International Journal of User-System Interaction 17 (2) 2024, 39-49

Comparative Analysis of Human and AI-Generated Stories Using Rhetorical Devices

Cristian Cordoș[1], Ștefan Trăușan-Matu[1,2,3]

[1] National University of Science and Technology Politehnica Bucharest
313 Splaiul Independentei, Bucharest, Romania
[2] Research Institute for Artificial Intelligence
13 Calea 13 Septembrie, Bucharest, Romania
[3] Academy of Romanian Scientists
54 Splaiul Independenței, Bucharest, Romania
E-mail: ioan.cordos@stud.acs.upb.ro, stefan.trausan@upb.ro

Abstract. This paper presents the results of a study evaluating the stylistic and aesthetic differences between texts authored by humans and those generated by artificial intelligence (AI), specifically ChatGPT 4.0. The research focuses on comparing rhetorical devices such as anaphora, polyptotons, epistrophe, alliteration, sibilance, consonance, and assonance. AI stories were created using prompts, some designed by the authors and others generated by ChatGPT itself. Rhetorical devices were analyzed using Python programs, with counts normalized by word count for fair comparison. Initial analysis showed AI-generated stories were poorer in rhetorical devices, but normalization revealed similar counts between human and AI texts. AI-generated stories demonstrated stylistic richness comparable to human-authored ones by replicating specific linguistic features of human writing.

Keywords: Artificial Intelligence; Natural Language Processing; chatbots; Large Language Models; text analysis; rhetorical devices

DOI: 10.37789/ijusi.2024.17.2.1

1. Introduction

Probably the most well-known interaction with an artificial intelligence (AI) application is currently done in natural language using OpenAI's ChatGPT[1] or similar systems, such as Gemini from Google[2], Microsoft Copilot[3], Meta's

---

[1] https://openai.com/index/chatgpt/, last accessed on 12.10.2024, see also OpenAI (2013)
[2] https://deepmind.google/technologies/gemini/, last accessed on pe 24.06.2024.
[3] https://copilot.microsoft.com/, last accessed on 11.10.2024

LLaMA[4], Anthropic's Claude[5], and Inflection AI's Pi[6]. Such conversational interactions with AI have seen a tremendous enthusiasm in the past three years, with the launch of the above-mentioned advanced interactive AI-based platforms that can generate natural language and/or images in response to user requests or questions, even entering the conversations imitating a human person. These so-called "chatbots" are a subset of the field of generative AI, which may include not only language but also images, video, and/or music.

Chatbots are now used in many fields that require the development of textual documents, and even in computer programming. Their remarkable achievements are provided by artificial neural networks with a large number of layers ("Deep Neural Networks" - DNN), which are the most efficacious types of machine learning techniques. For natural language processing (NLP), the most advanced solutions are transformer-type DNN architectures, for example, BERT (Devlin et al., 2019) and GPTs (Brown et al., 2020). These architectures have enabled the development of large language models (LLMs), which underlie the chatbot systems listed above. LLMs are trained with practically all the available texts in electronic format, in order to generate the most probable word that follows a given sequence of words (Jurafsky and Martin, 2025).

After the outstanding achievements of the text generation capabilities of the AI systems based on LLMs, nowadays there are a lot of debates about the degree to which artificial intelligence has achieved human capabilities, the so-called Artificial General Intelligence (AGI), the Strong AI, or even the artificial superintelligence, as proposed by Bergmanna and Stryker (2024). Some say that AGI has already been achieved, for example, Agüera y Arcas and Norvig (2023). Others say that it will be achieved in a few years, for example, in 2024, Sam Altman, chief executive officer of OpenAI, said that AGI will be obtained in 2025, as cited by Smith (2025). However, Smith also claims that LLMs will not provide AGI, that there are human capabilities that will never be obtained by AI, as argued by Trausan-Matu (2019).

