CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Stylometric comparisons of human versus AI-generated creative writing
Authors	O'Sullivan, James
Publication date	2025
Original Citation	O'Sullivan, J. (2025) 'Stylometric comparisons of human versus AI-generated creative writing', Humanities and Social Sciences Communications, 12(1), 1708. https://doi.org/10.1057/s41599-025-05986-3
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1057/s41599-025-05986-3
Rights	© 2025, the Author. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-06 04:34:35
Item downloaded from	https://hdl.handle.net/10468/16884



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05986-3 OPEN

Stylometric comparisons of human versus AI-generated creative writing

James O'Sullivan[1]✉

This study employs stylometry to investigate whether the creative writing styles of humans and large language models (LLMs) such as GPT-3.5, GPT-4, and Llama 70b can be distinguished through quantitative analysis. A balanced dataset of short stories composed in response to predefined narrative prompts forms the basis of the analysis. Burrows' Delta, a widely used metric in computational literary studies, is applied to measure stylistic similarity and difference across texts. By focusing on the distribution of the most frequent words, Burrows' Delta allows for comparison that is largely independent of content and instead sensitive to latent stylistic fingerprints. The methodology combines this measure with clustering techniques, including hierarchical clustering and multidimensional scaling, to visualise relationships between texts and to test whether human and machine-generated stories cohere into distinct groups. 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. GPT-4 demonstrates greater internal consistency than GPT-3.5, suggesting refinement in the stylistic coherence of newer systems, yet both remain distinguishable from human writing. Llama 70b shows similar uniform clustering behaviour. Occasional overlaps occur, particularly between GPT-3.5 and human texts, but these are rare and insufficient to erase the broader distinction between categories. The findings indicate that, despite rapid advances in generative AI and its growing capacity to simulate creativity, LLMs retain detectable stylistic signatures that separate them from human authors. The study contributes to debates about authenticity, authorship, and the scope of machine creativity by moving beyond subjective literary judgment towards quantitative evidence. Stylometric analysis confirms that LLM outputs remain statistically and stylistically identifiable as machine-generated.

[1] University College Cork, Cork, Ireland. ✉email: james.osullivan@ucc.ie

Introduction

Comparing human and AI-generated texts is receiving increasing academic attention (Herbold et al. 2023; Casal and Kessler, 2023; Amirjalili et al., 2024). Assessing the reception of textual content produced by generative AI, specifically, large language models (LLMs), is an essential aspect of their performance evaluation, and as the ability for such tools to mimic human endeavour grows, so too will their impact on society and culture.

In many respects, the economic feasibility of the LLM industry rests on the assumption that humans are willing to delegate significant portions of their labour, particularly mundane and routine writing tasks, to machines. Drafting standardised emails and reports, generating meeting summaries, responding to frequently asked queries, preparing basic legal documents such as terms of service are all examples of writing tasks that often involve repetitive or predictable patterns, making them well-suited for automation by LLMs. Advocates for the mass adoption of generative AI frame the reduction of human effort in repetitive tasks as a liberation of creativity, allowing individuals to focus on higher-order intellectual or artistic pursuits.

Creative expression enjoys a measure of moral and cultural prestige when compared with functional writing, so researchers have also begun exploring the extent to which AI-generated *creative* texts, such as poetry and fiction, compare with human efforts (Gunser et al. 2021; Beguš, 2023). It is one thing for an LLM to reliably produce an email or summary report; the capacity to automate the production of literature raises profound ethical and philosophical concerns about authenticity, originality, and the very nature of authorship.

Unsurprisingly, there has been a marked increase in popular treatments of the topic, exemplified by sensational headlines such as 'AI poetry rated better than poems written by humans' (Creamer, 2024). While there is evidence to suggest the general public prefers AI-generated poetry, this is largely a reflection of broader attitudes towards complex literature: poems produced by LLMs lack complexity, 'they are better at unambiguously communicating an image, a mood, an emotion, or a theme to non-expert readers of poetry', and so hold more popular appeal (Porter and Machery, 2024).

