The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 4 | Issue 2 | February 2026
ISSN: 2583-973X (Online), SJIF IF: 5.6
Website: www.theacademic.in

Can Machines Write Literature? Evaluating AI-Generated Fiction against Human Authorship Standards

Dr. Arohi Sarin

Senior Assistant Professor, Department Of English, D. A-V. College, Kanpur Nagar

Dr. Shanker Singh Solanki

Senior Assistant Professor, Department of English Studies & Research, D. A-V. College, Kanpur Nagar

DOI : https://doi.org/10.5281/zenodo.18872748

ARTICLE DETAILS

Research Paper

Accepted: 17-02-2026

Published: 10-03-2026

Keywords:

artificial intelligence, creative writing, large language models, literary evaluation, computational creativity, natural language generation, narrative theory

ABSTRACT

The rapid development of large language models (LLMs) has prompted urgent questions about the nature, quality, and authenticity of machine-generated literary fiction. This paper presents a comprehensive multidimensional evaluation framework for assessing AI-generated fiction against established human authorship standards. Drawing on empirical studies, computational linguistics, reader-response theory, and narratology, we analyze the capacity of contemporary AI systems—including GPT-4, Claude 3, and PaLM 2—to produce fiction that meets literary criteria across seven key dimensions: narrative coherence, emotional depth, character authenticity, stylistic originality, thematic complexity, structural integrity, and reader engagement. Our analysis of 42 experimental studies involving over 15,000 human readers reveals that while AI models achieve near-human performance on surface linguistic measures, they exhibit significant deficiencies in emotional authenticity, subtext generation, moral ambiguity, and cultural specificity. We further examine the philosophical implications of machine authorship, the role of intentionality in literary meaning, and the ethical dimensions of AI creative writing in publishing contexts. Our findings suggest that AI-generated fiction currently occupies a distinct ontological category—neither fully literature nor

mere text—and that meaningful evaluation requires rethinking traditional authorship standards in light of computational creativity.

1. INTRODUCTION

The emergence of sophisticated large language models has fundamentally altered the landscape of creative writing. Systems capable of generating grammatically fluent, contextually coherent, and stylistically variable text have moved from academic novelty to commercial product within a remarkably compressed timeframe (Brown et al., 2020). As these systems become increasingly capable and accessible, they raise profound questions: Can artificial intelligence produce literature in any meaningful sense? What standards should govern the evaluation of machine-generated fiction? And what, ultimately, distinguishes a great novel from a statistically plausible sequence of words?

These are not merely technical questions. They engage the oldest debates in literary theory—concerning authorship, intention, authenticity, and the nature of aesthetic experience. When Walter Benjamin (1935) wrote about the 'aura' of the artwork, he could not have anticipated an era in which the distinction between original and reproduction would be challenged not by mechanical reproduction but by computational generation. Yet his concerns about authenticity and the erosion of singular creative vision have acquired new urgency in the age of transformer-based language models (Vaswani et al., 2017).

This paper addresses these debates through a systematic, multidimensional evaluation of AI-generated fiction. We begin by reviewing existing literature on computational creativity and natural language generation. We then present an analytical framework for literary evaluation, followed by an empirical comparison of AI-generated and human-authored fiction across multiple quality dimensions. We conclude with a discussion of the theoretical and ethical implications of machine authorship for the future of literary culture.

2. LITERATURE REVIEW

2.1 The History of Computational Creativity

Attempts to automate creative writing predate modern computing. The 'Stochastic Poet' constructed by Christopher Strachey in 1952 and the work of Theo Lutz on Stochastic Texts (1959) represent early experiments in algorithmically generated prose and poetry. These systems, however, operated through simple combinatorial rules and lacked the capacity for extended narrative coherence (Funk houser, 2007).

The rule-based systems of the 1970s and 1980s, including the TALE-SPIN story generator developed by Meehan (1977) and the AUTHOR system of Dunn (1983), introduced planning-based approaches that could generate simple stories with causal structure, though at the cost of rigid, predictable output.

2.2 Large Language Models and Fiction Generation

The GPT series of models developed by Open AI represents the current state of the art in general-purpose language generation. GPT-2, released in 2019, demonstrated the capacity to generate plausible short fiction with surprising coherence; GPT-3 (Brown et al., 2020) scaled this capability dramatically, producing outputs that many readers found indistinguishable from human writing in short-form tasks. GPT-4 further extended these capabilities, demonstrating improved factual accuracy, reduced hallucination rates, and greater sensitivity to contextual constraints (Open AI, 2023).

Clark et al. (2021) conducted a large-scale annotation study examining long-form narrative coherence, finding that AI-generated stories exhibited a characteristic 'coherence cliff'—high local coherence within paragraphs that deteriorated sharply over extended narratives. This finding is consistent with the theoretical observations of Bender et al. (2021), who argued that LLMs lack the 'grounded understanding' required for genuine narrative comprehension, operating instead through sophisticated surface-level pattern matching.

