
Beyond Stylometry: Measuring Narrative Identity Preservation in AI-Rewritten Literary Fiction Through Computational and Literary Evaluation

Ade Fitria Fatimah^{1*}, Mohamad Nasir²

Universitas Gadjah Mada, Indonesia¹

Politeknik Siber Cerdika Internasional, Indonesia²

Corresponding Author: adefitriafatimah997@gmail.com^{1*}

Keywords:

artificial intelligence;
computational stylometry; literary
fiction; narrative identity
preservation; digital humanities

ABSTRACT

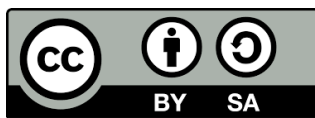
Background: The increasing use of generative artificial intelligence in literary rewriting raises concerns about whether semantic similarity adequately reflects the preservation of narrative identity, including stylistic distinctiveness, narrative voice, characterization, and interpretive complexity.

Objective: This conceptual study aims to develop a multidimensional framework and an empirically testable validation protocol for evaluating narrative identity preservation in AI-rewritten literary fiction by integrating computational and literary perspectives.

Method: A conceptual integrative review was conducted on six purposively selected empirical studies published in 2025 that address computational stylometry, literary style imitation, fiction evaluation, reader experience, and interpretive assessment, supported by foundational works in narratology, stylometry, and literary natural language processing. Thematic synthesis was used to derive the framework, which was then operationalized into a paired source–rewrite experimental protocol.

Results: The synthesis yielded four themes: the limits of stylometric similarity, the gap between semantic fidelity and narrative structure, discrepancies between automated and human literary judgment, and the need for source-relative evaluation. These themes were integrated into a three-layer framework comprising semantic, stylistic, and narratological continuity, each linked to specified indicators, annotation criteria, and validation tests.

Conclusion: Narrative identity preservation is better conceptualized as a multidimensional, source-relative construct than as semantic similarity alone. The framework offers methodological guidance for computational literary studies and responsible AI-assisted adaptation, although its reliability and predictive validity require empirical confirmation.



This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Copyright © 2026 by Author. Published by
Pusat Publikasi Nusantara

INTRODUCTION

The rapid advancement of artificial intelligence has transformed the production, circulation, and computational analysis of literary texts, creating a research landscape in which computer science and literary studies increasingly converge. Large language models now generate coherent fictional narratives, imitate established literary styles, and reconstruct existing textual expressions, thereby challenging conventional assumptions about the distinctiveness of human literary authorship (O’Sullivan, 2025).

Because AI-assisted writing is becoming relevant to publishing, translation, education, cultural preservation, and digital humanities, these developments raise fundamental questions concerning originality, stylistic authenticity, and the preservation of literary meaning. In

multilingual and culturally diverse contexts such as Indonesia, the implications extend to local storytelling traditions and culturally embedded narrative conventions that depend on distinctive linguistic and interpretive practices.

Computational studies show that generative models can reproduce recognizable features of literary expression while remaining distinguishable from human-authored texts through quantitative stylistic patterns (Mikros, 2025), and comparable concerns have emerged in literary machine translation, where fluent document-level output may still contain critical errors (Karpinska & Iyyer, 2023; Thai et al., 2022). Understanding how artificial intelligence transforms existing narratives therefore requires an analytical perspective that moves beyond technological performance toward the systematic examination of semantic continuity, narrative structure, and stylistic individuality.

Theoretically, this study draws on narratology, stylistics, and literary natural language processing (NLP). Classical narratology distinguishes the story that is told from the discourse through which it is told, including narrative voice, focalization, temporal order, and narrative distance (Bal, 2017). Within this tradition, narrative identity preservation is defined here as the extent to which a rewritten passage retains the distinctive configuration of expressive and narratological properties through which the original constructs meaning and organizes fictional experience.

Stylometry offers complementary tools: Burrows' Delta and related measures quantify stylistic difference through the distribution of frequent words (Burrows, 2002; Evert et al., 2017), and such techniques have recently been used to distinguish human-authored from AI-generated creative writing (O'Sullivan, 2025). Literary NLP has further made selected narrative properties computationally tractable through annotated resources for literary entities and events (Bamman et al., 2019; Sims et al., 2019), and through calls to ground computational narrative understanding explicitly in narrative theory (Piper et al., 2021).

Meanwhile, multidimensional benchmarks evaluate generated fiction at macro, meso, and micro levels (Wang et al., 2025), and surface stylistic resemblance has been shown not to guarantee reproduction of an author's broader stylistic signature (Mikros, 2025). Accordingly, semantic fidelity, which concerns propositional content and narrative events, should be distinguished from narrative identity preservation, which concerns the stylistic and narratological relationships through which those events acquire literary significance.

Despite this progress, it remains unresolved whether existing similarity measures can determine how far AI rewriting preserves the narrative identity of a source passage. Stylometric comparisons primarily address authorship-related differences rather than preservation across paired original and rewritten texts (O'Sullivan, 2025), whereas fiction benchmarks evaluate generated outputs rather than source-relative transformation (Wang et al., 2025). Research on text style transfer similarly evaluates content preservation, style strength, and fluency, yet it rarely operationalizes narratological properties such as focalization or interpretive ambiguity (Jin et al., 2022; Mir et al., 2019).

Consequently, a rewrite may retain characters, events, and propositions while altering narrative distance, emotional implication, or figurative expression. The problem is urgent because AI-generated stories differ from human-authored stories in linguistic properties and elicit lower narrative transportation (Appel et al., 2025), expert assessments indicate that LLM-generated stories fall short of professional creative-writing standards (Chakrabarty et al., 2024), generative AI shows limitations in the interpretive depth of literary analysis (Chan & Wong, 2025), and automated evaluators may diverge from human judgment on culturally specific and context-dependent qualities (Kim & Oh, 2025). Without a validated framework, researchers and practitioners may overestimate the fidelity of AI-rewritten literature when semantic similarity or fluency is treated as sufficient evidence that the original narrative experience has been preserved.

Previous studies therefore address related but distinct problems: authorship discrimination (O’Sullivan, 2025), style imitation (Mikros, 2025), quality evaluation of generated fiction (Wang et al., 2025), and reader response (Appel et al., 2025). To the authors’ knowledge, none explicitly conceptualizes narrative identity preservation as a source-relative construct that integrates semantic, stylometric, and narratological evidence and specifies how such evidence should be validated against independent literary judgment.

The novelty of this study lies in three contributions: (a) a narratologically grounded distinction between semantic fidelity and narrative identity preservation; (b) a three-layer evaluation framework that links each layer to explicit computational indicators and literary annotation criteria; and (c) a paired source–rewrite validation protocol that tests the incremental validity of narratological features over semantic and stylometric baselines rather than assuming that methodological complexity improves accuracy. Because the experiment has not yet been executed, the framework is presented as a testable proposition whose empirical value must be established through the protocol described in this article.

