MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Data Science & Artificial Intelligence
ISSN: 2996-671X

Evaluating AI-Generated Arabic Fiction: A Reproducible Matrix for Narrative Coherence, Character Continuity, and Ethical Narrative Brakes in The Temple

Noshokaty IE*, Hamid SR[2], Halim HM[3], Mohamed KO[4] and Omar MM[5]

[1]Professor and Head of the Department of Acoustics and Artificial Intelligence, Enosh Science Center, Egypt

[2]Assistant Professor and Quality Assurance Officer, Department of Quality Assurance and Academic Compliance, Enosh Science Center, Egypt

[3]Doctoral Researcher and Libraries & Research Support Officer, Enosh Science Center, Egypt

[4]MSc Student and Artificial Intelligence Research Assistant, Department of Artificial Intelligence, Enosh Science Center, Egypt

[5]Assistant Professor, Department of Psychiatry, Medical Studies Department for Children, Ain Shams University, Egypt

***Corresponding author:** Ibrahim El Noshokaty, Professor and Head of the Department of Acoustics and Artificial Intelligence, Enosh Science Center, Egypt, Tel: +201148444177; Email: ibrahim@enoshmink.com

Research Article

Volume 4 Issue 1

Received Date: June 30, 2026

Published Date: August 25, 2026

DOI: 10.23880/oajda-16000168

Abstract

This article proposes a reproducible measurement matrix for evaluating AI-generated fiction, using the Arabic novel The Temple as an applied case study. The study addresses a central methodological challenge in the assessment of generative literary outputs: how can narrative coherence, character continuity, and ethical narrative development be evaluated without relying solely on subjective literary impressions? To answer this question, the paper adopts a mixed-method design combining structural-semiotic analysis with supporting quantitative indicators. Qualitatively, the analysis examines coherence anchors, character value arcs, point-of-view stability, recurring symbolic motifs, and the narrative function of documents and system logs within the fictional world. Quantitatively, the study applies page-level metrics, including entity mentions, page coverage, longest consecutive presence runs, and TF-IDF cosine similarity between adjacent pages as a proxy for local textual cohesion. The findings suggest that The Temple maintains a recognizable network of symbolic and causal anchors, particularly the 110 Hz motif, the stone, the temple, Geneva, and the investigation frame. Entity tracking also indicates sustained continuity for major characters and system-related entities across the narrative. The paper further introduces the concept of the Ethical Narrative Brake (ENB), defined as a set of fictional-world constraints that limits irresponsible disclosure, reduces binary moral judgment, and redirects narrative action toward accountable preservation of knowledge. The proposed matrix is not presented as a final aesthetic judgment, but as a repeatable evaluative framework that supports human critical reading through transparent computational proxies.

Keywords: Artificial Intelligence; AI-Generated Fiction; Arabic Fiction; Computational Text Analysis; Narrative Coherence; Character Continuity; Ethical Narrative Brake

Introduction

Generative artificial intelligence has changed the conditions under which literary texts are produced, revised, and evaluated. Narrative writing is no longer limited to the model of an individual human author as the sole source of sentence formation, event progression, and character development. Since the introduction of the Transformer architecture [1], large language models have developed the capacity to generate extended textual sequences, maintain local semantic patterns, and simulate dialogue, style, and thematic recurrence across long spans of text [2-4]. This shift raises a practical methodological question for literary studies, digital humanities, and artificial intelligence evaluation: how can the quality of AI-generated fiction be assessed in a way that is both critically meaningful and operationally repeatable? Recent work on large language model evaluation has emphasized that generative systems should not be assessed only through surface fluency or isolated task performance. Reliable evaluation requires transparent procedures, reproducible settings, and attention to limitations such as inconsistency, hallucination, bias, and weak grounding [5-7]. These concerns become more complex in literary generation, where the value of a text cannot be reduced to factual correctness or grammatical accuracy. A novel must also sustain narrative coherence, preserve character continuity, manage point of view, and develop ethically meaningful conflicts across scenes. AI-generated fiction therefore presents at least three major evaluation challenges. The first concerns narrative coherence: whether the text maintains causal, symbolic, and thematic anchors that allow the reader to follow the fictional world without unexplained fragmentation. This issue can be linked to discourse coherence theories that examine how textual units are connected through relations such as cause, elaboration, contrast, and evidence [8]. It also relates to computational approaches to coherence, where lexical overlap, semantic similarity, and entity transitions are used as operational proxies for textual continuity [9-11].