2.3 Theoretical Frameworks for Literary Evaluation

The evaluation of literary quality is itself a contested theoretical domain. Formalist approaches, drawing on the work of Russian Formalists and New Critics, prioritize structural properties such as unity, coherence, and the strategic deployment of literary devices (Brooks, 1947; Wellek & Warren, 1949). Reader-response theories, associated with scholars such as Stanley Fish (1980) and Wolfgang Iser (1978), shift attention to the reader's affective and cognitive engagement with texts. Narrative theory, developed by Genette (1980), Chatman (1978), and more recently Fludernik (1996), provides detailed analytical frameworks for examining point of view, temporality, focalization, and narrative voice.

3. METHODOLOGY

3.1 Research Design

This study employs a mixed-methods approach combining systematic literature review, quantitative analysis of evaluation data from 42 published empirical studies (2018–2024), and original qualitative

analysis of selected human-authored and AI-generated texts. We examine fiction generated by three leading LLMs (GPT-4, Claude 3, and PaLM 2) and compare it to human-authored fiction across seven evaluative dimensions derived from our theoretical framework. Expert evaluators (literary scholars with PhDs in literature or creative writing) rated each text on a 10-point Likert scale across all dimensions. A separate panel of general readers (n=600) provided engagement and preference ratings.

3.2 Evaluation Dimensions

Our seven-dimensional evaluation framework was derived through a combination of theoretical analysis and expert consultation. The dimensions are: (1) Narrative coherence—the degree to which plot, causality, and temporal structure cohere across the full extent of a text; (2) Emotional depth—the capacity to evoke genuine emotional responses in readers and to represent complex interior emotional states convincingly; (3) Character authenticity—the degree to which characters behave consistently, unpredictably, and in accordance with psychologically plausible motivations; (4) Stylistic originality—evidence of distinctive, idiosyncratic style that cannot be attributed to simple imitation of training data; (5) Thematic complexity—the capacity to sustain, develop, and complicate central themes across a text; (6) Structural integrity—the deliberate deployment of narrative structure, including pacing, framing, and formal experimentation; and (7) Reader engagement—measured through self-reported engagement, reading time, and inclination to recommend.

Table 1: Comparative Quality Ratings — Human Authors vs. AI Models Across Seven Literary Dimensions (Expert Panel, n=45)

Evaluation Criterion	Human Authors	GPT-4 (AI)	Claude 3 (AI)	Expert Rating (1–10)
Narrative Coherence	9.4	7.8	8.1	H: 9.4 / AI: 7.9
Emotional Depth	9.1	6.9	7.3	H: 9.1 / AI: 7.1
Character Authenticity	8.9	6.5	6.8	H: 8.9 / AI: 6.6
Stylistic Originality	8.7	5.9	6.2	H: 8.7 / AI: 6.1
Thematic Complexity	9.0	6.3	6.7	H: 9.0 / AI: 6.5

Structural Integrity	8.8	7.5	7.7	H: 8.8 / AI: 7.6
Reader Engagement	9.2	7.1	7.4	H: 9.2 / AI: 7.2

4. RESULTS AND ANALYSIS

4.1 Surface Fluency vs. Deep Literary Quality

Our analysis confirms the widely reported finding that contemporary AI models have achieved near-human performance on surface linguistic measures. Across the 42 studies examined, AI-generated texts were rated comparably to human texts on measures of grammatical correctness, vocabulary diversity, and local sentence-level coherence. In several studies, AI texts actually exceeded human texts on formal measures of readability (Kincaid et al., 1975), producing prose that was syntactically cleaner and more uniformly clear than the sometimes deliberately obscure or experimentally complex writing of literary authors.

However, the performance gap widens dramatically when evaluators move from surface measures to deeper literary qualities. As Table 1 illustrates, human authors score significantly higher than AI systems on all seven dimensions of our framework, with the largest gaps appearing in emotional depth (9.1 vs. 7.1), character authenticity (8.9 vs. 6.6), and stylistic originality (8.7 vs. 6.1). These gaps are statistically significant ($p < 0.001$) across all dimensions and robust to variation in evaluator background and genre conventions.

The finding regarding stylistic originality is particularly significant. While AI models can produce texts that superficially resemble the style of specific human authors (Floridi & Cowls, 2019), they struggle to produce styles that are genuinely novel—styles that could not be explained as combinations of existing human styles present in the training data. This limitation is theoretically expected, given that language models are fundamentally trained to predict what text a human might have written (Marcus, 2022). They model the distribution of existing human text rather than generating from any novel expressive intention.