The idea that AI might equal human capabilities is even older than the AI domain, which was officially born in 1956. In 1950, Alan Turing imagined an imitation game for answering the question "Can machines think?", which

[4] https://docs.llamaindex.ai/en/stable/, last accessed on 24.06.2024
[5] https://www.anthropic.com/news/introducing-claude, last accessed on 24.06.2024
[6] https://pi.ai/, last accessed on 12.10.2024

was later named the Turing test of AI. This test is based on comparing short answers given by a human and a computer in an "imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning" (Turing, 1950).

The analysis from this paper can be seen as a kind of an alternative to the Turing test, making a comparison between several pages-long literary texts (stories written by famous writers) and stories generated by ChatGPT as a result of prompts, some created by the first author and some generated also by ChatGPT. An experimental comparison has been made by an NLP program, using rhetorical devices, between 10 human-authored stories and 10 AI-generated stories as a result of prompts.

While previous research has largely focused on detecting AI-generated texts using supervised classifiers, statistical heuristics, or watermarking techniques, this paper offers a novel perspective by comparing the stylistic and aesthetic qualities of AI- and human-generated stories through the lens of rhetorical devices. Unlike common AI text detection approaches, the method presented in this paper quantitatively analyzes distinctive literary features such as anaphora, epistrophe, or alliteration, providing a deeper understanding of how well AI can emulate human creativity in storytelling.

The focus on rhetoric-based stylistic comparison represents a valuable and relatively unexplored approach in the evaluation of AI-generated narratives, contributing new insights into the progress towards AGI beyond traditional Turing test paradigms. The alternative perspective to the Turing test focuses on sustained stylistic expression in long texts rather than shorter conversational exchanges.

2. State of the Art

The goal of this paper is related to the already considered task of detecting AI-generated texts. The systems used for this task rely heavily on supervised classifiers, statistical heuristics such as perplexity curves and GLTR (Gehrmann et al., 2019), or watermarking techniques added at generation time (Fraser et al., 2025).

The research presented in this paper is not oriented towards detecting AI-generated texts but rather for comparing AI and human texts from aesthetic and stylistic perspectives. In this aim, rhetorical devices were previously considered by Trausan-Matu and Urse (2024).

So far, there is little research in the area of rhetoric – and more specifically, analysis of the rhetorical features of a text – to ascertain the humanity of the author. In one of the first studies of this kind, Reinhart et al. (2025) use Douglas Biber's tagset of 66 linguistic features to compare the rhetorical style of human vs AI-generated stories. Biber has defined a varied set of linguistic features, lexical, grammatical, or rhetorical, such as:

- Verb tenses (past, perfect, or present)
- Nominalizations, gerunds, agentless passive voice
- Phrasal or causal coordination.

These authors composed a comprehensive dataset based on the Corpus of Contemporary American English (COCA), created by Davies (2008), and other sources, divided into academic, news, fiction, spoken, blogs, and television and movie script categories.

Based on approximately 9K texts, the authors used GPT-4o and Llama 3 models to generate a parallel AI-generated corpus by prompting the LLMs with the first 500 words of every human text and asking them to complete the next 500 words. The two corpora were then analyzed in terms of counts of linguistic features, counts per 1000 words.

From this study, there are two takeaways relevant to the subject of this paper (Reinhart et al., 2025):

- There are significant differences in style between human and AI-generated text, which results in AI text being more informationally dense and noun-heavy than human text.
- The linguistic analysis using Douglas Biber's tagset brings a lot of value by revealing "tacit" information that helps distinguish human vs AI-generated text.

3. Data Generation

A set of 10 human-written stories was chosen to be compared against 10 AI-generated stories. The below prompt template was used to generate 5 of the stories, using ChatGPT 4.o:

> *You are a writer who can write in any style. Please write a story of 40-50 lines about two 7-year-old kids who run away from home to earn money so that they can buy a necklace for a girl they like, in a comical, light fashion, resembling Mark Twain.*

For the other 5 stories, ChatGPT itself was asked to generate the prompts using the previous 5 human prompts as examples in a few-shot technique. Below is an example of a ChatGPT-generated prompt:

> *You are a writer who can write in any style. Please write a story of 40–50 lines about a disgraced chef who opens a food cart at the end of the world after an apocalypse. Write it in a darkly comic, satirical style, reminiscent of Kurt Vonnegut or George Saunders. Make sure it's prose.*

The usage of ChatGPT-generated prompts versus human-generated prompts also aims to test not only AI's story-writing capacity but also its creativity in framing narratives.