To date, few comparisons of human writing versus LLMs have proceeded with the use of computational linguistics, particularly in the context of style. The quantitative analysis of style, or stylometry, is particularly useful in comparing human versus AI texts because it emphasises *how* a text has been written, rather than what a text is about. Stylometry ignores the content of a piece, focusing instead on the latent authorial fingerprint—a linguistic signature detectable through the frequency with which particular words are chosen—that is present in all writings. Stylometry is essentially a quantitative form of authorship attribution, designed to indicate with a measure of statistical certainty the most likely author of a document or set of documents.

Using stylometry to compare creative writing produced by humans and LLMs strips away notions of 'complexity', bypasses subjective interpretations of literary merit or the emotional depth of a text, framing the comparison in terms of formal linguistic properties. Doing so raises the question of whether or not LLMs can mimic human creativity beyond subjective matters of taste towards more objective elements of literary production.

Methodology

Stylometry uses quantitative techniques to cluster texts based on their style (Burrows, 2002; Argamon, 2008; Evert et al., 2017; Neal et al., 2017). It operates by analysing measurable linguistic features such as word frequency, sentence length, syntactic structures, or punctuation patterns, which collectively form a unique 'fingerprint' of an author's or text's style. These features are extracted and transformed into numerical data, enabling the comparison and attribution of texts through statistical methods. It is a widely used technique in corpus and computational linguistics, particularly in the digital humanities, where it is often utilised as a form of distant reading in digital literary studies (Hoover, 2007; Rybicki and Heydel, 2013; Eder, 2017; O'Sullivan et al., 2018; Caddy, 2021; Savoy, 2023). Previous studies have used stylometry to compare human writing and LLMs (Zaitsu and Jin, 2023; Opara, 2024; Wang et al., 2024; Przystalski et al., 2024), but none have looked at explicitly literary texts, nor have they applied Burrows' Delta.

Traditionally, computational literary scholars conduct stylometric analyses using Burrows' Delta (Burrows, 2002), which works by focusing on the most frequent words (MFW) in a corpus—typically function words—as these words are believed to reveal consistent stylistic tendencies while being less influenced by thematic content. The frequencies of these words in each text are calculated and normalised using z-scores, which standardise the data to account for differences in text length and variability. The Delta value is then determined by calculating the average absolute difference in z-scores for the MFW between texts. A lower Delta value indicates greater stylistic similarity. This study deliberately excludes explicit measures of sentence length, syntactic complexity, and punctuation patterns. While alternative stylometric methods that include syntactic or structural analysis can provide valuable insights, relying on most frequent word analysis is largely considered best practice in computational literary studies.

While Burrows' Delta remains a foundational method in stylometry, techniques like Cosine Delta refine the metric by incorporating cosine similarity, which is less affected by variations in text length and scale (Jannidis et al., 2015). Comparative analyses of Burrows' Delta and its advanced variants have been conducted to evaluate their effectiveness in authorship attribution and stylistic analysis. A landmark study by Evert et al. (2017) systematically tested various Delta methods, including Burrows' original formulation, across corpora in multiple languages. The findings indicate that while Burrows' Delta remains robust, certain variants demonstrate improved performance in specific contexts. Machine learning methods further advance the field by employing algorithms like support vector machines to model stylistic patterns adaptively (Savoy, 2020).

For the purposes of this study, and considering the controlled corpus under examination, Burrows' Delta represents a robust technique for the analysis. The dataset has been gathered from an existing open dataset created by Nina Beguš (2024) for her behavioural and computational analyses of human and AI-generated texts, comprising short stories composed by GPT-3.5, GPT-4 and Llama, as well as crowdsourced human equivalents. Beguš's dataset is available through the Open Science Framework platform (2023) and is particularly suitable for this analysis as it offers a balanced representation of popular AI models. The complete Beguš corpus comprises 250 stories, crowdsourced through Amazon Mechanical Turk, as well as 80 stories generated by OpenAI's GPT-3.5 and GPT-4 LLMs, and 50 stories generated by Meta's open source Llama 3-70b model. Participants were tasked with writing 150–500-word stories in response to two 'please continue the story' prompts:

1. A human created an artificial human. Then this human (the creator/lover) fell in love with the artificial human.
2. A human (the creator) created an artificial human. Then another human (the lover) fell in love with the artificial human.