Table 2: Summary of Key Empirical Studies on AI Fiction Evaluation (2019–2023)

Study / Researcher	Year	Methodology	Key Finding
Elkins & Chun	2020	Blind Reading	Readers could not reliably distinguish AI

		Study	prose from human prose in 50% of trials
Ippolito et al.	2020	Turing-Test Style Evaluation	AI text scored lower on creativity but comparable on fluency
Gero & Chilton	2019	Metaphor Generation Test	Humans preferred human-generated metaphors 68% of the time
Clark et al.	2021	Crowd sourced Annotation	AI stories lacked long-range coherence across extended narratives
Bender et al.	2021	Critical Discourse Analysis	AI models reproduce patterns without grounded understanding
Chakrabarty et al.	2022	Creative Writing Evaluation	GPT-3 outperformed humans on surface fluency but not depth
Yang et al.	2023	Automated Metric Analysis	BLEU/ROUGE scores insufficient for literary quality evaluation

4.2 Specific Literary Deficiencies in AI Fiction

Beyond the aggregate dimensional comparisons, our analysis identifies specific recurring deficiencies in AI-generated fiction that distinguish it from human authorship at a qualitative level. Table 3 presents a comparison across specific literary elements, based on expert coding of 120 text samples (60 human, 60 AI-generated).

Table 3: Comparative Performance on Specific Literary Elements — Expert Coding of 120 Text Samples

Literary Element	Human Score (%)	AI Score (%)	Gap (%)	Significance
Metaphor & Figurative Language	87	61	26	High
Dialogue Naturalness	83	72	11	Moderate
Plot Unpredictability	81	54	27	High

Subtext & Implication	88	49	39	Very High
Cultural Specificity	84	57	27	High
Moral Ambiguity	86	52	34	Very High
Temporal Pacing	82	68	14	Moderate
Voice Consistency	90	74	16	Moderate

The most striking gap in Table 3 concerns subtext and implication (39 percentage points), reflecting a fundamental limitation of current LLMs. Human authors routinely communicate meaning through what is unsaid—through strategic omission, ironic juxtaposition, and the gap between characters' stated beliefs and their observable behavior. This capacity for meaningful silence requires a model of the reader's interpretive processes and an intention to exploit the gap between surface meaning and deeper significance (Sternberg, 1978). Current AI systems, which generate text by predicting the most likely next token given the context, have no mechanism for deliberately suppressing information for rhetorical effect.

Similarly substantial gaps appear in moral ambiguity (34 percentage points) and plot unpredictability (27 percentage points). AI-generated fiction tends toward resolution and clarity; stories generated by GPT-4 and Claude 3 in our sample consistently moved toward morally legible conclusions, avoiding the ethical ambiguity and unresolved tension that characterize much of the literary canon from Dostoevsky to Toni Morrison. This tendency likely reflects the models' training on large corpora of popular fiction, which tends toward more conventional narrative resolutions, as well as the reinforcement learning from human feedback (RLHF) process, which may inadvertently penalize morally unsettling content (Ouyang et al., 2022).

5. DISCUSSION

The most philosophically fundamental question raised by AI-generated fiction concerns intentionality. Human literary authorship is understood as an intentional act: the author intends to communicate specific meanings, to evoke particular emotions, to challenge certain assumptions, and to produce specific aesthetic effects (Hirsch, 1967). This intentionality is not merely a psychological fact about the author's mental states; it shapes how readers interpret texts and how critics evaluate them. The concept of the

'implied author' (Booth, 1961) and the related notion of authorial voice both presuppose an intentional agent behind the text.

AI systems do not have intentions in any philosophically robust sense. They generate text by computing probability distributions over possible continuations of a given prompt; the 'choices' they make reflect statistical regularities in training data rather than communicative intentions. This does not mean that AI-generated texts cannot be meaningful—readers may construct meanings for any text, regardless of its origin (Barthes, 1967)—but it does raise the question of whether such meanings are properly attributed to the text itself or entirely to the reader's interpretive activity.

7. CONCLUSION

This paper has examined the capacity of AI-generated fiction to meet established standards of human literary authorship across seven key dimensions. Our analysis reveals a consistent and theoretically interpretable pattern: AI systems have achieved near-human performance on surface linguistic qualities while remaining significantly below human standards on the deeper dimensions of literary quality that constitute genuine literature—emotional authenticity, character depth, stylistic originality, and thematic complexity.

These limitations are not simply engineering problems that will inevitably be solved by larger models and more training data. They are rooted in fundamental differences between statistical text prediction and the intentional, experiential, and culturally situated practice of human authorship. Literature, at its most significant, is a form of testimony—a record of human consciousness grappling with the conditions of existence. This testimonial dimension of literature may be irreducibly dependent on the kind of embodied, phenomenological subjectivity that current AI systems lack.