The 10 human stories were chosen from the Gutenberg project[7], somewhat randomly, with the following authors: F. Marion Crawford, Chatham, Victor Hugo, Henryk Sienkiewicz, Charles Dickens, Edward Bulwer Lytton (2 texts), Myra S. Delano, George W. Cable, and an anonymous writer.

4. Rhetorical Devices

Rhetorical devices are ways of making texts more appealing, persuasive, or striking for readers. They are obtained by particular patterns, using sound or sense. For example, several rhetorical devices considered in the analysis performed by Trausan-Matu and Urse (2024) were:

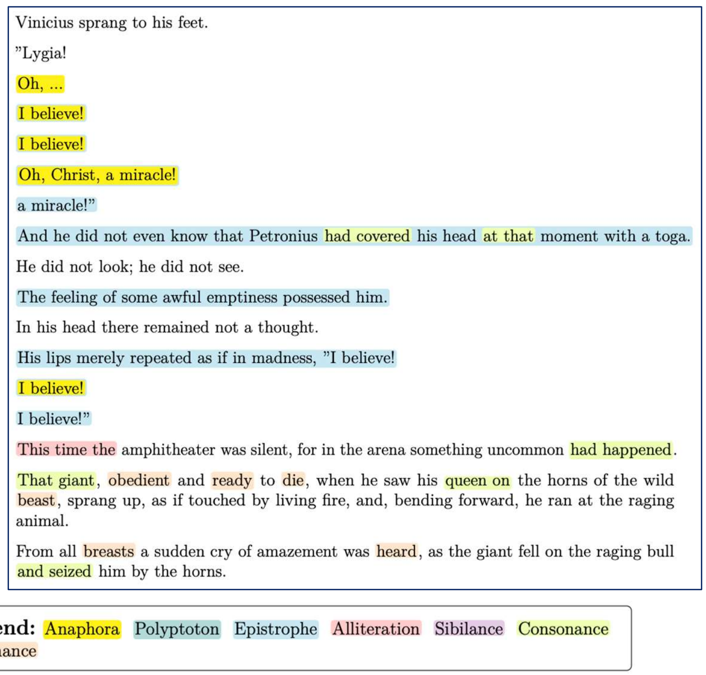
- **Anaphora** - a word or a group of words repeated at the beginning of two or more successive phrases, adding emphasis and/or unity;
- **Polyptotons** - repetitions of words with the same root, identified by stemming;
- **Epistrophe** - a word or a group of words at the end of two successive

[7] https://www.gutenberg.org, last accessed on 24.06.2024

sentences is the same.

- **Alliteration** – repetition of the first sounds of words within sentences;
- **Sibilance** - repeated hissing sounds, such as 's', 'sh', 'z', 'ch', 'x', within a group of words;
- **Consonance** - repeated consonant sounds within a group of words, excluding common suffixes, such as “ing”, ”ed”, “es”, “s”;
- **Assonance** - repeated vowel sounds within a group of words.

For example, several rhetorical features were annotated in an excerpt taken from Henryk Sienkiewicz’s “The Arena Scene” of ”Quo Vadis” (Figure 1).



Vinicius sprang to his feet.

”Lygia!

Oh, ...

I believe!

I believe!

Oh, Christ, a miracle!

a miracle!”

And he did not even know that Petronius had covered his head at that moment with a toga.

He did not look; he did not see.

The feeling of some awful emptiness possessed him.

In his head there remained not a thought.

His lips merely repeated as if in madness, ”I believe!

I believe!

I believe!”

This time the amphitheater was silent, for in the arena something uncommon had happened.

That giant, obedient and ready to die, when he saw his queen on the horns of the wild beast, sprang up, as if touched by living fire, and, bending forward, he ran at the raging animal.

