Stylometric Approach to AI-generated Texts. An Analysis of Contemporary French-Language Literature

Adam Pawłowski
University of Wrocław
pl. Uniwersytecki 1
50-137 Wrocław, Poland
adam.pawlowski@uwr.edu.pl

Tomasz Walkowiak
Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław, Poland
tomasz.walkowiak@pwr.edu.pl

Abstract

The article focuses on a stylometric analysis of authentic literary texts and thematically related texts generated by large language models. The texts under study represent a fairly broad cross-section of twentieth-century French literature. Five models were used to generate the texts (ChatGPT 4-o, GPT 4-o mini, DeepSeek v.3, c4ai-command-r-plus, and c4ai-command-a). The original human-written stories of approximately 20,000 characters were summarized, and new narratives were then generated on the basis of these abstracts. In terms of plot and style, they were intended to resemble the originals. The research carried out with TF-IDF of the most frequent words showed that texts generated by specific LLMs and written by humans cluster relatively well as distinct groups. The experiments also showed that the "authorial" specificity of machine-generated texts partly matches the original clustering of human-written source texts.

1 Introduction

In today's communication landscape, machine-generated texts have secured a permanent place. Large Language Models (LLMs) are widely used to produce function texts on a variety of topics (formal letters, instructions, applications, etc.). LLMs translate, summarize, and assist humans in information retrieval and problem solving. They can also engage in quasi-natural conversations. One area where machine-generated texts have not yet reached a quality comparable to that of human-created ones is literary fiction. One might even say that literature has become the last shore defended against the expansion of machines, belonging to humans who reassure each other that there will never be an "electronic Aeschylus, Shakespeare, Molière or Balzac." Whether this will actually never happen remains unknown. However, it is certain that some authors are already using LLMs as intelligent assistants to generate their literary texts. A comparative study of human- and machine-generated literary texts is therefore justified and necessary. On the one hand, such experiments serve as test of model performance; on the other hand, they reveal whether there still exists a boundary — and, if so, where it lies — separating human creativity from machine-generated AI outputs.

2 State-of-the-Art

Since the publication of the groundbreaking study by Mosteller and Wallace in 1964, it has been well established that reliable results in stylometric research are obtained by analyzing content-free function words rather than content-bearing lexemes (Mosteller and Wallace, 1964). This is because function words are resistant to conscious manipulation by the author and are largely independent of the subject matter of a text. A subsequent stage in the development of stylometric research involved the widespread adoption of advanced multivariate methods and the extension of the linguistic category of the word to include n-grams, defined as arbitrary sequences of characters.

Today, it is assumed that the aim of the interdisciplinary field of stylometry (Sara El Manar El Bouanani, 2014; Päpcke et al., 2022; Eder et al., 2015) is to explore the relationship between the statistical features of texts and their meta-characteristics, such as authorship (Juola, 2006; Stamatatos, 2009; Kestemont, 2014) or the literary period of the work (Rabaev et al., 2023).

In recent years, increasing attention has been paid to the study of texts generated by artificial intelligence. The computational methods employed in this line of research do not differ substantially from those used during the period when only human-authored texts were under investigation. However, the principal challenge lies in the se-

Proceedings of SIGHUM 2026, pages 221–226

lection of the textual material. Existing studies tend to rely on the most readily available corpora, which typically lack a clearly defined stylistic profile. Consequently, the conclusions drawn from such analyzes are of limited relevance to the study of creative or literary texts.

A representative example of this trend can be found in (Przystalski et al., 2026), where Wikipedia materials were processed without taking into account the fact that the encyclopedic style – particularly in the context of collaborative writing – does not exhibit a distinctive authorial stylistic signature and that Wikipedia contributors increasingly rely on AI-assisted tools when composing entries.

A more valuable contribution to the analysis of literary style in the context of creative writing is offered by (O'Sullivan, 2025). The author examined creative texts produced by both human writers and machines and concluded as follows: "The results reveal clear and consistent stylistic distinctions. Human-authored texts form broader, more heterogeneous clusters, reflecting the diversity of individual expression, writing ability, and interpretive engagement with the prompts. In contrast, LLM outputs, while fluent and coherent, display a higher degree of stylistic uniformity, clustering tightly by model." (ibid.)