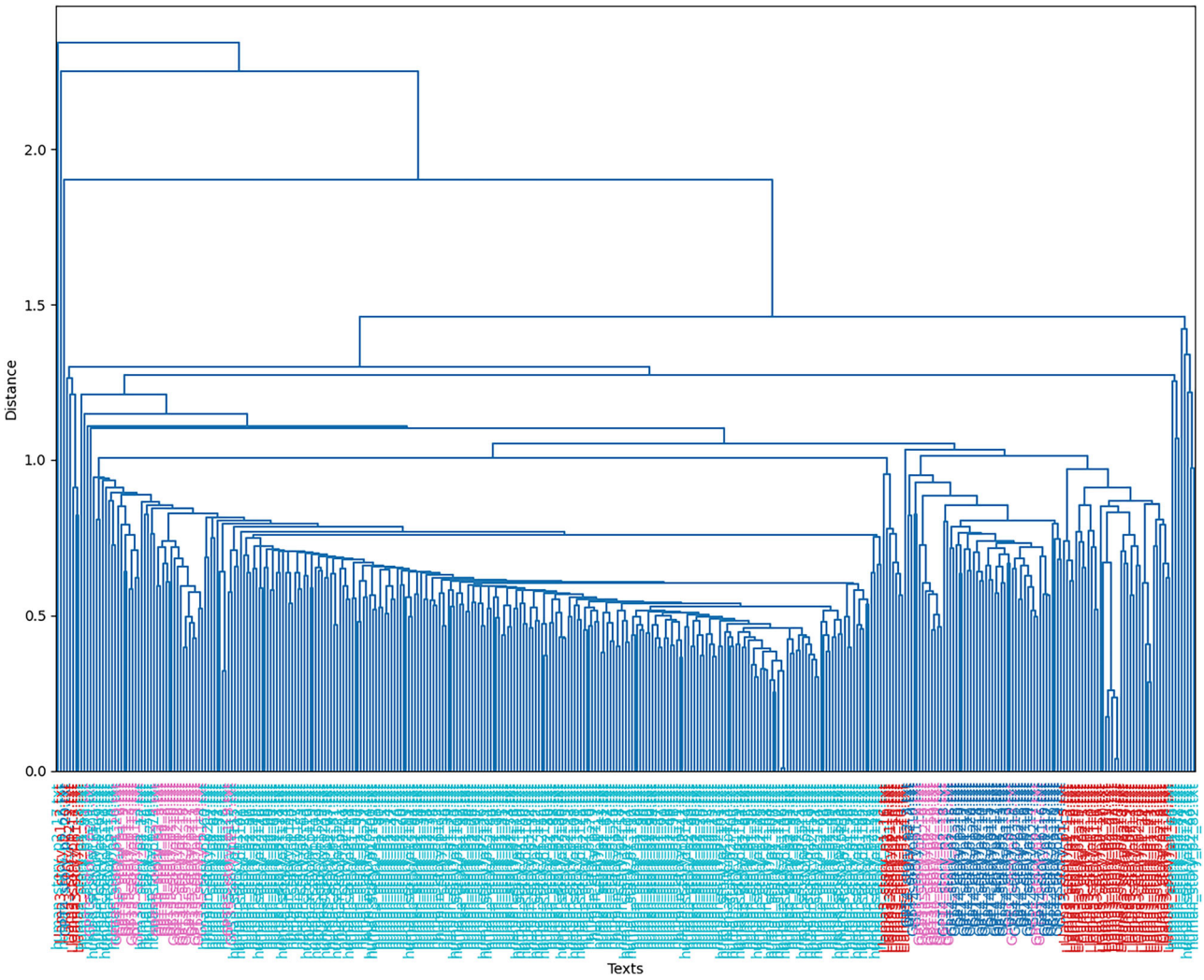


Fig. 1 Hierarchical clustering based on stylistic distances calculated using Burrows' Delta and grouped using the average linkage method. Colour-coded labels indicate groups derived from filename metadata, reflecting the source of each text (e.g., GPT-3.5, GPT-4, Llama 70b, or human-authored).