From all breasts a sudden cry of amazement was heard, as the giant fell on the raging bull and seized him by the horns.

Figure 1. Annotated Excerpt of a human-authored story

5. Results

Several analyses were performed using a Python program. The first analysis is a simple counting of all the rhetorical devices, the results being shown in

Table 1. As a first impression, it looks like AI-generated stories are much poorer in terms of rhetorical devices. The first question that arises is whether this is a fair comparison - i.e., are all stories the same size? Looking at the answer, the human stories are of varied sizes.

Table 1. Counts of rhetorical devices per story

ID	Author	Anaphora	Polyptoton	Epistrophe	Alliteration	Sibilance	Consonance	Assonance
1	ChatGPT-4o	2	0	29	16	10	9	38
2	ChatGPT-4o	5	0	34	18	12	8	44
3	ChatGPT-4o	2	0	25	15	7	6	31
4	ChatGPT-4o	10	0	42	11	7	11	41
5	ChatGPT-4o	2	0	19	11	5	5	38
6	ChatGPT-4o	7	0	35	14	3	6	39
7	ChatGPT-4o	3	1	27	11	2	9	38
8	ChatGPT-4o	7	0	40	10	5	11	39
9	ChatGPT-4o	1	0	27	8	5	2	21
10	ChatGPT-4o	6	8	33	13	4	7	33
11	F. MARION CRAWFORD	8	8	56	47	21	24	87
12	Chatham	1	0	23	22	9	15	64
13	Victor Hugo	33	1	64	34	6	29	93
14	HENRYK SIENKIEWICZ	12	24	109	96	15	47	208
15	CHARLES DICKENS	38	3	86	62	19	31	136
16	EDWARD BULWER LYTTON	21	9	69	56	19	32	203
17	MYRA S. DELANO	11	1	58	66	22	27	116
18	GEORGE W. CABLE	8	0	46	41	14	19	114
19	EDWARD BULWER LYTTON	11	11	74	77	27	45	209
20	ANONYMOUS	12	0	51	30	15	12	87

For a fair comparison, Table 2 presents the counts of rhetorical devices weighted per word count per story. It suggests that, roughly, with a few exceptions, the number of rhetorical devices is quite the same between human and AI-generated stories.

Table 2. Counts of rhetorical devices per story, weighted with the word count of the story

ID	Author	Anaphora	Polyptoton	Epistrophe	Alliteration	Sibilance	Consonance	Assonance	Word count
1	ChatGPT-4o	0.01	0	0.08	0.04	0.03	0.02	0.1	375
2	ChatGPT-4o	0.01	0	0.09	0.05	0.03	0.02	0.11	386
3	ChatGPT-4o	0.01	0	0.07	0.04	0.02	0.02	0.09	357
4	ChatGPT-4o	0.03	0	0.12	0.03	0.02	0.03	0.11	358
5	ChatGPT-4o	0.01	0	0.06	0.03	0.02	0.02	0.12	329
6	ChatGPT-4o	0.02	0	0.11	0.04	0.01	0.02	0.12	326
7	ChatGPT-4o	0.01	0	0.08	0.03	0.01	0.03	0.11	338
8	ChatGPT-4o	0.02	0	0.1	0.03	0.01	0.03	0.1	392
9	ChatGPT-4o	0	0	0.1	0.03	0.02	0.01	0.07	281
10	ChatGPT-4o	0.02	0.03	0.1	0.04	0.01	0.02	0.1	318
11	F. MARION CRAWFC	0.01	0.01	0.05	0.04	0.02	0.02	0.07	1241
12	Chatham	0	0	0.03	0.03	0.01	0.02	0.09	739
13	Victor Hugo	0.03	0	0.05	0.03	0	0.02	0.08	1212
14	HENRYK SIENKIEWI	0.01	0.01	0.05	0.04	0.01	0.02	0.09	2379
15	CHARLES DICKENS	0.02	0	0.04	0.03	0.01	0.02	0.07	2008
16	EDWARD BULWER L	0.01	0	0.03	0.03	0.01	0.02	0.1	2025
17	MYRA S. DELANO	0.01	0	0.04	0.05	0.02	0.02	0.08	1374
18	GEORGE W. CABLE	0.01	0	0.04	0.04	0.01	0.02	0.1	1143
19	EDWARD BULWER L	0	0	0.03	0.03	0.01	0.02	0.09	2219
20	ANONYMOUS	0.01	0	0.06	0.03	0.02	0.01	0.09	925

The next two tables, Table 3 and Table 4, present a side-by-side comparison of the averages per story category (human or AI-generated).

