index ranges from 7.95 to 12, indicating that Gemini's output is around college-level difficulty.

The Flesch-Kincaid Grade Level shows a similar trend, increasing from 6.36 for human writing to 10.02 for Gemini. This suggests that human writing is still easy for middle school readers, while AI-generated texts are more suited for high school juniors and seniors or early college students. The Gunning Fog Index highlights the difference in complexity between human writing and AI-generated text. Human writing scores a 9.3, which is at a 9th-grade level, while ChatGPT and Gemini score 11 and 12.8, respectively. Interestingly, Gemini's score is getting close to the level of difficulty usually found in academic papers meant for young adults between 18 and 20 years old.

The trends indicate that GenAI models frequently produce narratives that are more challenging to comprehend and denser, despite seeming fluent and grammatically accurate on the surface. The intricacy may enhance the content's appeal to older audiences, however also poses issues for ordinary readers, particularly in educational or public communications. This heightened density frequently arises from extended sentences and a greater ratio of intricate vocabulary, hence augmenting the cognitive burden necessary for understanding. Although this may augment the narrative's depth, it can also impede the speed and accessibility crucial for captivating a wide audience. Moreover, the propensity for verbosity and redundancy in GenAI outputs might undermine clarity, hindering users' ability to seamlessly grasp essential concepts. Resolving these challenges is essential for GenAI systems to function well in educational, journalistic, or public service settings where readability is crucial.

Writing Style: Lexical Density and Diversity

Exploring the aspects of writing style, particularly focusing on lexical density and diversity, is essential for enhancing the richness and complexity of text. By analyzing the variety of vocabulary and the frequency of word usage, one can better understand how these elements contribute to effective communication and engagement in writing. The analysis of writing style through metrics like lexical density and diversity helps to highlight the differences in quality between narratives created by humans and those generated by AI. The lexical density, which looks at the ratio of content-carrying words like nouns, verbs, adjectives, and adverbs, shows a steady increase across the samples: 58.2% in human text, 60% in ChatGPT, and 64.3% in Gemini. Based on what I've learned, human writing typically sits in the "above average" density range (50%–59%), which means it strikes a good balance between being informative and easy to read. On the other hand, both AI models show "moderately high" density, which could improve the amount of information but might also overwhelm readers.

Lexical diversity measures how rich a vocabulary is by looking at the ratio of unique words to the total number of words, which helps to further distinguish the samples. The human text scored 41.1%, which falls into the "average" range. This indicates a good mix of repetition and variety, a quality seen in skilled narrative writing that helps to reinforce themes without being repetitive. ChatGPT has a score of 46.6%, which is still considered average but shows some potential for more variety. On the other hand, Gemini scored 53.2%, placing it in the

"above average" category and suggesting it has a wider vocabulary. Having more diversity in writing can make it better, but too many different words can confuse things, especially in stories where sticking to a theme is important.

Implications for AI Storytelling

The results highlight the potential and the existing limitations of GenAI systems when it comes to creative writing tasks. On one hand, texts generated by AI show strong performance in terms of lexical density and diversity, getting close to or even exceeding human standards. This shows how the models are getting better at copying the complex style that comes with professional writing. However, the high readability scores point out a continuing issue: GenAI systems tend to prefer complex syntax and lengthy structures instead of clear and straightforward writing. The longer sentences, more complex words, and increased word lengths make the text harder to read, even though it stays grammatically correct and stylistically fluent.

Additionally, even though GenAI's increased use of complex vocabulary and varied language indicates a level of sophistication, it might unintentionally make the story harder to follow and understand, especially for general audiences or younger readers. This tension between richness and readability shows the basic trade-offs in AI writing: the push for detailed content versus the necessity for clarity and engagement. The story written by humans shows a good mix of complexity and variety, something that AI models still haven't quite mastered. Writers, using their own experiences and careful style choices, effectively handle wordiness, rhythm, and theme to create stories that connect with readers on both emotional and intellectual levels. On the other hand, GenAI models work by making predictions based on probabilities and recognizing patterns. They are good at sounding fluent on the surface, but they have a hard time with creating deeper connections in narratives and making things easier to understand.

Still, the advancements shown by tools such as ChatGPT and Gemini indicate that AI narrative skills are improving. As large language models keep developing, it will be really important to improve how they balance complexity and clarity. Future development should aim to help AI systems change their writing styles based on what the audience needs. This could involve adding features that adjust readability or allow users to set their preferred difficulty levels.