The prompts were informed by Beguš' previous work on Pygmalionesque fiction's 'Pygmalionesque type', where the humanoid's creator is also her lover, and the 'agalmatophiliac type' where the creator and lover are distinct figures (Beguš, 2023). The same prompts were issued to the large language models.

There are, as will always be the case, limitations with this corpus as a suitable dataset for stylometric analysis. Its reliance on predefined prompts constrains the creative scope of the texts, potentially biasing both human and AI outputs toward specific narrative structures and styles. The crowdsourced human stories, while representative of diverse authors, may vary significantly in quality and adherence to the prompts due to differences in participants' writing experience and engagement, and such 'contrived' texts might not be widely construed as legitimately 'literary'. Despite these challenges, the dataset remains an invaluable resource for examining stylistic and narrative differences between human and AI-generated texts.

Burrows' Delta was applied to the Beguš corpus using Natural Language Toolkit Python scripts (O'Sullivan, 2024). The first script in this analysis employs hierarchical clustering to represent relationships between texts as a dendrogram, a tree-like diagram that visually organises data into clusters based on their stylistic similarities (see Fig. 1). Using the Burrows' Delta measure of distance, the script computes a pairwise comparison of texts and produces a symmetric distance matrix. Hierarchical clustering is then applied to this matrix using the average linkage method, which calculates the average distance between all pairs of points in two clusters during each merging step. Average linkage is particularly well-suited for stylometric applications as it balances between compact clustering and the avoidance of chaining effects that can obscure interpretability.

The second script employs MDS to create a scatter plot that maps texts into a two-dimensional space, where their positions reflect stylistic proximity (see Fig. 2). Like the dendrogram script, it begins by calculating the Burrows' Delta matrix to capture pairwise stylistic distances. MDS then projects these high-dimensional relationships into a lower-dimensional space, allowing for intuitive visualisation. The scatter plot enables a different perspective on the data, highlighting clusters of similar texts and providing a spatial representation of outliers. Each text is represented as a point, labelled with its filename for clarity, and colour-coded based on the same group identifiers used in the dendrogram script.

Together, these two visualisation methods provide complementary approaches to stylometric analysis. While the dendrogram offers a hierarchical view of clustering relationships, MDS excels at providing a spatial layout that emphasises overall proximity and distance between texts.

Limiting the analysis to the 100 MFW reduces the impact of less frequent, content-specific words that are more indicative of a