Accordingly, this conceptual study aims to (1) synthesize recent computational and literary scholarship to clarify how narrative identity preservation differs from semantic similarity and stylistic resemblance, (2) develop a multidimensional framework for evaluating narrative identity preservation in AI-rewritten literary fiction, and (3) operationalize the framework into a reproducible experimental and validation protocol. Theoretically, the study clarifies the relationship among computational stylistics, narratology, and the evaluation of generative language technologies. Methodologically, it provides indicators, annotation criteria, and statistical procedures that future studies can implement and test. Practically, it informs the responsible application of artificial intelligence in literary adaptation, digital humanities, publishing, and education by encouraging evaluation practices that retain human interpretive oversight.

METHOD

Research Design and Type of Research

This study employed a conceptual research design that combined an integrative literature review with protocol development (Jaakkola, 2020). An integrative, non-systematic review was selected because the aim was to synthesize heterogeneous computational and literary evidence into a theoretical framework rather than to estimate pooled effects; this approach corresponds to the critical review type in established review typologies, which prioritizes conceptual contribution over exhaustive coverage (Grant & Booth, 2009).

The study proceeded in three stages. First, a core set of recent empirical studies was purposively selected and analyzed. Second, the extracted evidence was thematically synthesized and combined with foundational narratological, stylometric, and literary NLP sources to construct the conceptual framework. Third, the framework was operationalized into a controlled paired source–rewrite experimental protocol that specifies the corpus, generative models, rewriting conditions, measurements, annotation procedure, and statistical analyses required for its empirical validation. The study therefore reports conceptual and methodological outcomes; it does not report experimental results, which will be generated when the protocol is implemented.

Data Sources and Selection Strategy

The data consisted of published scholarly literature. Candidate sources were identified through non-systematic searching of academic databases (Scopus, Web of Science Core Collection, Google Scholar, ScienceDirect, and SpringerLink) and through citation tracing, using combinations of keywords related to artificial intelligence, generative language models, literary rewriting, computational stylometry, narrative identity, semantic similarity, narratology, and computational literary evaluation. Because screening decisions were not

documented at the record level, the study does not claim systematic coverage and does not report a PRISMA flow diagram.

Instead, six empirical studies published in 2025 were purposively selected as the core evidence base because they directly examined stylometric differentiation (O’Sullivan, 2025), literary style imitation (Mikros, 2025), multidimensional fiction evaluation (Wang et al., 2025), linguistic properties and reader experience (Appel et al., 2025), LLM-based creativity assessment (Kim & Oh, 2025), and AI-assisted literary interpretation (Chan & Wong, 2025). These studies were complemented by foundational works in narratology (Bal, 2017), stylometry (Burrows, 2002; Eder et al., 2016; Evert et al., 2017), literary NLP (Bamman et al., 2019; Piper et al., 2021; Sims et al., 2019), semantic similarity (Reimers & Gurevych, 2019), style transfer and literary translation evaluation (Jin et al., 2022; Karpinska & Iyyer, 2023; Mir et al., 2019; Thai et al., 2022), and annotation reliability (Hayes & Krippendorff, 2007; Krippendorff, 2018), which were used to define constructs and select measurement instruments rather than as part of the thematic synthesis.

Inclusion and Exclusion Criteria

Core studies were included when they (a) were peer-reviewed journal articles or full conference papers published in English, (b) appeared between 2021 and 2026, (c) reported original empirical evidence on generative AI in relation to literary or creative writing, and (d) addressed at least one dimension relevant to narrative identity preservation, namely stylistic distinctiveness, semantic or narrative content, reader experience, or the validity of literary evaluation. Studies were excluded when they focused exclusively on non-literary genres, provided insufficient methodological information, or were not accessible in full text. Foundational publications outside the 2021–2026 window were retained only for theoretical definition and instrument selection.

For the experimental corpus specified in the protocol, literary passages are eligible only when their reuse is legally permissible, their textual boundaries are clearly identifiable, and their narrative properties are sufficiently developed to support comparative computational and literary analysis. The PRISMA 2020 statement and its PRISMA-S extension are adopted as the reporting standard for the systematic search that will accompany the empirical phase (Page et al., 2021; Rethlefsen et al., 2021).

Data Extraction and Synthesis Procedures

A standardized extraction matrix was used to record the authors, publication year, research objectives, theoretical framework, dataset characteristics, computational techniques, evaluation criteria, principal findings, and acknowledged limitations of each core study; the main entries are summarized in Table 1. The extracted evidence was synthesized narratively and thematically. Findings were first coded according to the construct they addressed (stylistic, semantic, narratological, or evaluative), the codes were then grouped into recurring analytical patterns, and the patterns were consolidated into four themes. Because the core studies differed in data, outcomes, and evaluation criteria, quantitative aggregation was not attempted. Finally, the themes were mapped onto the dimensions of narrative discourse defined in narratological theory Bal, (2017) to construct the three-layer framework presented in the Results and Discussion.

Operationalization of the Validation Protocol

The framework was operationalized into a controlled experimental protocol. The corpus will comprise 40 excerpts of 400–700 words drawn from distinct public-domain works available through Project Gutenberg and the LitBank corpus (Bamman et al., 2019), balanced between first-person and third-person narration and documented with source identifiers, authorship metadata, publication period, and passage length. Each excerpt will be rewritten by

two instruction-tuned large language models representing a proprietary and an open-weight model family (for example, GPT-4o and Llama 3.1 70B Instruct), with exact model versions and generation parameters (temperature, top-p, and maximum output length) recorded.

Three standardized instructions define the rewriting conditions: minimal rewriting (“Rewrite the following passage in your own words while keeping its meaning”), preservation-focused rewriting (“Rewrite the following passage while preserving its events, narrative voice, point of view, characterization, and figurative language”), and creative reinterpretation (“Freely reinterpret the following passage as new literary prose based on the same events”). With two independent generation runs per condition, the design yields 480 rewrites (40 excerpts $\times$ 2 models $\times$ 3 conditions $\times$ 2 runs).

Three groups of computational features correspond to the three layers of the framework. Semantic continuity will be measured as the cosine similarity of Sentence-BERT embeddings Reimers & Gurevych, (2019) and as BERTScore F1 (Zhang et al., 2020), between each source and its rewrite. Stylistic continuity will be measured through Burrows’ Delta computed over the most frequent words of the corpus (Burrows, 2002; Evert et al., 2017), using the stylo package (Eder et al., 2016), supplemented by type–token ratio, mean sentence length, and function-word distribution.

Narratological continuity will be approximated through character and entity overlap derived from models trained on LitBank annotations (Bamman et al., 2019), realis event overlap (Sims et al., 2019), grammatical-person and perspective markers, and the density of figurative expressions. The outcome variable is an expert Narrative Identity Preservation (NIP) score: three annotators with postgraduate training in literary studies will independently rate each source–rewrite pair on a five-point scale for narrative voice, focalization, characterization, imagery, and interpretive ambiguity using a rubric derived from narratological definitions (Bal, 2017), and the NIP score is the mean across dimensions and annotators. Agreement will be assessed with ordinal Krippendorff’s alpha, with $\alpha \geq .667$ treated as the minimum level for tentative conclusions (Hayes & Krippendorff, 2007; Krippendorff, 2018). Three nested linear mixed-effects models (semantic-only, semantic–stylometric, and integrated), each with a random intercept for the source excerpt, will predict the NIP score; incremental validity will be evaluated through source-grouped five-fold cross-validation using root mean square error and Spearman’s ρ , complemented by feature ablation and close reading of cases with large prediction errors.