Conclusion

In comparison to texts created by human writers, the study shows that GenAI, as represented by ChatGPT and Gemini models, typically produces information with a higher verbosity. Additionally, across a range of reading metrics, including Linsear Write, SMOG Index, and Flesch-Kincaid Grade Level, the AI-generated stories have consistently higher readability scores than those authored by humans. Nuanced variations can be seen when writing style is analyzed using metrics like lexical diversity and density. In summary, GenAI demonstrates encouraging potential in rewriting stories with improved verbosity and competitive readability metrics, its efficacy in storytelling still needs investigation and improvement. Subsequent investigations may explore user satisfaction

and user experience studies alongside optimizing LLMs related to narrative creativity while preserving human-like characteristics that appeal to a wide range of viewers.

References

Al-Alami, S. E. (2025). Factors affecting language and narrative styles in prose fiction: A stylistic perspective. *Forum for Linguistic Studies, 7*(2). https://doi.org/10.30564/fls.v7i2.8144

Beguš, N. (2024). Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-03868-8

Castro, C. D., & Halliday, M. A. K. (1995). *An introduction to functional grammar*. *Language, 71*(4), 831. https://doi.org/10.2307/415759

Cheung, L. M. E., & Shi, H. (2025). Co-creating stories with generative AI. *Australian Review of Applied Linguistics*. Advance online publication. https://doi.org/10.1075/aral.24101.che

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283–284. https://doi.org/10.1037/h0076540

Durak, H. Y., Eğin, F., & Onan, A. (2025). A comparison of human-written versus AI-generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini, and BingAI. *European Journal of Education, 60*(1). https://doi.org/10.1111/ejed.70014

Fang, X., Ng, D. T. K., Leung, J. K. L., & Chu, S. K. W. (2023). A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies, 28*(11), 14361–14397. https://doi.org/10.1007/s10639-023-11741-5

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

George, D., & Mallery, P. (2018). *IBM SPSS statistics 25 step by step*. Routledge. https://doi.org/10.4324/9781351033909

Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill.

Habib, M. Y., & ElTarabishi, M. (2024). The role of artificial intelligence in shaping the future of media production and the application of algorithm bias theory in storytelling. *Journal of Media and Interdisciplinary Studies, 3*(8), 1–28. https://doi.org/10.21608/jmis.2024.299137.1033

Hinton, P., McMurray, I., & Brownlow, C. (2014). *SPSS explained*. Routledge. https://doi.org/10.4324/9781315797298

Hodges, J. L., Morse, P. M., & Kimball, G. E. (1952). Methods of operations research. *Journal of the American Statistical Association, 47*(258), 315. https://doi.org/10.2307/2280756

Johnson, W. (1944). A program of research. *Psychological Monographs, 56*(2), 1–15. https://doi.org/10.1037/h0093508

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel*. U.S. Navy.

Mariani, M., & Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research, 175*, 114542.

https://doi.org/10.1016/j.jbusres.2024.114542

McLaughlin, G. H. (1969). SMOG grading—A new readability formula. *Journal of Reading, 12*(8), 639–646.

Nick, T. G. (2007). Descriptive statistics. In *Methods in molecular biology* (pp. 33–52). https://doi.org/10.1007/978-1-59745-530-5_3

Python. (n.d.). In *Alalqab.com*. Retrieved March 5, 2026, from https://alalqab.com/en/Python_(programming_language)

Sardinha, T. B. (2023). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics, 4*(1), 100083. https://doi.org/10.1016/j.acorp.2023.100083

Smith, E. A., Senter, R. J., & Grether, W. F. (1967). *Automated readability index*. Aerospace Medical Research Laboratories.

Van Rossum, G. (2007). Python programming language. In *Proceedings of the USENIX Annual Technical Conference*. USENIX Association.

Wang, S., Liu, X., & Zhou, J. (2022). Readability is decreasing in language and linguistics. *Scientometrics, 127*(8), 4697–4729. https://doi.org/10.1007/s11192-022-04427-1

Wu, Q., & Xu, A. (2025). Poe or Gemini for fostering writing skills in Japanese upper-intermediate learners: Uncovering the consequences on positive emotions, boredom to write, academic self-efficacy, and writing development. *British Educational Research Journal*. Advance online publication. https://doi.org/10.1002/berj.4119

Yang, S., Chen, S., Zhu, H., Lin, J., & Wang, X. (2024). A comparative study of thematic choices and thematic progression patterns in human-written and AI-generated texts. *System*, 103494. https://doi.org/10.1016/j.system.2024.103494

Zhao, D. (2024). The impact of AI-enhanced natural language processing tools on writing proficiency: An analysis of language precision, content summarization, and creative writing facilitation. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-024-13145-5