The second challenge concerns character continuity: whether characters preserve stable roles, motivations, relationships, and value trajectories across scenes. In long-form generative writing, a character may appear consistent at the sentence level while gradually shifting in motivation, voice, or moral position without sufficient narrative preparation. For this reason, character evaluation requires both qualitative coding of narrative function and quantitative tracking of entity presence across the text. The third challenge concerns ethical narrative design. Generative systems may produce language that appears morally reflective while still allowing sudden binary judgments, irresponsible disclosure, or unmotivated shifts in blame and responsibility. This study therefore treats ethics not only as a theme in the story, but also as a structural feature of the fictional world. In this context, the concept of the Ethical Narrative Brake (ENB) is introduced to describe fictional-world constraints that limit harmful disclosure, reduce simplistic moral binaries, and redirect narrative action toward accountable preservation of knowledge. Within this framework, The Temple provides a suitable applied case study. The novel was generated within a pre-designed Arabic fictional world that includes recurring symbolic anchors, ethical boundaries, investigation scenes, and system-level references. The present article does not treat the novel as evidence that AI writing is inherently superior or inferior to human writing. Instead, it uses the text as a test case for developing a reproducible measurement matrix that can be applied to other AI-generated or AI-assisted novels.

The study is guided by three research questions:

- How is narrative coherence produced in an AI-generated Arabic novel when direct human intervention in sentence formulation and event direction is minimized?
- Do the major characters and system entities develop cumulatively across the narrative, or do they shift through unexplained jumps?
- How can qualitative critical reading be combined with page-level quantitative indicators to create a repeatable evaluation matrix for AI-generated fiction?

Materials and Methods

Study Design

This study adopts a mixed-method design that combines qualitative structural-semiotic analysis with supporting quantitative indicators. The qualitative component examines narrative anchors, character value arcs, scene transitions, point-of-view stability, and the ethical function of system-level constraints within the fictional world. The quantitative component provides operational proxies for reference continuity and local textual cohesion. The two analytical tracks are intentionally complementary: quantitative indicators are used to support and document interpretive reading, but they are not treated as substitutes for critical judgment.

Study Material

The study material is the complete Arabic novel The Temple, consisting of approximately 173 pages. The novel was selected as an applied case study because it was generated within a pre-designed Arabic fictional world that includes recurring symbolic anchors, investigation scenes, ethical boundaries, and system-level references. Several internal reference points were used as thresholds for

Copyright© Noshokaty IE, et al.

tracking coherence, including the first appearance of Enosh, the stone inscription, the 110 Hz audio file, Geneva-related material, investigation scenes, system logs, and the late-stage ethical conflict between disclosure and preservation.

Units of Analysis

Two units of analysis were defined to improve replicability. The scene or narrative segment was used as the qualitative unit because each scene contains an event structure involving dialogue, decision, response, and outcome. The page was used as the quantitative unit because it provides a stable operational unit for counting entity presence, measuring continuity, and computing local similarity between adjacent textual segments. This dual-unit design allows the study to connect interpretive scene-level analysis with repeatable page-level measurements.

Entity Definition and Disambiguation

To avoid entity confusion, the study distinguishes explicitly between Enosh and ENOSH. Enosh is treated as a fictional character inside the narrative world, whereas ENOSH is treated as a technical system or generative platform. The two labels were not merged during normalization or counting. Value-arc coding was applied to fictional characters such as Enosh, Maha, Ibrahim, Yassin, and Elias, but not to ENOSH as a technical system. ENOSH was instead coded according to system function, including monitoring, compensation, mode switching, output control, and narrative redirection.