3 Goal of the Study

The experiment conducted aimed to apply selected stylometric measures to a corpus of twentieth-century literary fiction texts, which included samples written by humans (see the Appendix) and generated automatically by LLMs. Particular attention should be paid to the way the machine corpus was constructed. Our intention was to avoid comparing random text samples whose similarity or lack thereof would be difficult to interpret. To this end, we prepared a corpus of text pairs (human vs machine) related in terms of theme and style, which recount the same plots and employ the same narrative schemes.

The aim of the study was to verify how NLP and stylometric methods would cluster these texts, and to identify some hidden relations. This is certainly not an answer to the question of whether "a machine can write books" or whether AI-written narratives are as good as human literature; however, this measurement indirectly indicates the specificity of the literature produced by LLMs. The study was conducted on prose texts written in French by authors from Europe (15 authors) and Africa (7 authors).

It is worth emphasizing that the task formulated here is demanding, as it goes beyond standard research aimed at distinguishing between machine-generated and human-written texts. Since the essence of literature is the creative fictionality based on the principle of mimesis, there is no truth criterion that could be applied to evaluate machine-generated texts. This radically distinguishes literary stylometry from studies of texts written under one's own identity (e.g., letters, blogs, posts on social networks), as well as from functional texts (e.g., administrative or informational). The quality of literature has always been judged by the reader and his/her expectations rather than by its truthfulness or similarity to any specific pattern. There are numerous examples of works considered as worthy in one literary period but rejected by subsequent generations of readers. There are also works praised in one culture or language but dismissed in another. For this reason, a mixed "machine-human" corpus of texts linked by topics and (supposedly) style constitutes a valuable testing material for our research.

We formulated two hypotheses. The first posits that stylometric methods, which make it possible to identify the author of a literary text on the basis of style, will also prove effective in the analysis of machine-generated literature. The second hypothesis posits that texts generated by LLMs exhibit certain common features of a specific 'collective identity'. This implies that, at a sufficiently high level of abstraction, human-written texts and texts produced by various LLMs will form distinct clusters.

4 Methods

The study was conducted in three stages. First, we compiled a corpus of contemporary French-language prose from the twentieth century (Pawłowski et al., 2025) (cf. Figure 1). From contemporary novels available in public resources (libraries and repositories) worldwide, we extracted samples of approximately 15,000 to 20,000 characters long. An average of ten samples were generated from each novel. In total, we produced 744 "human-authored" source samples that were further used to generate the corresponding machine-produced texts. The choice of twentieth-century texts, often protected by copyright and difficult to

access, was deliberate. Large language models are trained primarily on contemporary language, and we assumed that the resulting machine-generated texts would be of higher quality than those based on nineteenth-century literature, which is available in free-access repositories. In the second phase,

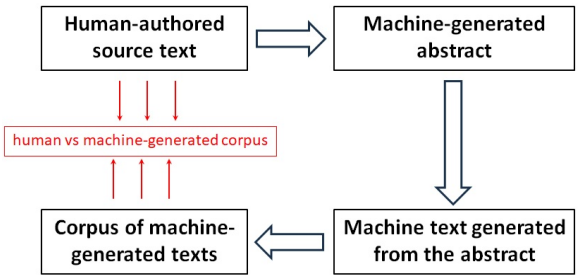

Figure 1: Data processing pipeline

we generated the texts corresponding to the source samples. Using each LLM, we produced 500-word summaries, and subsequently, on the basis of these summaries, full "literary" texts on the same topic were reproduced. An example of a summarization prompt for a novel by the Senegalese writer Fatou Diome was as follows:

Voici le texte de Fatou Diome, écrivaine sénégalaise d'expression française. Je voudrais obtenir un résumé de ce texte d'une longueur de 500 mots.

An example of a prompt used to generate an entire micro-story by the same author was as follows:

Voici le résumé de texte de Fatou Diome, écrivaine sénégalaise d'expression française. Sur la base de ce résumé, je voudrais obtenir un récit cohérent d'une longueur d'environ 2560 mots, rédigé dans le style de Fatou Diome.

It is worth noting that prompts defined in this way are highly elaborate and go beyond standard AI-based text generation methods. They comprise not only metadata (in the above query: the author's surname, gender, and nationality), but also a detail-rich content description embedded in the abstract.