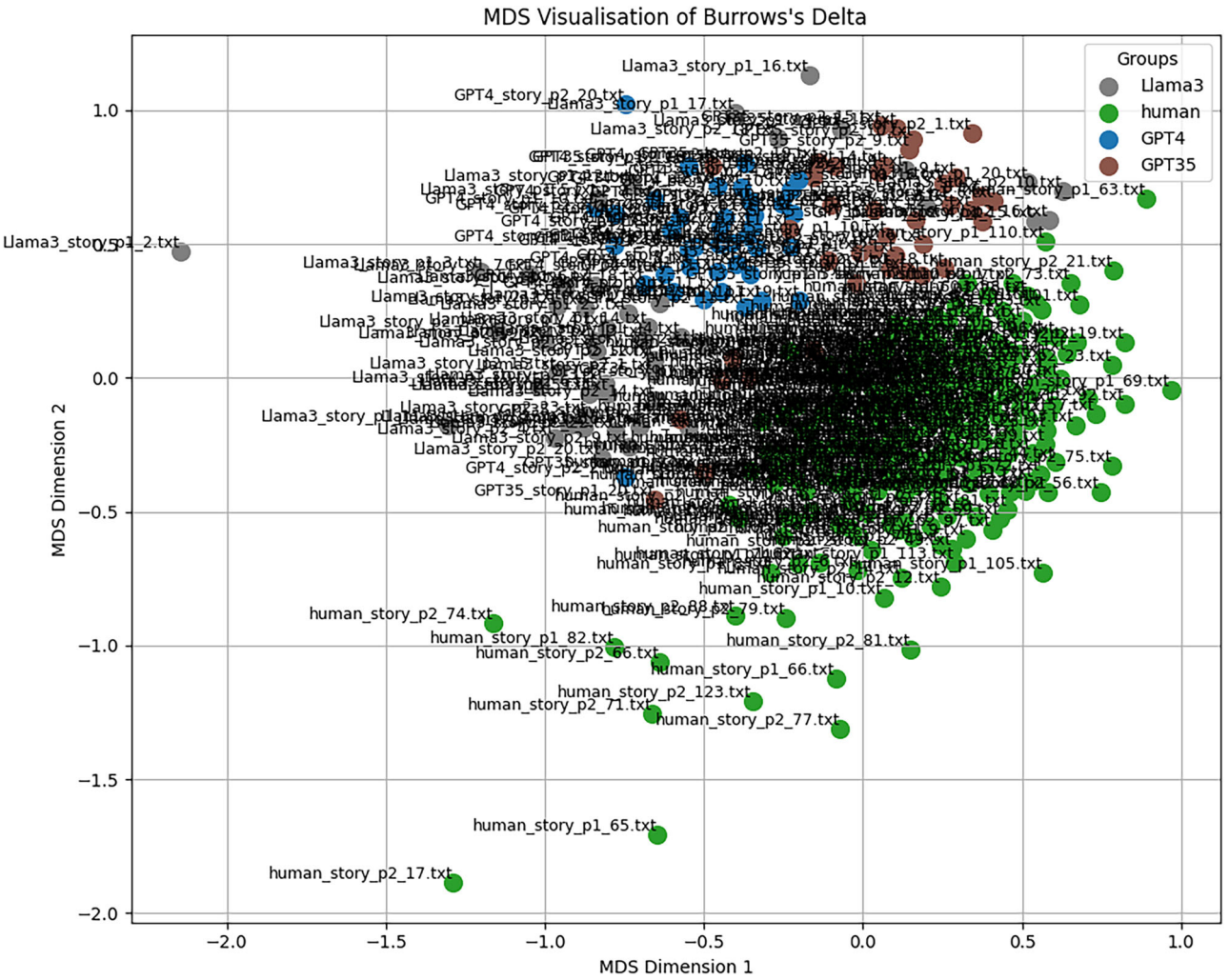


Fig. 2 Multidimensional Scaling (MDS) scatter plot, wherein points represent individual texts, and their positions in two-dimensional space reflect pairwise stylistic distances calculated using Burrows' Delta. Colour-coded points correspond to groups derived from filename metadata.

text's content than writing style. Such words are typically function words, which are ideal for statistical analyses of authorial signatures, because 'they are so frequent that they account for a large proportion of the running words (tokens) of a text, and also because it [is] assumed that their frequencies should be especially resistant to intentional authorial manipulation, and so should reveal authorial habits that remain relatively constant across a variety of texts' (Hoover, 2009, 35).

Findings

Conducting a stylometric analysis of texts produced by humans and LLMs, using Burrows' Delta, shows distinct clusters. These findings suggest that, despite the increasing sophistication of LLMs, the content they produce exhibits stylistic tendencies that remain distinguishable from human writing.

While there are some exceptions, hierarchical clustering demonstrates clear divisions between human and AI-generated texts, with each group forming distinct and interpretable clusters (see Fig. 1). Texts generated by the same LLM—GPT-3.5, GPT-4, or Llama 70b—consistently cluster tightly together, reflecting the uniform stylistic patterns characteristic of these models. Furthermore, these clusters are grouped within a broader generative AI category, which remains stylistically distinct from the clusters formed by human-authored texts.

In contrast, human-authored texts display a much greater degree of variability, forming broader and less compact clusters. This variability likely reflects the diversity of individual writing styles, levels of creativity, and engagement with the predefined prompts. Human writers, even when working within the constraints of a structured task, introduce personal stylistic flourishes, which contribute to the overall heterogeneity of the group. The clustering patterns observed in the dendrogram highlight not only the uniformity of AI-generated texts but also the richer stylistic diversity characteristic of human creativity.