What is clear is that the question 'Can machines write literature?' cannot be answered without first answering harder questions about what literature is for, what authorship means, and what readers deserve to know about the texts they read. These are not merely technical questions. They are questions about the nature of human communication, the value of authentic expression, and the ethical responsibilities of those who produce and distribute fictional texts in an age of artificial intelligence. They demand the sustained, collaborative attention of technologists, literary scholars, ethicists, and the reading public alike.

REFERENCES

- Agüera y Arcas, B. (2022). Do large language models understand us? Daedalus, 151(2), 183–197.
- Alhussain, A., & Alhazza, A. (2021). Automatic story generation: A survey of approaches. ACM Computing Surveys, 54(5), 1–38.
- Authors Guild. (2023). Authors Guild AI survey results. Authors Guild Publications.
- Barthes, R. (1967). The death of the author. Aspen, 5–6. (Translated by S. Heath)
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of FAccT 2021 (pp. 610–623). ACM.
- Benjamin, W. (1935/2008). The work of art in the age of mechanical reproduction. Penguin.
- Booth, W. C. (1961). The rhetoric of fiction. University of Chicago Press.
- Brooks, C. (1947). The well-wrought urn: Studies in the structure of poetry. Reynal & Hitchcock.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
- Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799.
- Chakrabarty, T., Padaki, K., Iyengar, S., & Muresan, S. (2022). It's not rocket science: Interpreting figurative language in narratives. Transactions of the Association for Computational Linguistics, 10, 589–606.
- Chatman, S. (1978). Story and discourse: Narrative structure in fiction and film. Cornell University Press.

- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. In Proceedings of ACL-IJCNLP 2021 (pp. 7282–7296).
- Colton, S., Pease, A., & Ritchie, G. (2012). The evaluation of creative systems. In Proceedings of the International Conference on Computational Creativity.
- Dennett, D. C. (1991). Consciousness explained. Little, Brown.
- Dreyfus, H. L. (1972). What computers can't do: A critique of artificial reason. Harper & Row.
- Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing test? Journal of Cultural Analytics, 5(2).
- Fish, S. (1980). Is there a text in this class? The authority of interpretive communities. Harvard University Press.
- Floridi, L. (2019). The logic of information: A theory of philosophy as conceptual design. Oxford University Press.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).
- Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(3), 1771–1796.
- Fludernik, M. (1996). Towards a 'natural' narratology. Routledge.
- Funkhouser, C. T. (2007). Prehistoric digital poetry: An archaeology of forms, 1959–1995. University of Alabama Press.
- Genette, G. (1980). Narrative discourse: An essay in method. Cornell University Press.
- Gero, K. I., & Chilton, L. B. (2019). Metaphoria: An algorithmic companion for metaphor creation. In Proceedings of CHI 2019 (pp. 1–12). ACM.

- Grimmelmann, J. (2023). The mouse that roared: Bringing order to the AI copyright maze. Yale Law Journal Forum.
- Hirsch, E. D. (1967). Validity in interpretation. Yale University Press.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
- Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In Proceedings of ACL 2020 (pp. 1808–1822).
- Iser, W. (1978). The act of reading: A theory of aesthetic response. Johns Hopkins University Press.
- LeCun, Y. (2022). A path towards autonomous machine intelligence. Open Review Preprint.
- Malhotra, R. (2021). Algorithmic authorship and literary authenticity: A critical examination. Digital Humanities Quarterly, 15(3).
- Marcus, G. (2022). Deep learning is hitting a wall. Nautilus, 95.
- Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon.
- Meehan, J. R. (1977). TALE-SPIN, an interactive program that writes stories. In IJCAI (Vol. 77, pp. 91–98).
- Nagel, T. (1974). What is it like to be a bat? Philosophical Review, 83(4), 435–450.
- Nguyen, H. T. (2022). A multi-criteria evaluation framework for AI creative writing systems. ACM Transactions on Intelligent Systems and Technology, 13(4), 1–28.
- OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

- Rettberg, J. W. (2019). Ways of knowing with data visualizations. In C. Burdick et al. (Eds.), Digital_humanities (pp. 98–115). MIT Press.
- Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424.
- Sternberg, M. (1978). Expositional modes and temporal ordering in fiction. Johns Hopkins University Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
- Waddell, C. (2023). AI-generated content and the future of literary publishing. Publishers Weekly.
- Wellek, R., & Warren, A. (1949). Theory of literature. Harcourt, Brace.
- Wordsworth, W. (1800/1991). Preface to the Lyrical Ballads. In W. Wordsworth & S. T. Coleridge, Lyrical Ballads. Routledge.
- Yang, D., Bhatt, U., Schölkopf, B., & Lobanov, E. (2023). Beyond BLEU: Toward a multidimensional framework for evaluating AI-generated creative text. Computational Linguistics, 49(1), 45–91.