Texts generated by the models rarely reached the expected length in the first generation cycle. A likely reason for this is a limitation inherent to LLMs, which have difficulty sustaining long-form narratives. Consequently, in many cases, we employed follow-up prompts explicitly requesting an increase in the length of the generated narrative. Since we used five LLMs, the final corpus comprised 3,999 samples: 744 human-authored texts and 3,255 machine-generated texts. Some API calls did not respond, which is why the number of machine-generated texts is lower than the expected $744 \times 5 = 3{,}720$. The models evaluated included two OpenAI models (GPT-4o and GPT-4o-mini), two models developed by Cohere Labs (c4ai-command-r-plus and c4ai-command-a), and one model from DeepSeek AI (DeepSeek-V3).

In the third phase, stylometric feature vectors were constructed for each text. We relied on a standard stylometric representation based on the frequencies of the most frequent words, weighted with TF–IDF (Salton G, 1988). For exploratory analysis, the resulting vectors were projected onto a two-dimensional space using the PaCMAP (Wang et al., 2021) dimensionality reduction technique. Next, using the classifier of the closest neighbors (k-NN) with $k = 5$, we evaluated three classification tasks: (i) detection of human text, distinguishing human-authored texts from AI-generated ones; (ii) detection of text origin, identifying whether a text was written by a human or by one of the language models used; and (iii) analysis of authorship, aimed at attributing texts to their respective authors. Finally, we quantified text similarity within the same author and source (human or specific LLM) by computing the average distance between each text's stylometric vector and its nearest neighbor from the same author and source.

5 Results

The results obtained should be considered valuable and, in some cases, even surprising. As expected, attribution of authorship for human authors yielded convincing results. However, relatively high performance was also achieved in the AI-author corpus (Table 1). Notably, reliable identification of AI-authored texts becomes possible only when at least 1,000 features are used, whereas in the human-authored corpus stylistic variation is more pronounced: high classification accuracy is observed even with a relatively small number of features. This indicates that human authors are more strongly individualized than the closely related group of AI authors.

Figure 2 shows that human texts can be effectively distinguished from machine-generated ones, despite the use of highly elaborate prompts (an instruction and a 500-word abstract). Figure 3 proved particularly striking. It revealed that LLMs exhibit their own stylistic identities: texts generated by different LLMs on the basis of identical prompts and abstracts formed distinct clusters. Using anthropocentric terminology, one would say that each

LLM possesses its own specific knowledge representation ("image of the world") and communicates through its own "idiolect". Texts generated by certain LLMs (e.g. GPT-4.0 and GPT-4.0-mini) cluster closely together and appear to have been trained on similar datasets, while others are positioned further apart.

Table 2 is also noteworthy. It shows that texts written by each human author are more similar to each other than texts written on the same topics by AI. LLMs seem to have some inherent constraints of their "individuality", understood as the ability to emulate the distinctive stylistic features of specific human authors. These constraints apparently impose limits to any nuanced stylistic imitation. LLMs thus appear to exhibit a form of dominant collective identity and a sort of collective authorial fingerprint. By contrast, human authors display greater stylistic coherence within samples of one author, while differing more markedly between authors.

A very interesting question is the division of machine-generated texts into two groups of unequal size, as shown in Figures 2 and 3. Clearly explaining this phenomenon was not straightforward. The analysis of the data contained in both clusters convinced us that this division cannot be explained by semantic criteria (such as gender, year of publication, genre, European versus African origin, etc.). The explanation turned out to be entirely different. As noted above, some of the generated texts were too short, so LLMs were asked to produce new narratives of the required length. The division into two groups thus indicates a clustering of texts completed after the initial prompt and those generated in response to a second prompt. This secondary "urgency" prompt was in fact radically different from the first, as it incorporated the entire context — not only the original abstract but also the text already generated. This leads to an important methodological lesson. While we treated the "expanding" prompts as primary instructions, the LLMs interpreted them as composite prompts, taking into account as an intermediate product the previously generated text. As a result, we inadvertently violated the epistemological principle of conducting experimental research under identical conditions, known as *ceteris paribus* principle.