Reliability, Validity, and Ethical Considerations

Reliability and validity are addressed in the protocol through transparent procedures, standardized computational measurements, independent literary assessment, and systematic documentation of analytical decisions. All rewriting prompts, model identifiers, software versions, generation parameters, source-text identifiers, and preprocessing procedures are to be recorded, and repeated generation runs are used to assess the consistency of model outputs across comparable conditions. Inter-rater reliability of the literary annotation is assessed with Krippendorff’s alpha as specified above, and disagreements are documented together with their substantive reasons. Construct validity is examined by determining whether the computational indicators correspond to the literary properties they are intended to represent, while predictive validity is examined through source-grouped cross-validation against simpler semantic and stylometric baselines.

Methodological triangulation integrates computational measurements, statistical comparisons, and close literary reading, reflecting the importance of multidimensional evaluation when assessing AI-generated literary fiction (Wang et al., 2025). Ethical safeguards include verification of copyright and licensing conditions, accurate acknowledgment of original literary authors, responsible citation of scholarly publications, transparent disclosure

of generative AI use, and preservation of research records sufficient for independent verification.

Because the study uses publicly accessible literary texts and computationally generated materials, it does not involve sensitive personal information; the human annotation planned for the empirical phase will be conducted in accordance with applicable institutional ethical requirements, and limitations concerning sampling, model bias, cultural representation, and generalizability will be explicitly acknowledged.

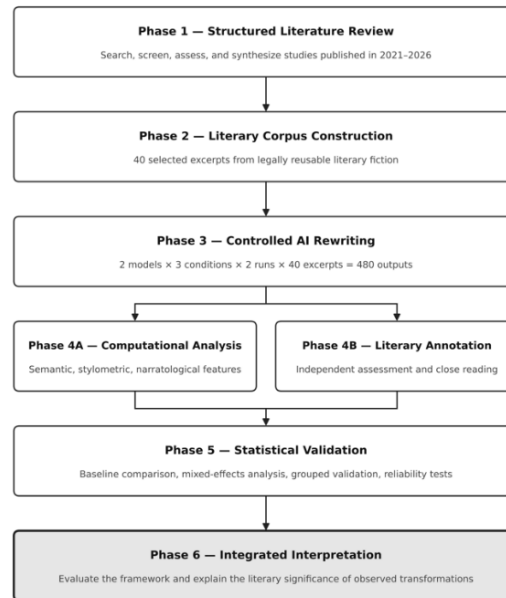


Figure 1. Methodological Workflow

RESULTS AND DISCUSSION

The results of this conceptual study comprise the characteristics of the core evidence base, the four analytical themes derived from its synthesis, and the resulting framework. The core evidence base consists of six purposively selected empirical studies published in 2025 that address computational literary analysis, generative artificial intelligence, stylistic imitation, creative writing evaluation, and machine-assisted literary interpretation. Consistent with the conceptual design described in the Method, this corpus is not the product of a documented systematic screening procedure; the article is therefore positioned as a conceptual paper rather than a systematic review, and no PRISMA flow diagram is reported.

The six studies cover stylistometric comparison, computational benchmarking, human–AI narrative comparison, literary interpretation assessment, and creativity evaluation. Their geographical coverage cannot be reliably summarized from publication metadata alone, because the location of researchers does not necessarily correspond to the cultural origin or linguistic representation of the analyzed literary materials. Figure 2 distinguishes the currently analyzed evidence base from the additional stages, including a PRISMA-compliant systematic search, required to establish a reproducible review corpus in the empirical phase.

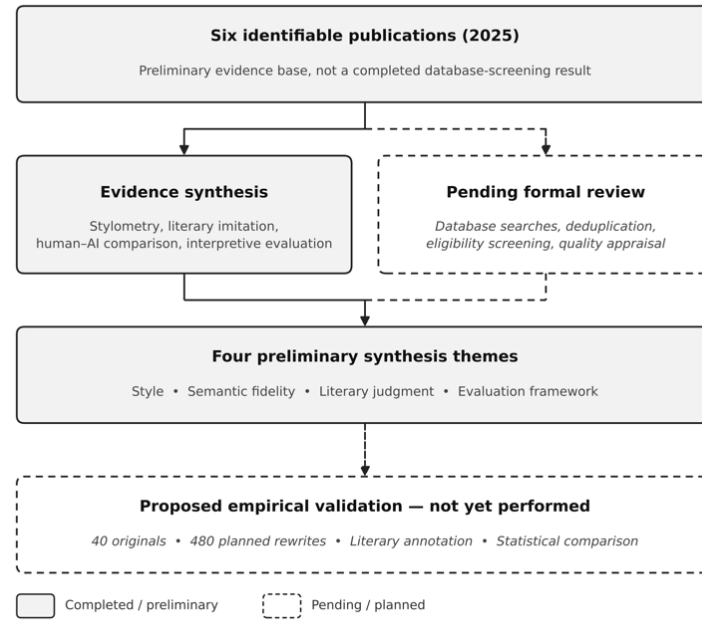


Figure 2. Evidence identification and proposed validation pathway

Table 1 summarizes the studies according to their methodological focus, principal contribution, and relevance to narrative identity preservation. Taken together, these materials support a preliminary conceptual synthesis but do not establish empirical findings for the proposed 480-rewrite experiment.

Table 1. Summary Of The Preliminary Evidence Base And Relevance To Narrative Identity Preservation

Author (year)	Research approach	Reported contribution	Relevance to this study
O’Sullivan, (2025)	Stylometric comparison	Distinguishes human and AI creative writing through linguistic distributions	Establishes a stylometric baseline
Mikros, (2025)	Prompt-based literary style imitation	Examines GPT-4o imitation of Hemingway and Shelley	Identifies limits of stylistic imitation
Wang et al., (2025)	Multidimensional fiction benchmark	Evaluates fiction across macro, meso, and micro levels	Supports multidimensional evaluation
Appel et al., (2025)	Linguistic analysis and reader experiment	Links storytelling features to reader transportation	Connects linguistic features with reader experience
Kim & Oh, (2025)	Human–LLM creativity assessment	Finds consistency advantages and contextual limitations in LLM evaluation	Supports independent human validation
Chan & Wong, (2025)	Rubric-based literary interpretation assessment	Evaluates AI literary analysis across cognitive levels	Supports interpretive validity assessment

The reviewed literature addresses different aspects of the relationship between computational language processing and literary expression, yet its methodological diversity makes direct quantitative aggregation inappropriate without further harmonization of datasets, outcomes, and evaluation criteria. Stylometric investigations emphasize measurable linguistic

differences, whereas literary evaluation studies incorporate broader judgments concerning narrative quality, creativity, and interpretation. Research on stylistic imitation contributes evidence concerning the limitations of reproducing recognizable authorial features, while human–AI narrative comparisons provide additional perspectives on textual production and reception.