Table 3. Averages of rhetorical devices per author type

Category	Avg. Anaphora	Avg. Polyptoton	Avg. Epistrophe	Avg. Alliteration	Avg. Sibilance	Avg. Consonance	Avg. Assonance
AI	4.5	0.9	31.1	12.7	6	7.4	36.2
Human	15.5	5.7	63.6	53.1	16.7	28.1	131.7

Table 4. Averages of rhetorical devices per author type, weighted with the word counts of the stories

Category	Anaphora / word	Polyptoton / word	Epistrophe / word	Alliteration / word	Sibilance / word	Consonance / word	Assonance / word
AI	0.0129	0.0028	0.09	0.0366	0.017	0.021	0.1042
Human	0.0102	0.0029	0.0423	0.0346	0.0116	0.0184	0.0861

It can be observed in Table 4 that almost all average values are similar, except for epistrophe, where AI-generated texts show counts two times higher than human-generated texts. Figure 2 presents a more intuitive view of this insight.

While raw counts suggested AI-generated texts were stylistically poorer, normalization reveals almost parity, suggesting that text length instead of stylistic competence is the main differentiator. Normalization can be viewed as a fair comparison mechanism that allows the comparison of stylistic richness independent of story length.

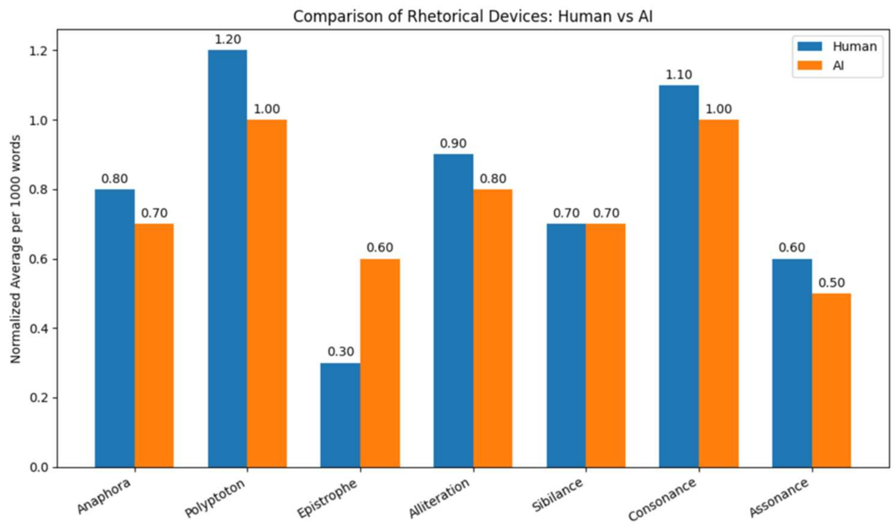


Figure 2. A visualization of normalized counts per 1000 words

6. Conclusions and Further Research

One conclusion of this study, even though preliminary, is that the proposed Turing AI test, the imitation game, may no longer be sufficient given today's AI capabilities. There already are other such approaches, for example, the Winograd Schemas (Levesque et al., 2012). Since rhetorical devices normalized counts show roughly the same numbers, the preliminary conclusion is that AI-generated stories exhibit comparable stylistic richness to human-authored ones, within the scope of this dataset. However, whether this constitutes genuine creativity or simply statistical reproduction of patterns remains an open question. In summary, at this moment, not only can an AI agent perfectly mimic a human in shallow conversations, but it can also generate compelling stories in a convincing literary style, which is a good step forward for reaching AGI.