	k-NN accuracy [%]				
				authorship	
n	**hum**	**cyber**	**all**	**hum**	**ai**
5	82.72	45.04	9.95	21.10	9.95
10	91.35	62.39	16.65	32.26	14.65
20	94.57	74.07	27.48	56.99	21.44
50	96.52	81.12	32.93	66.26	25.93
100	96.85	81.77	37.08	73.39	29.43
200	96.37	81.55	41.31	76.61	33.70
500	95.27	78.87	50.19	80.91	43.47
1000	93.95	67.02	71.49	87.37	67.80
2000	90.60	55.71	80.47	88.58	78.13
3000	89.77	48.96	83.90	89.38	83.13
4000	89.45	45.59	86.25	89.52	85.19
5000	88.82	42.06	86.82	88.98	86.02

Table 1: k-NN classification accuracy (k = 5) for different numbers of the most frequent words (n) in three tasks: hum – human vs. LLM texts (2 classes), cyber – human vs. individual LLMs (6 classes), and authorship – author identification (22 classes). For authorship, results are reported for all texts (all), human-only (hum), and LLM-generated texts (ai).

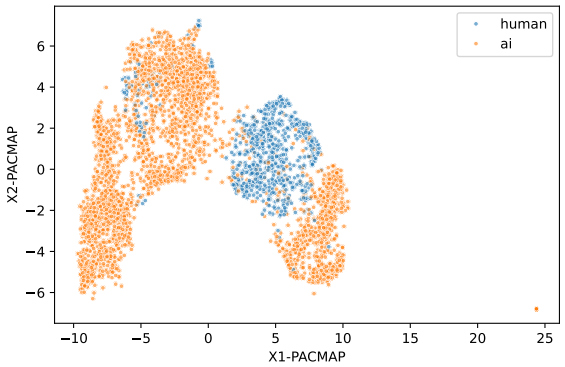

Figure 2: PaCMAP visualization based on TF-IDF vectors of the 100 most frequent words, illustrating a strong and consistent separation between human-authored texts and AI-generated outputs.

	number of features			
model	**500**	**1000**	**3000**	**5000**
c4ai-comm-a	0.425	0.486	0.598	0.659
c4ai-comm-r	0.402	0.465	0.590	0.654
deepseek-v3	0.486	0.549	0.679	0.748
gpt-4o	0.421	0.494	0.637	0.706
gpt-4o-mini	0.384	0.451	0.568	0.625
human	0.366	0.419	0.516	0.568

Table 2: Average distance between texts and its nearest neighbor from the same author and source (one of LLMs or human), evaluated for varying numbers of the most frequent words. Human-authored texts consistently show the lowest mean distances.

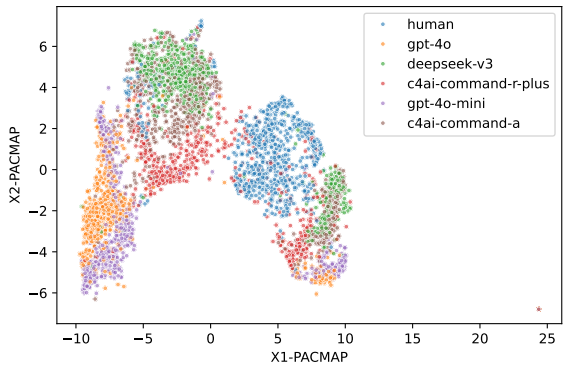

Figure 3: PaCMAP visualization using TF-IDF vectors of the 100 most frequent words, comparing multiple language models with human-authored texts. Outputs from each model separate into two clusters, while human texts remain grouped in one cohesive cluster.

6 Conclusions

The study provides a reliable extension of existing research on the relationship between human- and machine-generated texts in the field of literary fiction, particularly prose. The analysis was carried out on thematically and stylistically related human and machine-authored texts, representing narrative fiction. The study demonstrated that machine-generated texts, produced by various models, exhibit a sort of "collective character". Despite the use of elaborate prompts, the texts that mimicked different authors were more similar to each other than the original human-authored texts. In the case of LLMs, this appears as though a single mind were attempting to imitate multiple voices and writing styles. Unlike AI models, humans exhibit a more distinctive authorial voice and personality. The research also revealed that LLMs possess their own "identities," which include specific representations of knowledge and characteristic modes of conveying it, which could be called idiolects. In general, these findings suggest that despite the inherent limitations of the human species, authors of literary fiction remain more original and creative than LLMs.