Qualitative Coding

The qualitative codebook contains four main fields: present entities, anchors present, pivot event, and decision/outcome. Character development was coded along three value axes: interrogation versus mercy, disclosure versus protection, and authority versus accountability. Scene transitions were labeled as cumulative when they were supported by prior narrative evidence, symbolic preparation, or causal development. They were labeled as unexplained jumps when no traceable causal, symbolic, or motivational preparation could be identified within the surrounding narrative context.

Supporting Quantitative Indicators

For each page, the analysis recorded word count, sentence count, average sentence length, entity mentions, anchor counts, and local cohesion. Entity tracking included total mentions, page coverage, and the longest consecutive presence run for each major character or system entity. Local cohesion was estimated using TF-IDF cosine similarity between page n and page n+1:

$$Sim(n, n+1) = \frac{\left(v_n \cdot v_{\{n+1\}}\right)}{\left(\left\|v_n\right\|\left\|v_{\{n+1\}}\right\|\right)}$$

where v_n and $v_{\{n+1\}}$ are TF-IDF vectors representing two adjacent pages. The mean local cohesion score was calculated by averaging similarity values across all adjacent page pairs. This metric is used as a lexical and topical proxy for local cohesion; it is not treated as a full semantic, literary, or aesthetic judgment. Future studies may replace or supplement this measure with deeper semantic similarity methods such as BERTScore [11] or contextual embedding-based similarity.

Ethical Narrative Brake and Fuzzy-Logic Parameterization

The Ethical Narrative Brake (ENB) was defined as a set of constraints and values embedded inside the fictional world. These constraints include access restrictions, the right to silence, evidence preservation, and mechanisms that prevent public shaming or irresponsible disclosure. ENB was coded when the narrative redirected a potentially harmful disclosure into a controlled, accountable, or preservation-oriented solution. To avoid treating ethical decisions as binary choices, the ENB layer was operationalized through a fuzzy-logic-inspired parameterization. In this approach, narrative decisions were not coded simply as "ethical" or "unethical," or as "disclosure" versus "non-disclosure." Instead, each ethically sensitive scene was evaluated through graded parameters that influence whether the narrative remains coherent or collapses into abrupt moral judgment.

The main parameters were:

- **Risk intensity:** the degree to which a disclosure could produce harm or escalation.
- **Evidence certainty:** the degree to which the available evidence is stable, complete, or narratively verified.
- **Potential harm:** the expected damage caused by uncontrolled exposure, public shaming, or premature judgment.
- **Responsibility level:** the extent to which the acting character or system entity carries responsibility for limiting harm.
- **Right-to-know:** the degree to which disclosure is justified by narrative accountability.
- **Preservation need:** the degree to which evidence should be stored, protected, or delayed rather than publicly released.
- **Narrative stability:** the degree to which the decision preserves causal continuity, character motivation, and the internal logic of the fictional world.

Copyright© Noshokaty IE, et al.

These parameters were treated as gradual values rather than absolute categories. For example, a scene with high evidence of certainty but also high potential harm would not automatically lead to public disclosure. Instead, the ENB mechanism would favor a controlled preservation-oriented solution, such as storing the evidence, limiting access, or delaying exposure until narrative responsibility becomes clearer. This fuzzy parameterization helped explain how The Temple avoids sudden binary moral shifts and maintains narrative stability during ethically intense scenes.

In this study, ENB is therefore treated as a structural feature of narrative design rather than a general moral theme. It indicates whether the generated fictional world contains internal mechanisms that reduce binary moral judgment, regulate disclosure, and redirect conflict toward responsible preservation of knowledge. The use of fuzzy-logic-inspired parameters supports the broader aim of the matrix: to evaluate whether an AI-generated narrative can sustain ethical complexity without losing coherence, collapsing character motivation, or forcing the plot into unjustified moral shortcuts.

Detection Mechanism of Ethical Narrative Brakes

The identification of Ethical Narrative Brakes (ENB) emerged inductively during the structural and scene-level analysis. The study did not begin by assuming that such brakes existed as fixed narrative devices. Instead, they were detected through repeated patterns in ethically sensitive scenes where the plot approached a possible collapse into public exposure, binary blame, revenge logic, or uncontrolled disclosure. In these scenes, the narrative repeatedly introduced an internal constraint that redirected the action away from immediate judgment and toward a more controlled form of responsibility.