The hierarchical clustering reveals that texts generated by GPT-4 cluster more tightly than those produced by GPT-3.5, reflecting an increased consistency in stylistic patterns within GPT-4's outputs. This tighter clustering suggests that GPT-4 exhibits a more refined stylistic coherence, likely a result of improvements in its training data and architecture. While GPT-3.5 texts show a degree of stylistic variation, indicative of less controlled outputs, GPT-4 demonstrates a higher degree of predictability and uniformity in its writing style. This increased cohesion might also reflect GPT-4's enhanced capacity to follow prompts more consistently and generate text that aligns closely with its training distribution. However, despite this tighter clustering, GPT-4 remains distinct from human-authored texts, which exhibit greater stylistic diversity and broader clusters. The comparison underscores how advancements in LLMs, such as GPT-4, are

refining stylistic consistency but still fall short of capturing the nuanced variability characteristic of human creativity.

The clustering reveals occasional overlaps between GPT-3.5 texts and human-authored texts, with isolated incidences of GPT-3.5 outputs appearing within or near human clusters. These overlaps suggest that GPT-3.5, while generally less stylistically consistent than GPT-4, sometimes produces outputs that mimic human stylistic tendencies closely enough to disrupt the broader distinctions between the two groups. This behaviour may reflect GPT-3.5's relative lack of refinement compared to GPT-4, which can lead to greater variability in its outputs. Such overlaps are exceptions rather than the norm, as GPT-3.5 still forms a recognisable and separate cluster overall, emphasising that while its outputs occasionally blur the line, it does not consistently replicate the stylistics of human writing.

The MDS scatter plot offers a particularly intuitive visualisation with which to discern the clusters (see Fig. 2). By projecting high-dimensional stylistic data into two dimensions, the MDS plot allows for a straightforward examination of proximity and separation between groups. This approach emphasises the spatial relationships between texts, highlighting patterns of similarity and divergence in a more immediately accessible way.

In the MDS plot, human-authored texts occupy a dispersed and less concentrated area, reflecting their greater stylistic diversity and variability. This dispersion suggests that human writing, even within the constraints of predefined prompts, retains a unique individuality that is less present in AI-generated outputs. In contrast, texts produced by each LLM form tighter clusters, revealing the underlying uniformity within the outputs of individual models, as well as the limited stylistic range characteristic of AI systems trained to optimise coherence and fluency over diversity. The distinct positioning of LLM-generated clusters relative to human-authored texts further emphasises the stylistic divide between the two groups, underscoring the human tendency toward richer stylistic variation.

Conclusion

This study employs stylometric analysis, specifically Burrows' Delta, to explore stylistic differences between human-authored and AI-generated creative texts, using short stories prompted by specific narrative scenarios. The results indicate clear stylistic clusters separating human texts from those generated by various large language models, including GPT-3.5, GPT-4, and Llama 70b. These findings suggest that, as far as *creative* writing is concerned, current AI systems still produce text with detectable stylistic signatures distinct from human writing.

The focus on creative texts is important: despite the patterns observed here, stylometric methods should not be treated as tools for academic integrity enforcement. While robust at the corpus level, stylometric analysis relies on statistical regularities unsuited to measuring, say, student assessments:

> Students are not stylistically coherent subjects. Their writing evolves across time and task, reflecting shifting competencies, disciplinary norms, and sometimes the invisible labour of support structures like grammar checkers, tutors, or peer feedback. Some write in their second or third language; others write under duress and anxiety, illness, or distraction. Their stylistic fingerprints are smudged by the very real conditions of learning. And yet stylometric detection tools often assume that deviation from an imagined norm—whether that be the student's prior submissions or a constructed corpus of 'authentic' student work—is a sign of suspicion. This is where the method fails, and not just technically, but ethically (O'Sullivan, 2025).

The use of stylometric evidence as an arbiter of authorship authenticity in educational contexts completely misses its purpose:

> Stylometry is extremely useful for guiding interpretive questions of culture and contributing to matters like historical authorship attribution, but it is wholly inadequate as the sole basis for accusation in matters of academic integrity. To treat it otherwise is to misunderstand the method and to misapply it in ways that can have serious consequences for students and educators alike (O'Sullivan, 2025).