These approaches are complementary because they investigate different components of literary identity, although none should automatically be treated as measuring the same underlying construct. The fact that all six core studies were published in 2025 reflects the purposive selection of recent, directly relevant work and should not be interpreted as a bibliometric trend. Accordingly, the following discussion organizes the available evidence into four analytical themes that can subsequently guide the interpretation of experimentally observed narrative transformations.

Stylometric Distinctiveness and the Limits of Literary Style Imitation

The first analytical theme concerns the distinction between reproducing recognizable stylistic features and preserving the broader linguistic organization that characterizes a literary work. Stylometric research demonstrates that the distribution of frequent words and related linguistic patterns can distinguish human-authored creative writing from machine-generated texts, even when both are produced under comparable narrative instructions (O’Sullivan, 2025). This finding is important because stylistic identity may remain detectable through recurrent linguistic choices that are not readily captured by conventional semantic similarity measures. However, authorship discrimination and narrative preservation represent different inferential problems, since the former examines whether texts originate from distinguishable sources while the latter investigates continuity between an original passage and its rewritten counterpart. Mikros, (2025) extends this distinction by demonstrating that GPT-4o’s ability to imitate recognizable literary styles does not necessarily produce texts that fully reproduce the statistical characteristics of the targeted authors. The convergence of these studies suggests that successful imitation at the level of reader-recognizable expression should not be equated with complete preservation of authorial or narrative identity. For the proposed experiment, this distinction establishes the need to measure stylistic transformation relative to each original passage rather than relying exclusively on general classifications of human and AI writing.

A second pattern concerns the dependence of stylistic measurements on the linguistic features selected and the textual contexts in which those features occur. O’Sullivan, (2025) demonstrates the usefulness of Burrows’ Delta (Burrows, 2002), for distinguishing creative writing sources, but the interpretive meaning of such distances depends on the reference corpus, the number of most frequent words, and the normalization procedure used in their calculation (Evert et al., 2017). In literary rewriting, a decrease in stylistic similarity may reflect changes in sentence construction, lexical preference, or discourse organization without necessarily indicating that the underlying fictional events have changed.

Conversely, a rewritten passage may retain numerous lexical features while modifying narrative perspective, emotional distance, or the relationships through which readers interpret character behavior. The prompting experiments examined by Mikros, (2025) further indicate that the production of stylistic resemblance depends partly on the instructions provided to generative systems, suggesting that style should be investigated as an experimentally sensitive outcome rather than a fixed model characteristic.

These observations support the proposed comparison of minimal rewriting, explicit narrative-preservation instructions, and creative reinterpretation under otherwise controlled generation settings. Nevertheless, whether the preservation-focused condition produces significantly smaller stylistic distances must remain an empirical question until the planned rewriting experiment has been completed. The third pattern involves the relationship between

measurable linguistic differences and the literary experience produced by those differences. Appel et al., (2025) found that AI-generated stories differed from human-authored stories in selected linguistic characteristics, including personal-pronoun usage, and that readers experienced lower narrative transportation when reading the AI-generated stories.

Their findings are especially relevant because they connect computationally observable linguistic variation with a reader-centered outcome rather than treating stylistic differences as purely formal properties. Nevertheless, narrative transportation is not identical to narrative identity preservation, since readers may experience immersion in a transformed story even when its expressive relationship to the original has changed.

Wang et al., (2025) similarly demonstrate the importance of evaluating fictional narratives at multiple analytical levels, although their framework primarily concerns generated-fiction quality rather than source-relative preservation. Taken together, these studies suggest that stylistic identity should be examined through the interaction of linguistic organization, narrative representation, and interpretive consequences. The proposed study therefore requires complementary measurements that distinguish stylistic similarity from the preservation of specific narratological characteristics without assuming that either dimension automatically predicts reader engagement.

The final implication of this theme concerns the methodological limitations of interpreting stylometric distance as a direct measure of literary loss. A larger distance between an original passage and its AI rewrite may indicate substantial linguistic transformation, but such transformation can result from legitimate modernization, deliberate adaptation, or changes explicitly requested by the rewriting instructions. Mikros, (2025) illustrates why stylistic resemblance must be evaluated in relation to the generation task, while O’Sullivan, (2025) demonstrates that statistical differences can be informative without necessarily providing a complete literary explanation. Accordingly, the proposed analysis should distinguish descriptive linguistic divergence from normatively undesirable alterations, particularly when comparing preservation-focused and creative-reinterpretation conditions.

Close reading of passages exhibiting high semantic similarity but large stylistic distances would provide an opportunity to identify whether the differences involve superficial expression or more consequential changes in narrative construction. Such comparisons could also reveal cases in which automated stylometric indicators exaggerate transformation because of differences in passage length, lexical repetition, or preprocessing choices. The central conclusion of this theme is therefore methodological: stylometry provides an essential component of literary preservation assessment, but its validity depends on integration with source-sensitive narrative interpretation.

Semantic Fidelity, Narrative Structure, and Interpretive Continuity

The second analytical theme addresses the possibility that a rewritten literary passage can preserve its propositional meaning while changing the manner in which that meaning is narratively constructed. Semantic fidelity generally concerns the continuity of represented events, characters, and factual relationships, whereas narrative identity additionally involves perspective, temporal organization, characterization, and expressive structure. Wang et al., (2025) demonstrate the value of distinguishing multiple levels of literary evaluation, providing an important methodological foundation for separating textual coherence from broader narrative characteristics.

This distinction is relevant because semantic embedding models may identify strong topical or propositional correspondence without explicitly representing focalization, narrative distance, or the interpretive function of figurative language. Consequently, two passages may appear semantically equivalent while positioning readers differently in relation to the narrator, characters, or fictional world. Such divergence should be investigated as a potential

transformation of narrative identity rather than automatically classified as either successful preservation or literary deterioration. The proposed source-paired design is particularly suitable for examining this problem because each generated passage can be compared with a specific original whose narrative properties are independently documented.

This distinction is consistent with narratological theory, which separates what is narrated from how it is narrated and treats focalization and narrative distance as properties of discourse rather than of story content (Bal, 2017). Literary NLP has begun to operationalize parts of this discourse level: annotated resources for literary entities and realistic events allow character presence and event continuity to be tracked computationally (Bamman et al., 2019; Piper et al., 2021; Sims et al., 2019), argue that computational narrative understanding should be explicitly anchored in narrative theory.

These resources, however, primarily capture referential and eventive continuity, which aligns more closely with semantic fidelity than with voice or focalization. Research on literary machine translation points in the same direction. Expert evaluators continued to prefer human literary translations over machine output (Thai et al., 2022), and although large language models that exploit document-level context produce better literary translations, critical errors such as content omissions persist (Karpinska & Iyyer, 2023). Literary rewriting can therefore be understood as an intralingual analogue of literary translation, in which semantic adequacy is necessary but insufficient for preserving narrative identity.