Limitations

Literature, in its various forms of prose, poetry, and drama, is not intended for machines but for humans. Its purpose is to provide an esthetic and intellectual experience that leads to what has been termed *catharsis* since Aristotle. True humanists therefore argue that the ultimate measure of a literary work's value is human judgment, grounded in sensitivity, knowledge, and culture. However, methods developed within natural language processing (NLP) are not compatible with readers' reception because algorithms do not "read" texts through human eyes. Consequently, an effective assessment of the similarity between machine-generated and human-authored texts should indeed be undertaken by humans. However, humans have a sluggish rate of processing textual information, despite the complexity and perceptiveness of the human mind. Therefore, it is difficult to imagine an efficient, technical evaluation of hundreds or thousands of literary samples in terms of quality, "humanness," "machineliness," or similarity to other texts when such an evaluation is carried out solely by human assessors. Consequently, only NLP tools can guarantee reliable and reproducible results when researching large volumes of human and machine-generated texts.

There are also other potential limitations to our inferences. Dimensionality reduction, which lies at the core of the methods we employ, is akin to a long journey – and the journey itself often transforms the traveler. It begins with data represented in, say, a 1,000-dimensional space and ends with a two-dimensional visualization. Therefore, dimensionality reduction entails an inevitable loss of information and a simplification of reality. This is further compounded by the profound subjectivity inherent in the creative literary process and its evaluation. Language is not a material substance, nor is it a collection of elementary particles to which methods validated in the study of inorganic matter can be straightforwardly applied. The meaning of words and sentences shifts depending on the context and communicative situation. Consequently, the results presented here do not provide strong evidence to confirm or refute the earlier formulated hypotheses. Rather, they serve as signposts or milestones that indicate the main orientation points within a new hybrid machine-human communicative space.

Appendix

In this study, we processed a corpus consisting of text samples by the following authors: Louis Aragon, Mariama Bâ, Georges Bernanos, Calixthe Beyla, Tanella Boni, Ken Bugul, Louis-Ferdinand Céline, Albert Cohen, Fatou Diome, Romain Gary (Émile Ajar), Jean Giono, Jean-Marie Gustave Le Clézio, Andreï Makine, Scholastique Mukasonga, Paul Pavlovitch, Georges Perec, Raymond Que-

neau, Aminata Sow Fall, Michel Tournier, Boris Vian, Marguerite Yourcenar.

Acknowledgements

The work was financed by CLARIN-PL: Common Language Resources and Technology Infrastructure (POIR.04.02.00-00C002/19, 2024/WK/01, FENG.02.04-IP.040004/24).

References

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2015. Stylometry with r: A package for computational text analysis. *The R Journal*, 8:107–121.

Patrick Juola. 2006. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Frederick Mosteller and David Wallace. 1964. *Inference and disputed authorship : The Federalist*. Addison-Wesley series in behavioral science : quantitative methods. Addison-Wesley, Reading Mass. [etc.

James O'Sullivan. 2025. Stylometric comparisons of human versus ai-generated creative writing. *Humanities and Social Sciences Communications*, 12.

Adam Pawłowski, Ewa Kalinowska, and Tomasz Walkowiak. 2025. Au-delà de l'attribution d'auteur : la stylométrie permet-elle d'identifier l'« identité collective » des textes littéraires et le sexe des auteurs ? analyse comparée de la fiction narrative d'europe et d'afrique. *Romanica Cracoviensia*, 25:203.

Karol Przystalski, Jan K. Argasiński, Iwona Grabska-Gradzińska, and Jeremi K. Ochab. 2026. Stylometry recognizes human and llm-generated texts in short samples. *Expert Systems with Applications*, 296:129001.

Simon Päpcke, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. 2022. Stylometric similarity in literary corpora: Non-authorship clustering and Deutscher Novellenschatz. *Digital Scholarship in the Humanities*. Fqac039.

Irina Rabaev, Marina Litvak, Vladimir Younkin, Ricardo Campos, Alípio Mário Jorge, and Adam Jatowt. 2023. The competition on automatic classification of literary epochs. In *Proceedings of the IACT - The 1st International Workshop on Implicit Author Characterization from Texts for Search and Retrieval held in conjunction with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), Taipei, Taiwan, July 27, 2023*, volume 3477 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org.

Buckley C. Salton G. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Ismail Kassou Sara El Manar El Bouanani. 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12):22–29.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.