The detection procedure followed four steps. First, ethically sensitive scenes were marked when the narrative involved accusation, disclosure, investigation, evidence, blame, or the possibility of public harm. Second, the scene was examined for the presence of a restraining mechanism, such as access limitation, silence, delayed release, evidence preservation, system logging, or refusal of public shaming. Third, the effect of this mechanism on the plot was evaluated: a brake was recorded only when the constraint changed the direction of the scene, reduced moral simplification, or prevented an abrupt narrative jump. Fourth, the scene was rechecked against the fuzzy ENB parameters, including risk intensity, evidence certainty, potential harm, responsibility level, right-to-know, preservation need, and narrative stability.

Accordingly, an ENB was not coded merely because the scene contained moral language. It was coded only when the narrative structure itself performed a braking function. This means that the brake had to influence the causal development of the scene, preserve character motivation, or maintain the internal logic of the fictional world. Through this procedure, the analysis showed that The Temple does not rely only on ethical statements, but repeatedly builds internal mechanisms that slow down judgment, regulate disclosure, and protect the narrative from collapsing into simplistic moral binaries.

Replicability Procedure: To support repeatability, all coding decisions were organized in a structured measurement matrix. Page-level quantitative counts were separated from scene-level qualitative interpretation, and the entity labels were fixed before analysis began. The distinction between Enosh as a fictional character and ENOSH as a technical system was preserved throughout the coding process, and the two labels were not merged during normalization, counting, or interpretation. Quantitative indicators were calculated using the same page boundaries across the full text. These indicators included entity mentions, page coverage, anchor frequency, longest consecutive presence runs, and TF-IDF cosine similarity between adjacent pages. Qualitative judgments were based on predefined coding fields, including present entities, anchors, pivot event, decision/outcome, transition type, value-arc movement, and ENB activation. The ENB layer was also made repeatable by applying the same fuzzy-logic-inspired parameters to ethically sensitive scenes. These parameters included risk intensity, evidence certainty, potential harm, responsibility level, right-to-know, preservation need, and narrative stability. A brake was coded only when the scene contained a detectable restraining mechanism that influenced narrative direction, reduced moral simplification, or preserved causal and character continuity. This procedure allows the same matrix to be reused with other AI-generated or AI-assisted novels after adjusting the entity list, anchor terms, scene boundaries, and language-specific normalization rules. The purpose is not to produce a universal aesthetic score, but to provide a transparent and repeatable evaluation protocol that combines human critical reading with documented computational proxies.

Results

Narrative Coherence Matrix

Table 1 presents the operational matrix used to evaluate narrative coherence in The Temple. The matrix focuses on recurring anchors, causal linkage, document and log integration, voice stability, thematic consistency, pacing, and

Copyright© Noshokaty IE, et al.

closure. These indicators were used to determine whether the narrative world maintains traceable continuity across scenes rather than relying on isolated or disconnected episodes.

Indicator	Operational definition	Measurement method	Scale	Evidence /example
Global anchor	A center of meaning repeatedly reactivated by the narrative	Track 110 Hz, stone, temple, and Geneva references	1 = absent; 5 = repeated	Stone inscription and 110 Hz file (pp. 6-8)
Causal linkage	Events produce later events rather than isolated episodes	Code cause-effect relations in pivotal scenes	1 = fragmented; 5 = strong	ethical layer affects binary judgment (pp. 38-42)
Logs as narration	Documents or logs drive plot decisions	Check whether logs alter decisions or outcomes	1 = decorative; 5 = plot-driving	mode change and system tags in investigation scenes
Voice / POV stability	Narrative voice is stable or transitions are clear	Track changes in point of view and speaker role	1 = confusing; 5 = clear	Temple voice differs from technical log language
Thematic consistency	Key themes remain traceable across the narrative	Count and interpret testimony, mercy, justice, and protection	1 = scattered; 5 = stable	Justice without mercy is criticized near the climax
Pacing / arc	Stakes rise toward a recognizable climax	Map beginning, escalation, turning point, and result	1 = flat; 5 = progressive	Public leakage is redirected by temple intervention
Closure	Main questions are resolved or intentionally left open	Compare opening questions with final resolution	1 = unresolved; 5 = purposeful	Evidence is preserved rather than publicly displayed

Table 1: Narrative coherence matrix.