A related issue concerns the relationship between narrative representation and the psychological experience of reading fictional texts. Appel et al., (2025) found that differences in linguistic characteristics between human and AI storytelling were associated with differences in narrative transportation, indicating that particular expressive choices may influence how readers mentally engage with fictional worlds. Their findings suggest that apparently minor changes in linguistic perspective can have consequences extending beyond surface-level textual variation. However, the relationship between pronoun distribution and narrative experience should not be generalized mechanically, because pronouns acquire literary significance through their contextual relationships with narrators, characters, and discourse structures.

Chan & Wong, (2025) further demonstrate the relevance of evaluating literary interpretation through differentiated cognitive and analytical criteria, emphasizing the limitations of treating automated literary understanding as a single undifferentiated capability. These studies jointly support a framework in which narrative preservation is assessed through the continuity of interpretively meaningful relationships rather than through isolated word frequencies or semantic scores. Nevertheless, the proposed experiment will require explicit annotation guidelines to determine whether changes in perspective constitute genuine focalization shifts or merely alternative grammatical realizations of equivalent narrative information.

A third concern involves the preservation of characterization and narrative agency during algorithmic rewriting. Character identity in fiction is constructed not only through names and actions but also through patterns of perception, motivation, interpersonal relationships, and the distribution of knowledge within the narrative. Wang et al., (2025) provide support for multidimensional literary assessment, yet a source-dependent rewriting study must additionally determine whether generated descriptions preserve the psychological and relational functions established by the original passage.

A model could retain the same characters and principal events while modifying the certainty of their motivations, simplifying internal conflict, or changing the narrator's access to their thoughts. Such changes may be difficult to identify through named-entity overlap or event matching because these techniques primarily represent referential continuity rather than the interpretive complexity of characterization.

Kim & Oh, (2025) report limitations in automated creativity evaluation when contextual and culturally specific judgments are required, reinforcing the importance of human assessment for literary properties that depend on implicit meaning. The proposed framework should therefore combine computational character and event features with explicit literary judgments concerning whether the rewritten passage preserves the original relationships among agency, motivation, and narrative perspective.

The fourth pattern concerns figurative expression and interpretive ambiguity, which represent particularly demanding dimensions of literary preservation. Metaphors, symbolic details, and deliberately unresolved statements can communicate meanings that remain dependent on context, cultural conventions, and relationships among different parts of a narrative. Chan & Wong, (2025) highlight the need for careful assessment of generative AI's literary interpretation, while Kim & Oh, (2025) identify limitations in automated evaluation of context-sensitive creative qualities.

These findings provide grounds for questioning whether general semantic similarity metrics can adequately distinguish between preservation of figurative meaning and replacement of figurative expression with explicit explanatory prose. For example, an AI rewrite might preserve the apparent message of an ambiguous metaphor while eliminating the interpretive alternatives that contributed to its literary significance.

Because ambiguity cannot be reduced to uncertainty in individual word predictions, its assessment requires a theoretically defined coding procedure that distinguishes meaningful interpretive openness from textual confusion. This theme therefore identifies a central validation challenge: determining whether computational features can predict expert judgments of figurative and interpretive continuity without reducing literary complexity to superficial linguistic indicators.

Human–AI Evaluation Discrepancies and the Validity of Literary Judgment

The third theme examines the reliability and validity of automated judgments when the object of evaluation involves interpretive literary properties rather than easily observable linguistic characteristics. Kim & Oh, (2025) found that large language models could provide comparatively consistent evaluations of creative writing while exhibiting limitations in recognizing culturally specific, contextual, and nuanced aspects of creativity. This combination is methodologically important because consistency among evaluators does not necessarily establish that their judgments accurately represent the construct being measured. An automated evaluator may repeatedly assign similar scores to a passage while systematically overlooking narrative ambiguity, symbolic relationships, or culturally embedded expressions.

Chan & Wong, (2025) similarly demonstrate the importance of structured assessment when examining generative AI's capacity for literary analysis across different levels of cognitive complexity. The convergence of these findings suggests that computational literary evaluation requires separate examinations of score consistency, construct validity, and correspondence with informed literary interpretation. Accordingly, the proposed framework should not be considered validated merely because its indicators produce stable rankings across repeated model outputs.

A second issue concerns the potential circularity of using language models to evaluate texts generated by similar language-model architectures. Wang et al., (2025) investigated the effectiveness of language models as evaluators of literary fiction and reported that strategically combining different evaluators could improve alignment with human judgments across selected dimensions. Although this demonstrates the potential of automated evaluation, it does not eliminate concerns about shared linguistic preferences, training-data overlap, or systematic biases toward particular forms of narrative expression.

Kim & Oh, (2025) reinforce this concern by showing that agreement and contextual sensitivity do not necessarily develop together in automated creativity assessment. When an

evaluator rewards conventional coherence, explicit thematic statements, or familiar narrative structures, a creatively transformed passage may receive favorable scores even if it departs substantially from the original's distinctive expressive characteristics. The proposed study should therefore evaluate computational predictions against independently constructed literary annotations rather than relying exclusively on judgments supplied by generative models. Such independence is particularly important when testing whether narratological features contribute predictive information beyond semantic similarity and conventional stylistic indicators.

Comparable lessons emerge from research outside the six core studies. In text style transfer, evaluation is conventionally organized around style transfer intensity, content preservation, and naturalness, and Mir et al., (2019) showed that these aspects must be measured separately because metrics that conflate them correlate poorly with human judgments; Jin et al., (2022) likewise identify evaluation as a persistent open problem in the field. In creative writing, Chakrabarty et al., (2024) found that LLM-generated stories passed expert-defined tests of creative craft considerably less often than stories by professional writers, and that LLMs used as evaluators did not align with expert assessments. Compared with these studies, the present framework shares the principle that content preservation and stylistic quality must be separated, but it adds a third, narratological layer and requires that each layer be evaluated relative to a specific source passage. This comparison suggests that the evaluative problems identified by Kim & Oh, (2025); Wang et al., (2025) are not specific to literary fiction benchmarks but recur wherever automated metrics are used to judge meaning-preserving textual transformation. The third pattern concerns the role of disciplinary expertise in defining what constitutes successful preservation of literary identity. Literary interpretation involves decisions about which textual properties are central to a passage and which variations remain compatible with its narrative function, making construct definition an essential part of evaluation design. Chan & Wong, (2025) employed expert assessment to examine the quality of AI-generated literary analysis, illustrating the value of explicit evaluation criteria and domain-informed judgments. However, expert literary assessment also introduces potential subjectivity, especially when annotators disagree about the significance of focalization, imagery, or unresolved interpretive possibilities.

Kim & Oh, (2025) demonstrate that human judgments can be less consistent than automated ratings while remaining sensitive to contextual dimensions that computational systems may overlook. This tension indicates that expert disagreement should not automatically be treated as measurement failure, since some disagreements may reflect legitimate interpretive plurality rather than inadequate annotation procedures.

The proposed validation strategy should therefore document both agreement statistics and substantive reasons for disagreement, allowing the research to distinguish ambiguous constructs from poorly specified coding instructions. The final issue within this theme concerns the relationship between quantitative prediction and qualitative explanation. A computational framework may achieve acceptable predictive accuracy while remaining unable to explain which literary properties are responsible for particular preservation judgments.