This paper lends stylometric evidence to a very clear question: Can LLMs write creative prose in a style that is statistically indistinguishable from a human? The answer is no, at least, not yet.

However, caution must be exercised to avoid over-interpretation due to the controlled nature of the prompts and the inherent constraints of the dataset. Future studies should build upon these preliminary results by incorporating broader datasets, alternative prompts, and, of course, the most recent LLMs (including, most notably, GPT-5). Exploring qualitative aspects of representative texts alongside quantitative metrics could also deepen our understanding of how stylistic variations between human and AI-generated texts manifest practically.

Given the increasing use of generative AI in writing-intensive tasks, continued research in this area remains crucial, not only to understand stylistic intersections between human creativity and artificial intelligence but also to address broader ethical implications related to authenticity, originality, and responsible authorship.

Data availability

This study makes use of a publicly available dataset compiled by Beguš (2024), accessible via the Open Science Framework: https://doi.org/10.17605/OSF.IO/K6FH7. All stylometric analysis scripts used in this study, including those generating the dendrogram and multidimensional scaling plots, are available on GitHub: https://github.com/jamesosullivan/stylometry. Both resources are cited within the body of the paper where relevant.

Code availability

This study makes use of a publicly available dataset compiled by Beguš (2024), accessible via the Open Science Framework: https://doi.org/10.17605/OSF.IO/K6FH7. All stylometric analysis scripts used in this study, including those generating the dendrogram and multidimensional scaling plots, are available on GitHub: https://github.com/jamesosullivan/stylometry. Both resources are cited within the body of the paper where relevant.

Received: 9 December 2024; Accepted: 12 September 2025;
Published online: 11 November 2025

References

Amirjalili F, Neysani M, and Nikbakht A (2024) Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English Literature. Front Educ 9. https://doi.org/10.3389/feduc.2024.1347421

Argamon S (2008) 'Interpreting Burrows's Delta: geometric and probabilistic foundations. Lit Linguistic Comput. 23(2):131–147. https://doi.org/10.1093/llc/fqn003

Beguš, N (2023) Data for experimental narratives: a comparison of human crowdsourced storytelling and AI storytelling. OSF. https://doi.org/10.17605/OSF.IO/K6FH7