The coherence matrix indicates that The Temple depends on an early symbolic and acoustic anchor that is repeatedly reactivated in relation to the stone, the temple, the investigation frame, and Geneva-related material. This pattern supports causal and thematic continuity rather than simple lexical repetition. The recurring anchors also help stabilize the narrative world by connecting early symbolic signals with later ethical and investigative scenes.

Character Continuity and Development

Table 2 summarizes the character-continuity matrix. The analysis does not measure continuity only by counting name recurrence. Instead, it evaluates whether each character preserves a traceable role, motive, voice, relation, and responsibility across scenes.

Indicator	Operational definition	Measurement method	Scale	Evidence / example
Entity consistency	Name, role, and relationships do not shift without explanation	Create a character card across scenes	1 = contradictory; 5 = stable	Maha remains linked to investigation pressure
Motivation	Actions are connected to clear motives	Ask why this action occurs and why at this point	1 = arbitrary; 5 = justified	Pressure in interrogation relates to later ethical conflict
Turning point	A scene changes the direction of the character	Identify before/after state	1 = absent; 5 = strong	Enosh shifts toward SELF-PROTECTION
Moral arc	Ethical movement is gradual rather than sudden	Code coercion, mercy, and responsibility	1 = static; 5 = progressive	Rejection of the interrogation cage
Voice	Each character has recognizable language or behavior	Analyze dialogue, rhythm, and functional role	1 = similar; 5 = distinct	Temple voice is dense and judgment-like
Relations	Trust, conflict, and alliances evolve	Map relations before and after the climax	1 = fixed; 5 = dynamic	Maha-Enosh relation changes with access restrictions
Agency	Final responsibility is identifiable	Analyze decisive actions and final decisions	1 = vague; 5 = clear	Withdrawal of the version becomes a final responsibility decision

Table 2: Character-continuity and development matrix.

Copyright© Noshokaty IE, et al.

The character-continuity matrix shows that the major figures in The Temple are not only repeated by name but are also organized through changing pressures, decisions, and ethical positions. Enosh, for example, is not treated as a static technological figure. Instead, the character develops through a traceable movement from interrogation pressure toward protection and refusal of the interrogation cage. ENOSH, by contrast, is treated as a system entity and is therefore evaluated according to system function rather than moral character development.

Page-Level Entity Presence

Table 3 presents updated page-level entity statistics. These values are used as continuity proxies. They do not measure the depth of characterization by themselves, but they help determine whether a character, symbolic element, or system entity remains available for narrative tracking across the text.

Entity / element	Total mentions	Pages with mention	Longest run (pages)	Run range
The Temple	160	96	7	54-60
The Stone	100	60	8	19-26
Maha	192	92	24	10-33
Ibrahim	172	102	13	155-167
Enosh (character)	96	62	13	131-143
Yassin	97	47	12	132-143
Elias	81	45	9	83-91
ENOSH (system)	93	69	12	111-122

Table 3: Page-level entity and element statistics.

The entity statistics show that major human characters, symbolic elements, and the ENOSH system retain measurable continuity across the text. Maha and Ibrahim show especially broad page coverage, while Enosh and ENOSH remain distinguishable both functionally and quantitatively. This supports the methodological decision to keep Enosh and ENOSH separate during normalization, counting, and interpretation.

Local Cohesion

The mean TF-IDF cosine similarity between adjacent pages was approximately 0.111 (SD = 0.050; range = 0.000-0.284). This result should be read as a local lexical and topical cohesion proxy rather than a full measure of literary quality. Higher local similarity was observed in sections that rely on repeated anchors such as the temple, stone, investigation frame, and system logs.