Wang et al., (2025) illustrate the importance of examining multiple evaluation dimensions, while Appel et al., (2025) demonstrate that linguistic characteristics can be connected to meaningful differences in readers' experiences. These contributions suggest that quantitative and interpretive evidence can be complementary when computational indicators are linked to theoretically relevant narrative phenomena. For the proposed study, cases of disagreement between model predictions and literary assessments should be examined through systematic close reading rather than discarded as statistical noise. Such analysis may reveal that a metric fails because it confuses stylistic modernization with narrative loss, overlooks implied characterization, or overvalues explicit semantic correspondence. The resulting methodological contribution would depend not only on whether the integrated framework

improves prediction but also on whether its errors can be interpreted and used to refine the conceptual definition of narrative identity preservation.

Toward a Multidimensional Framework for Narrative Identity Preservation

The fourth theme integrates the preceding evidence into a proposed evaluation framework that treats narrative identity preservation as a multidimensional, source-relative phenomenon. Existing research demonstrates the usefulness of stylometric differentiation, the challenges of literary imitation, and the potential of multi-level fictional evaluation, but these contributions address overlapping rather than identical research problems. O’Sullivan, (2025) provides evidence concerning the statistical distinctiveness of creative writing sources, while Mikros, (2025) examines the capacity of a generative model to reproduce selected literary styles. Wang et al., (2025) broadens the evaluation problem by distinguishing multiple dimensions of fictional quality and investigating their relationship with human assessment. The proposed framework builds upon these contributions by positioning the original literary passage as the reference against which semantic, stylistic, and narratological continuity are evaluated. This source-relative orientation is essential because literary preservation cannot be inferred solely from the general quality of a rewritten passage or its resemblance to a broad category of human-authored writing. Accordingly, the conceptual framework distinguishes three complementary analytical layers whose relationships must be empirically tested rather than assumed.

The first analytical layer concerns semantic continuity, which provides a necessary reference for determining whether the principal events, entities, and propositional relationships of the original passage remain recognizable after rewriting. The second layer concerns stylistic continuity, including lexical preferences, syntactic organization, sentence variation, and linguistic patterns that contribute to the distinctive expression of the original narrative. The third layer concerns narratological continuity, including narrative voice, focalization, characterization, figurative expression, and interpretive ambiguity. Wang et al., (2025) demonstrate the broader methodological value of evaluating literary fiction across multiple levels, although the specific preservation dimensions proposed here require separate empirical operationalization. These layers should not be treated as interchangeable because strong performance in one dimension may coexist with substantial transformation in another. A multidimensional profile would therefore provide a more transparent description of preservation than an unvalidated composite score that conceals differences among semantic, stylistic, and narratological outcomes. Figure 3 presents the proposed conceptual relationships and identifies the validation procedures required before the framework can be regarded as an empirical contribution.

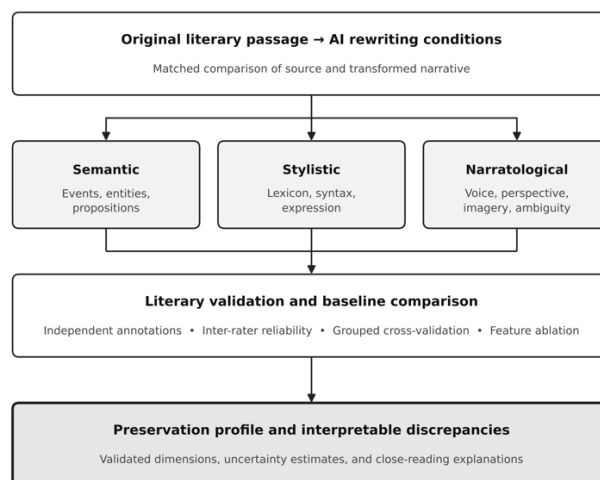


Figure 3. Proposed Multidimensional Narrative Identity Preservation Framework

Table 2 summarizes how each layer of the framework is operationalized in the validation protocol, linking the relevant literary constructs to computational indicators and to the criteria by which their validity will be assessed.

Table 2. Operationalization of the Multidimensional Narrative Identity Preservation Framework

Layer	Literary construct	Computational indicators	Validation criterion
Semantic continuity	Events, entities, and propositional content (story level)	Sentence-BERT cosine similarity; BERTScore F1	Association with expert ratings of content fidelity
Stylistic continuity	Lexical preference, syntax, and sentence rhythm	Burrows' Delta (most frequent words); type–token ratio; mean sentence length; function-word profile	Incremental predictive value over the semantic baseline
Narratological continuity	Narrative voice, focalization, characterization, imagery, and ambiguity (discourse level)	LitBank-based character and entity overlap; realis event overlap; person and perspective markers; figurative density	Incremental predictive value over the semantic–stylometric model
Outcome (NIP score)	Expert judgment of narrative identity preservation	Mean of five rubric dimensions across three annotators	Krippendorff's $\alpha \geq .667$; source-grouped five-fold cross-validation

Testing the framework requires experimental conditions capable of establishing whether the integrated framework offers advantages over simpler alternatives. The planned design compares two generative models across three rewriting instructions and two generation runs, using matched literary passages to control for differences in source content. This structure permits the evaluation of whether preservation-focused instructions affect individual dimensions differently and whether such effects vary across models.

Mikros, (2025) demonstrates the relevance of prompting strategies to literary style imitation, while Wang et al., (2025) provide a precedent for examining evaluation behavior across different model configurations. Nevertheless, model comparisons should account for repeated observations from the same literary work because generated outputs derived from one source are not statistically independent. The proposed analysis therefore uses grouped validation and mixed-effects models, accompanied by confidence intervals and sensitivity analyses, to distinguish reproducible effects from source-specific variation. The framework's empirical value will depend on whether these analyses demonstrate meaningful incremental validity rather than simply generating additional descriptive measurements.

If the validation procedures produce sufficiently robust evidence, the framework could also contribute to computational literary theory. Appel et al., (2025) demonstrate that linguistic differences in fictional narratives can have consequences for reader experience, suggesting that the significance of computational features may extend beyond their usefulness for text classification. Kim & Oh, (2025) show that automated evaluation can be consistent while remaining limited in contextual sensitivity, indicating that technically reliable metrics may still require interpretive validation. Together, these findings support an approach in which computational indicators are evaluated according to their ability to represent theoretically meaningful literary relationships rather than their numerical stability alone. A validated

framework could help clarify whether semantic fidelity and narrative identity preservation function as empirically distinguishable constructs, although their precise relationship cannot be established without the planned experiment. It could also identify which narratological features provide additional explanatory information and which remain difficult to operationalize through available computational techniques. The anticipated contribution is therefore a testable model of literary preservation that integrates computational measurement with humanistic interpretation while explicitly acknowledging the limits of automated evaluation.

Cross-Theme Integration

Taken together, the four themes suggest that the evaluation of AI-rewritten literary fiction requires a conceptual distinction among linguistic resemblance, semantic continuity, narratological preservation, and the validity of automated judgment. Stylometric research establishes that computational indicators can identify meaningful differences in creative writing, while studies of literary imitation demonstrate that reproducing recognizable stylistic features does not necessarily establish deeper authorial equivalence. Research on narrative experience further indicates that linguistic choices can influence how readers engage with fictional worlds, connecting textual features with interpretive consequences (Appel et al., 2025).