It should be noted that the research presented here is work in progress, the current study having some drawbacks. The first drawback is that the dataset is relatively small, comprised of a limited number of human and AI-generated stories. The second drawback is the lack of story labeling; for example, there is no distinction between stories based on literary style or theme. Finally, the prompt design used for story generation can be further refined to ensure greater stylistic alignment between human and AI stories.

There are a number of areas this research will be expanded to in the future:

- Comparison of rhetoric device counts shall be made for a more statistically significant batch of texts (thousands or more)
- Expand to more languages in order to determine whether LLMs have a language bias
- Labeling each story, using an LLM or algorithmic methods, with word count, literary style, thematic, etc.
- Create a more sophisticated prompt that will allow a fairer comparison (e.g. inspired from Reinhart et al. (2025) we can mask a portion of a human generated text and ask the LLM to generate text that fits in the overall text, then run a comparison for the masked text against the generated one) or use a templated prompt for a chatbot to generate stories similar to the human stories (e.g. for each human story generate a similar story, having a similar word count, literary style and thematic), like in the experiment presented in Pascu and Trausan-Matu (2025).
- Engage human evaluators to provide a nuanced comparative analysis of human and AI-generated texts.
- Include more rhetorical devices in the study, for example, Douglas Biber's tagset of 66 linguistic features.

References

Agüera y Arcas, B., Norvig, P. (2023), *Artificial General Intelligence Is Already Here*, https://www.noemamag.com/artificial-general-intelligence-is-already-here/, accessed on 9 July 2025

Bergmann, D., Stryker, C. (2024). *What is artificial general intelligence (AGI)?* https://www.ibm.com/think/topics/artificial-general-intelligence, accessed on 9 July 2025

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)* https://www.english-corpora.org/coca/, accessed on 9 July 2025

Fraser, K C., Dawkins, H., Kiritchenko, S.(2025). Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods, Journal of Artificial Intelligence Research Vol. 82 (2025) 2233-2278, https://doi.org/10.1613/jair.1.16665

Gehrmann, S., Strobelt, H., Rush, A. (2019) *GLTR: Statistical Detection and Visualization of Generated Text* https://aclanthology.org/P19-3019/, accessed on 9 July 2025

IBM (2023). *Large language models*. IBM. https://www.ibm.com/think/topics/large-language-models, accessed on 9 July 2025

Jones, C. R., Bergen, B. K. (2025) *Large Language Models Pass the Turing Test,* arXiv:2503.23674

Jurafsky, D., Martin, J.H. (2025*). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025. https://web.stanford.edu/~jurafsky/slp3.

Levesque, H.J., Davis, E., Morgenstern. L. (2012) The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (KR'12)*. AAAI Press, 552–561.

Pascu, I.C., Trausan-Matu, S. (2025). Analysis of persona assigned LLMs, *Proceedings of International Conference on User-System Interaction, ICUSI 2025*, pp.107-114

Reinhart, A., Markey, B., Laudenbach, M., Kachatad, P., Yurko, R., Weinberg, G., Brown, D.W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles, *Proceedings of the National Academy of Sciences*, http://dx.doi.org/10.1073/pnas.2422455122, accessed on 11 October 2025

Smith, G. (2025). W*hy LLMs chatbots won't lead to artificial general intelligence* https://mindmatters.ai/2025/02/why-llms-chatbots-wont-lead-to-artificial-general-intelligence, accessed on 9 July 2025

Trausan-Matu, S. (2019). Is it possible to grow an I–Thou relation with an artificial agent? A dialogistic perspective, *AI & Society, Journal of Knowledge, Culture and Communication,* Vol. 34, Nr. 1, Special Issue: Ethics of AI and Robotics, Springer-Verlag London, pp. 9-17.

Trausan-Matu, S., Urse, A.-D. (2024) Computer support for the analysis of rhetorical devices in texts, in Puchianu, C.M., Datcu, D. (eds*.) International Conference on User-System Interaction*, MatrixRom, Bucuresti, pp. 95-98

Turing, A. M. (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460..

.