The relatively modest mean similarity is expected in long-form fiction, where scene changes, dialogue shifts, and changes in narrative focus naturally reduce lexical overlap between adjacent pages. For this reason, TF-IDF similarity was used only as a supporting indicator. Its value lies in identifying zones of stronger or weaker local continuity, not in producing a final aesthetic judgment.

Exploratory Rating Summary

Table 4 summarizes the exploratory rating outcomes across three major evaluation axes: narrative coherence, character continuity and development, and originality in relation to Arabic narrative heritage. These ratings were used as a structured interpretive summary of the matrix results.

Axis	N	Mean	SD	Test statistic	p-value / method
Narrative coherence	8	4.25	0.463	t = 7.638	p = 0.000122; one-sample t-test vs 3
Character continuity and development	7	4	0	-	p = 0.015625; exact sign/binomial test vs 3
Originality and Arabic narrative heritage	6	4.333	0.516	t = 6.325	p = 0.001456; one-sample t-test vs 3

Table 4: Exploratory rating summary.

Copyright© Noshokaty IE, et al.

The exploratory rating summary suggests that the applied matrix scores The Temple above a neutral midpoint across coherence, character continuity, and originality/heritage axes. These inferential values should be interpreted cautiously because the rating matrix is small and was applied to a single case study. Future work should test the same matrix with multiple novels, multiple human raters, and inter-rater agreement measures.

Ethical Narrative Brake Findings

The analysis identified repeated Ethical Narrative Brake patterns in scenes involving disclosure, investigation, blame, and evidence handling. These brakes appeared when the narrative approached a possible collapse into public exposure, revenge logic, or binary moral judgment, but then introduced a restraining mechanism that redirected the plot toward preservation, limited access, delayed exposure, or accountable responsibility.

The most important ENB pattern was the movement from exposure to preservation. Rather than treating evidence as material for public shaming, the narrative repeatedly redirects evidence into controlled storage, system logging, or symbolic preservation. This pattern supports narrative stability because it prevents ethically intense scenes from becoming abrupt moral shortcuts. It also allows character motivation to remain traceable: decisions are not presented as sudden reversals, but as responses to graded conditions such as risk intensity, evidence certainty, potential harm, responsibility level, right-to-know, preservation need, and narrative stability.

These findings suggest that the ENB layer functions as a stabilizing mechanism within the fictional world. It does not merely add moral language to the narrative; it changes the direction of scenes, regulates disclosure, and protects the narrative from collapsing into simplistic judgment. In this sense, the fuzzy-logic-inspired ENB parameters help explain how The Temple sustains ethical complexity while maintaining causal and character continuity.

Discussion

The results suggest that AI-generated fiction can display measurable narrative coherence when generation occurs within a designed fictional world. In The Temple, coherence does not depend only on surface fluency or sentence-level continuity. It emerges from repeated anchors, causal reactivation, character tracking, and the integration of documents and system logs as active plot mechanisms. This finding is consistent with discourse-based views of coherence, where meaning is produced through functional relations between textual units such as cause, elaboration, evidence, and contrast [8]. It also aligns with computational approaches that treat lexical recurrence, semantic relatedness, and entity continuity as measurable indicators of textual cohesion [9,10]. The 110 Hz motif functions as an audio-symbolic anchor that links the stone inscription, the temple, the investigation frame, Geneva-related material, and the ethical climax. This supports the view that long-form AI-generated fiction can become more coherent when the fictional world contains stable symbolic and causal reference points. Since Transformer-based language models can generate extended textual sequences and maintain local semantic patterns across long spans of text [1,2], the challenge is not only whether such systems can produce fluent language, but whether the generated world can sustain continuity across characters, motifs, and causal transitions. The distinction between Enosh and ENOSH is methodologically important. Treating Enosh as a fictional character allows the study to code moral movement, turning points, voice, and character responsibility. Treating ENOSH as a technical system prevents a common analytical error: attributing a human-style character arc to a platform or system entity. ENOSH is therefore evaluated through system functions such as monitoring, compensation, mode switching, output control, and narrative redirection, rather than through emotional or moral development. This distinction strengthens the repeatability of the matrix because it prevents entity confusion during normalization, counting, and interpretation. It also reflects the need for transparent and reproducible evaluation protocols in the study of large language model outputs [3,6].