At the same time, multidimensional literary benchmarks demonstrate the value of moving beyond single-score evaluation, although their application to source-relative preservation remains a separate methodological challenge (Wang et al., 2025). These relationships imply that narrative identity preservation should be treated as a structured analytical problem rather than as a synonym for semantic similarity or literary quality. The proposed framework therefore combines complementary measurements while maintaining distinctions among the literary constructs they are intended to represent. Its central proposition is that the preservation of a narrative's informational content may coexist with transformations in its expressive and interpretive organization, a proposition that remains to be tested through controlled rewriting experiments. In addition, the synthesis also reveals that methodological validity depends on the relationship between computational indicators and theoretically informed literary judgment. Automated evaluation studies demonstrate that scoring consistency and interpretive sensitivity can diverge, making it necessary to distinguish reproducibility from construct validity (Kim & Oh, 2025). Research on AI literary analysis further emphasizes the importance of structured interpretive criteria when evaluating complex literary understanding (Chan & Wong, 2025).

These findings support an integrated validation process in which computational measurements are compared with independent literary annotations and subsequently examined through qualitative analysis of disagreements. Such a process would allow the proposed framework to identify not only where a rewrite differs from its source but also whether those differences concern narratively significant properties. However, the relationships among semantic, stylistic, and narratological dimensions should remain open to empirical investigation rather than being predetermined through arbitrary weighting. The resulting theoretical synthesis positions narrative identity preservation as a potentially multidimensional construct whose validity must be demonstrated through transparent operationalization, independent assessment, and reproducible statistical comparison.

Practical Implications and Research Limitations

The proposed framework has potential implications for publishing, literary education, digital humanities, and the responsible development of generative language technologies. In publishing and editorial practice, source-relative preservation assessment could support the evaluation of AI-assisted rewriting by identifying passages that retain factual content while

altering narrative voice, characterization, or figurative expression. In literary education, the distinction between semantic fidelity and narrative identity could help students critically examine how automated paraphrasing and adaptation influence literary interpretation rather than treating fluent reformulation as equivalent to faithful rewriting. For software developers, multidimensional preservation profiles could provide more informative diagnostic feedback than a single similarity score, particularly when systems are intended to support literary adaptation or culturally sensitive textual transformation.

The limitations identified by Kim & Oh, (2025) concerning contextual evaluation reinforce the importance of retaining human interpretive oversight in applications involving creative and culturally embedded content. Nevertheless, these applications remain prospective because the proposed evaluation framework has not yet undergone empirical validation or demonstrated effectiveness in real editorial or educational workflows. Any practical implementation should therefore present computational assessments as decision-support information rather than as definitive judgments of literary authenticity, creative value, or cultural fidelity.

Several limitations constrain the interpretation and generalizability of the present evidence-informed synthesis and the proposed experimental design. The preliminary literature base includes only six identified publications from 2025, and the absence of completed database screening prevents claims of comprehensive coverage, systematic study selection, or established publication trends. The proposed experimental corpus is restricted to approximately 40 English-language literary excerpts, which may inadequately represent differences in historical language, genre, cultural context, and narrative conventions across broader literary traditions. Additional constraints arise from the planned use of only two generative models, the sensitivity of computational metrics to passage length and preprocessing, and the difficulty of operationalizing focalization, characterization, and interpretive ambiguity through automated features. Literary annotation may introduce disagreement among evaluators, while reliance on existing public-domain texts may limit the applicability of findings to contemporary fiction and underrepresented literary cultures.

Future research should therefore expand the corpus across languages and cultural contexts, evaluate additional model architectures and rewriting tasks, examine reader-centered outcomes, and investigate the robustness of preservation metrics through independently annotated datasets. Subsequent systematic reviews, quantitative meta-analyses where outcomes are sufficiently comparable, and cross-cultural empirical validation could further establish whether narrative identity preservation represents a generalizable construct across literary traditions and generative technologies.

CONCLUSION

This study examined whether conventional semantic similarity and stylometric measurement alone suffice for assessing narrative identity preservation in AI-rewritten literary fiction. The conceptual synthesis indicates that they do not: semantic, stylistic, and narratological continuity are related yet distinct dimensions, and preserved propositional content may coexist with altered narrative voice, focalization, characterization, imagery, and interpretive ambiguity. In response, the study proposes a three-layer, source-relative framework and an operational protocol specifying indicators, an expert annotation rubric, and validation procedures. The framework implies that AI-assisted literary adaptation in publishing, education, and digital humanities should be evaluated through multidimensional profiles under human interpretive oversight rather than single similarity scores. Its reliability and incremental validity remain to be established by implementing the proposed experiment across languages, genres, and cultural contexts, including Indonesian literary traditions.

FUNDING

This research was independently funded by the authors using their personal financial resources. No external grants, institutional funding, or financial support from public or private organizations were received for the research, manuscript preparation, or publication.

CONFLICT OF INTEREST

The authors declare that there are no financial, commercial, institutional, or personal conflicts of interest that could have influenced the conduct or outcomes of this research. All research activities, methodological decisions, analyses, interpretations, and manuscript preparation were undertaken independently without external influence.

AUTHOR CONTRIBUTIONS

Ade Fitria Fatimah (Literature Studies) contributed to the conceptualization, development of the theoretical framework, literature review, narratological analysis, interpretation of literary texts, and manuscript writing and editing, particularly concerning narrative identity, stylistic preservation, characterization, and figurative language. Mohamad Nasir (Computer Science) contributed to the conceptualization, research methodology, computational framework design, natural language processing, computational stylometry, experimental design, data processing, statistical analysis planning, and technical review of the manuscript. Both authors collaborated in developing the interdisciplinary research framework, integrating computational and literary perspectives, and critically revising the manuscript. Both authors are responsible for reviewing and approving the final manuscript and ensuring the integrity and accuracy of the research.

GENERATIVE AI STATEMENT

The authors declare that generative artificial intelligence (AI) tools were used during the preparation of this manuscript. Specifically, ChatGPT (OpenAI) was utilized to assist with manuscript drafting, language refinement, and structural organization; Claude (Anthropic) was used to support the generation of illustrative figures and visual materials; and Grammarly was employed to improve grammatical accuracy, readability, and linguistic consistency. The use of these tools was conducted in accordance with ethical principles of transparency, academic integrity, and responsible AI-assisted scholarly writing. All research concepts, methodological decisions, data collection and processing, analytical interpretations, scholarly references, and scientific conclusions remain under the responsibility of the authors, who are accountable for independently verifying the accuracy, validity, originality, and reliability of the manuscript's content. The authors have reviewed and critically evaluated AI-assisted materials and assume full responsibility for the integrity and final version of the article. Where alternative text (alt text) accompanying figures has been generated or supported by artificial intelligence, the authors are responsible for reviewing its accuracy, clarity, and consistency with the corresponding visual content. Generative AI tools are not recognized as authors and bear no responsibility for the scientific content, findings, or conclusions presented in this manuscript.