Beguš N (2024) Experimental Narratives: a comparison of human crowdsourced storytelling and AI storytelling. Human Soc Sci Commun 11(1):1–22. https://doi.org/10.1057/s41599-024-03868-8
Burrows J (2002) Delta: a measure of stylistic difference and a guide to likely authorship. Lit Linguistic Comput. 17(3):267–287
Caddy SA (2021) The significance of literary outliers in nineteenth-century British fiction: a stylometric analysis. Arizona State University. https://keep.lib.asu.edu/items/161483
Casal JE, Kessler M (2023) Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. Res Methods Appl Linguist. 2(3):100068. https://doi.org/10.1016/j.rmal.2023.100068
Creamer E (2024) AI poetry rated better than poems written by humans, study shows. Guardian 18:2024, https://www.theguardian.com/books/2024/nov/18/ai-poetry-rated-better-than-poems-written-by-humans-study-shows Novembersec. Books
Eder M (2017) Elena Ferrante: a virtual author. In: Tuzzi A, Cortelazzo MA (eds) Drawing Elena Ferrante's profile, Padova University Press, Padova, pp 31–45. https://www.research.unipd.it/retrieve/handle/11577/3274069/238296/2018_Tuzzi_Cortelazzo_PUP_Ferrante_9788869381300.pdf#page=32
Evert S, Proisl T, Jannidis F, Reger I, Pielström S, Schöch C, Vitt T (2017) Understanding and explaining delta measures for authorship attribution. Digit Scholarsh Human 32(suppl_2):ii4–ii16. https://doi.org/10.1093/llc/fqx023
Gunser VE, Gottschling S, Brucker B, Richter S, Gerjets P (2021) Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text? In: Stephanidis C, Antona M, Ntoa S (eds) HCI International 2021—Posters, Springer International Publishing, Cham, pp 520–527. https://doi.org/10.1007/978-3-030-78635-9_67
Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A (2023) A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports 13:18617. https://doi.org/10.1038/s41598-023-45644-9
Hoover DL (2007) 'Corpus stylistics, stylometry, and the styles of Henry James. Style 41(2):174–203
Hoover DL (2009) Word frequency, statistical stylistics and authorship attribution. In What's in a Word-list? Investigating Word Frequency and Keyword Extraction, edited by Dawn Archer, Routledge, 35–51
Jannidis F, Pielström S, Schöch C, Vitt T (2015) Improving Burrows' Delta—an empirical evaluation of text distance measures. Paper presented at DH 2015—Global Digital Humanities, Western Sydney University, Sydney. https://dh-abstracts.library.cmu.edu/works/2275
Neal T, Sundararajan K, Fatima A, Yan Y, Xiang Y, Woodard D(2017) Surveying stylometry techniques and applications ACM Comput. Surv. 50(6):86:1–86:36. https://doi.org/10.1145/3132039
Opara C (2024) StyloAI: distinguishing AI-generated content with stylometric analysis. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial intelligence in education. posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky, Springer Nature Switzerland, Cham, 105–114. https://doi.org/10.1007/978-3-031-64312-5_13
O'Sullivan J (2024) Jamesosullivan/Stylometry. Python. https://github.com/jamesosullivan/stylometry
O'Sullivan J (2025) Stylometry should not be higher education's answer to ChatGPT. Substack newsletter. James O'Sullivan's Substack (blog). 23 April 2025. https://jamescosullivan.substack.com/p/stylometry-shouldnt-be-higher-educations-answer-to-chatgpt
O'Sullivan J, Bazarnik K, Eder M, Rybicki J (2018) Measuring Joycean influences on Flann O'Brien. Digit Stud/Champ Numér 8. https://doi.org/10.16995/dscn.288
Porter B, Machery E (2024) AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Sci Rep 14(1):26133. https://doi.org/10.1038/s41598-024-76900-1
Przystalski K, Argasiński J, Grabska-Gradzińska I, and Ochab J. (2024) Stylometry recognizes human and LLM-generated texts in short samples. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.4950812
Rybicki J, Heydel M (2013) The stylistics and stylometry of collaborative translation: Woolf's night and day in Polish. Lit Linguist Comput. https://doi.org/10.1093/llc/fqt027
Savoy J (2020) Machine learning methods for stylometry: authorship attribution and author profiling. Springer Nature, Cham
Savoy J (2023) Stylometric analysis of characters in Shakespeare's plays. Digit Scholarsh Human 38(3):1238–1246. https://doi.org/10.1093/llc/fqac092
Wang S, Cristianini N, Hood BM (2024) Stylometric comparison between ChatGPT and human essays. In: Abstracts of the 18th international AAAI conference on web and social media. https://doi.org/10.36190/2024.32
Zaitsu W, Jin M (2023) Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. PLoS ONE 18(8):e0288453. https://doi.org/10.1371/journal.pone.0288453

Author contributions

JOS conceptualised and designed the study, conducted the research, performed the analysis, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Ethical approval

This research did not require approval from an institutional review board or ethics committee. The study relied exclusively on a pre-existing, openly available dataset of anonymised human–and AI-generated short stories compiled by Beguš (2024) and hosted on the Open Science Framework. No new data were collected, and no interaction with human participants took place. According to Irish and EU research ethics frameworks, as well as the Declaration of Helsinki, secondary analysis of anonymised and publicly archived data does not require additional ethical approval.

Informed consent

Informed consent was not obtained for this study because no new human data were collected. The dataset used was created by Beguš (2024) via Amazon Mechanical Turk under a separate research protocol that included participant consent and anonymisation. The present research involved only secondary analysis of these anonymised materials, which are publicly accessible through the Open Science Framework. Under GDPR Recital 26 and Irish Health Research Regulations 2018, anonymised data that cannot be linked to identifiable individuals is outside the scope of consent requirements.

Additional information

Correspondence and requests for materials should be addressed to James O'Sullivan.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025