The Ethical Narrative Brake (ENB) provides the most important conceptual contribution of the study. ENB describes how a generated fictional world can include internal constraints that regulate disclosure, accountability, evidence handling, and moral judgment. In the case examined here, the narrative does not simply choose between public exposure and concealment. Instead, it repeatedly moves toward controlled preservation with responsibility. This is why the paper interprets sacrifice not as self-destruction, passive silence, or public accusation, but as knowledge-bearing under ethical constraint. This interpretation is relevant to broader concerns in AI evaluation, where generated outputs must be assessed not only for fluency, but also for risk, bias, harmful framing, and unreliable reasoning [5,7]. The fuzzy-logic-inspired ENB parameters help explain how this braking mechanism operates. Ethical decisions in the novel are not treated as binary choices between disclosure and non-disclosure, guilt and innocence, or justice and concealment. Instead, the narrative appears to balance graded conditions such as risk intensity, evidence certainty, potential harm,

Copyright© Noshokaty IE, et al.

responsibility level, right-to-know, preservation need, and narrative stability. This graded treatment is consistent with the basic logic of fuzzy-set theory, where membership is not restricted to absolute true/false categories but can vary by degree [12]. When potential harm is high, the narrative does not automatically move toward public exposure, even if evidence is present. It redirects the action toward preservation, restricted access, delayed release, or symbolic storage. This graded mechanism helps protect the narrative from abrupt moral collapse. Accordingly, ENB functions as a narrative stabilizer. It slows down scenes that might otherwise collapse into revenge logic, public shaming, or simplistic blame. It also protects character continuity because ethical decisions are presented as responses to accumulated pressures rather than sudden reversals. In this sense, the fuzzy ENB layer supports both ethical complexity and narrative coherence. It gives the fictional world a regulating mechanism that allows conflict to intensify without destroying the internal logic of the plot. The findings also support a broader hypothesis for AI narrative design: stronger internal constraints may improve character stability, reduce arbitrary moral shifts, and help long-form generated narratives maintain causal continuity. However, this remains a hypothesis rather than a universal claim. The present study is based on a single applied case, and the matrix should be tested on additional AI-generated and AI-assisted novels, different prompting strategies, different model families, and multiple languages. Future research should also include multiple human raters and inter-rater agreement measures to test whether the proposed coding categories remain stable across evaluators. Future versions of the matrix may also incorporate deeper semantic similarity measures such as BERTScore or contextual embedding-based comparison to complement TF-IDF-based local cohesion indicators [12]. Overall, the study shows that evaluating AI-generated fiction requires more than judging whether the language is fluent. A generated novel must be examined through its anchors, entity continuity, causal transitions, ethical constraints, and ability to maintain character motivation across long narrative spans. The proposed matrix offers one repeatable way to connect literary interpretation with computational proxies, while preserving the role of human critical judgment.

Limitations and Future Work

This study has several limitations. First, name frequency, page coverage, and consecutive presence runs are operational proxies for narrative continuity. They indicate whether a character, symbolic element, or system entity remains available for tracking across the text, but they do not by themselves prove emotional depth, dramatic complexity, or psychological development. These dimensions still require qualitative interpretation.

Second, TF-IDF cosine similarity captures lexical and topical continuity between adjacent pages, but it does not fully measure deep semantic coherence, aesthetic quality, narrative suspense, or symbolic meaning. A low similarity score may reflect a meaningful scene transition rather than a weakness in coherence, while a high score may reflect repeated vocabulary without deeper narrative development. For this reason, TF-IDF similarity was used only as a supporting indicator.

Third, the study is based on a single applied case, The Temple. Although the results show that the proposed matrix can document coherence, character continuity, and Ethical Narrative Brake patterns within one AI-generated Arabic novel, the findings should not be generalized to all AI-generated fiction without further testing. Future studies should apply the same matrix to additional AI-generated and AI-assisted novels across different languages, genres, prompting strategies, and model families.