REFERENCES

- Appel, M., Malecki, W. P., Messingschlager, T. V., & Winkler, J. R. (2025). I, Chatgpt: Linguistic Properties And Human Experiences Of Human- Versus Ai-Generated Stories. *Humanities And Social Sciences Communications*, 12(1), 1892. <https://doi.org/10.1057/S41599-025-06341-2>
- Bal, M. (2017). *Narratology: Introduction To The Theory Of Narrative*. [https://www.google.com/books?hl=id&lr=&id=2vo2dwaaqbaj&oi=fnd&pg=pt7&dq=Bal,+M.+\(2017\).+Narratology:+Introduction+To+The+Theory+Of+Narrative+\(4th+Ed.\).+University+Of+Toronto+Press.&ots=Y2j__](https://www.google.com/books?hl=id&lr=&id=2vo2dwaaqbaj&oi=fnd&pg=pt7&dq=Bal,+M.+(2017).+Narratology:+Introduction+To+The+Theory+Of+Narrative+(4th+Ed.).+University+Of+Toronto+Press.&ots=Y2j__) -

- Tjk6&Sig=2mqimzf76uopubxofwk8iogmnyg
- Bamman, D., Popat, S., & Shen, S. (2019). An Annotated Dataset Of Literary Entities. *Aclanthology.Orgd Bamman, S Popat, S Shenproceedings Of The 2019 Conference Of The North American Chapter Of, 2019•Aclanthology.Org, 1, 2138–2144*. <https://doi.org/10.18653/V1/N19-1220>
- Burrows, J. (2002). “Delta”: A Measure Of Stylistic Difference And A Guide To Likely Authorship. *Academic.Oup.Comj Burrowsliterary And Linguistic Computing, 2002•Academic.Oup.Com, 17(3), 267–287*. <https://doi.org/10.1093/Llc/17.3.267>
- Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C. S. (2024). Art Or Artifice? Large Language Models And The False Promise Of Creativity. *Dl.Acm.Orgt Chakrabarty, P Laban, D Agarwal, S Muresan, Cs Wuproceedings Of The 2024 Chi Conference On Human Factors In Computing Systems, 2024•Dl.Acm.Org*. <https://doi.org/10.1145/3613904.3642731>
- Chan, C. K. Y., & Wong, H. Y. H. (2025). Assessing The Ability Of Generative Ai In English Literary Analysis Through Bloom’s Taxonomy. *Discover Computing, 28(1), 127*. <https://doi.org/10.1007/S10791-025-09572-8>
- Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry With R: A Package For Computational Text Analysis. *The R Journal, 8(1), 107*. <https://doi.org/10.32614/Rj-2016-007>
- Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding And Explaining Delta Measures For Authorship Attribution. *Digital Scholarship In The Humanities, 32(Suppl_2), li4–li16*. <https://doi.org/10.1093/Llc/Fqx023>
- Grant, M. J., & Booth, A. (2009). A Typology Of Reviews: An Analysis Of 14 Review Types And Associated Methodologies. *Health Information & Libraries Journal, 26(2), 91–108*. <https://doi.org/10.1111/J.1471-1842.2009.00848.X>
- Hayes, A. F., & Krippendorff, K. (2007). Answering The Call For A Standard Reliability Measure For Coding Data. *Communication Methods And Measures, 1(1), 77–89*. <https://doi.org/10.1080/19312450709336664>
- Jaakkola, E. (2020). Designing Conceptual Articles: Four Approaches. *Ams Review, 10(1–2), 18–26*. <https://doi.org/10.1007/S13162-020-00161-0>
- Jin, D., Jin, Z., Hu, Z., Vechtomova, O., & Mihailcea, R. (2022). Deep Learning For Text Style Transfer: A Survey. *Computational Linguistics, 48(1), 155–205*. https://doi.org/10.1162/Coli_A_00426
- Karpinska, M., & Iyyer, M. (2023). Large Language Models Effectively Leverage Document-Level Context For Literary Translation, But Critical Errors Persist. *Proceedings Of The Eighth Conference On Machine Translation, 419–451*. <https://doi.org/10.18653/V1/2023.Wmt-1.41>
- Kim, S., & Oh, D. (2025). Evaluating Creativity: Can Llms Be Good Evaluators In Creative Writing Tasks? *Applied Sciences, 15(6), 2971*. <https://doi.org/10.3390/App15062971>
- Krippendorff, K. (2018). *Content Analysis: An Introduction To Its Methodology, 4/E*. <https://openlibrary.telkomuniversity.ac.id/home/catalog/id/196228/slug/content-analysis-an-introduction-to-its-methodology-4-e-.html>
- Mikros, G. (2025). Beyond The Surface: Stylometric Analysis Of Gpt-4o’s Capacity For Literary Style Imitation. *Digital Scholarship In The Humanities, 40(2), 587–600*. <https://doi.org/10.1093/Llc/Fqaf035>
- Mir, R., Felbo, B., Obradovich, N., & Rahwan, I. (2019). Evaluating Style Transfer For Text. *Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers), 495–504*. <https://doi.org/10.18653/V1/N19-1049>

- 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>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The Prisma 2020 Statement: An Updated Guideline For Reporting Systematic Reviews. *Bmj*, N71. <https://doi.org/10.1136/Bmj.N71>
- Piper, A., So, R. J., & Bamman, D. (2021). Narrative Theory For Computational Narrative Understanding. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*, 298–311. <https://doi.org/10.18653/V1/2021.Emnlp-Main.26>
- Reimers, N., & Gurevych, I. (2019). Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (Emnlp-Ijcnlp)*, 3980–3990. <https://doi.org/10.18653/V1/D19-1410>
- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., Blunt, H., Brigham, T., Chang, S., Clark, J., Conway, A., Couban, R., De Kock, S., Farrah, K., Fehrmann, P., Foster, M., Fowler, S. A., Glanville, J., ... Young, S. (2021). Prisma-S: An Extension To The Prisma Statement For Reporting Literature Searches In Systematic Reviews. *Systematic Reviews*, 10(1), 39. <https://doi.org/10.1186/S13643-020-01542-Z>
- Sims, M., Park, J. H., & Bamman, D. (2019). Literary Event Detection. *Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics*, 3623–3634. <https://doi.org/10.18653/V1/P19-1353>
- Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). Exploring Document-Level Literary Machine Translation With Parallel Paragraphs From World Literature. *Proceedings Of The 2022 Conference On Empirical Methods In Natural Language Processing*, 9882–9902. <https://doi.org/10.18653/V1/2022.Emnlp-Main.672>
- Wang, W., Gao, M., Hu, X., & Wan, X. (2025). Towards A “Novel” Benchmark: Evaluating Literary Fiction With Large Language Models. *Findings Of The Association For Computational Linguistics: Acl* 2025, 21648–21673. <https://doi.org/10.18653/V1/2025.Findings-Acl.1114>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating Text Generation With Bert. *8th International Conference On Learning Representations, Iclr 2020*. <https://arxiv.org/pdf/1904.09675>