Fourth, the qualitative coding was structured but should be expanded through multi-rater evaluation. Future work should involve several human coders and calculate inter-rater agreement to test whether the coding categories, especially value arcs, transition types, and ENB activation, remain stable across different evaluators.

Finally, future computational extensions may strengthen the matrix by incorporating deeper linguistic and semantic tools. These may include Coh-Metrix for cohesion and readability features, BERTScore or contextual embedding similarity for semantic comparison, topic modeling for thematic development, and narrative-role tracking for analyzing character functions across scenes. Such extensions would allow the framework to move from a single applied case study toward a broader, reusable protocol for evaluating long-form AI-generated fiction.

Conclusion

This study proposed a reproducible measurement matrix for evaluating AI-generated fiction and applied it to The Temple as an Arabic case study. The matrix combines qualitative structural-semiotic coding with supporting quantitative proxies, including entity mentions, page coverage, consecutive presence runs, anchor tracking, and adjacent-page TF-IDF cosine similarity. The applied results indicate that the novel maintains a recognizable network of symbolic and causal anchors, together with measurable continuity for major characters, symbolic elements, and system entities.

Copyright© Noshokaty IE, et al.

The study contributes to AI evaluation and digital humanities research in two main ways. First, it provides a practical framework for evaluating AI-generated narrative beyond impressionistic literary judgment. By separating page-level quantitative indicators from scene-level qualitative interpretation, the matrix offers a transparent method for documenting narrative coherence, character continuity, and local textual cohesion. Second, the study introduces the Ethical Narrative Brake (ENB) as a conceptual and operational tool for distinguishing between ethical design embedded within the fictional world and ethical language that is merely simulated at the surface level.

The fuzzy-logic-inspired ENB parameters further show how ethically sensitive scenes can be evaluated through graded conditions such as risk intensity, evidence certainty, potential harm, responsibility level, right-to-know, preservation need, and narrative stability. This approach helps explain how a generated narrative may avoid abrupt moral collapse, irresponsible disclosure, or simplistic binary judgment while preserving causal and character continuity.

The proposed matrix is not intended to produce a universal aesthetic score or replace human critical reading. Rather, it offers a repeatable evaluation protocol that can support interpretation with documented computational proxies. Future studies should test the framework across additional AI-generated and AI-assisted novels, different languages, different model families, and multiple human raters. Such work may help develop more reliable methods for assessing long-form generative fiction in Arabic and multilingual contexts.

Declarations

Funding

The authors received no external financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare that they have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The coding sheets, page-level indicators, and supporting measurement outputs can be provided as supplementary material or made available from the corresponding author upon reasonable request.

Ethics Approval

Not applicable. This study analyzes a fictional literary text and does not involve human participants, clinical data, personal data, or experimental intervention.

AI Use Disclosure

During the preparation of this manuscript, the authors used ChatGPT and Google Colab to support language refinement, condensation of author-written text, consistency checking, and organization of computational indicators. The study design, conceptual framework, coding decisions, analysis, interpretation, and conclusions were developed and verified by the authors. All AI-assisted outputs were critically reviewed, revised, and approved by the authors, who take full responsibility for the final content of the manuscript.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. (2017) Attention is all you need. Advances in Neural Information Processing Systems 30.

2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems 33: 1877-1901.

3. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, et al. (2021) On the opportunities and risks of foundation models. arXiv.

4. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, et al. (2024) Large language models: A survey. arXiv.

5. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

6. Laskar MTR, Alqahtani S, Bari MS, Rahman M, Khan MAM (2024) A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. arXiv.

7. Kenthapadi K, Sameki M, Taly A (2024) Grounding and evaluation for large language models: Practical challenges and lessons learned (Survey). arXiv.

8. Mann WC, Thompson SA (1988) Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3): 243-281.

Copyright© Noshokaty IE, et al.

9. Graesser AC, McNamara DS, Louwerse MM, Cai Z (2004) Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers 36(2): 193-202.

10. Barzilay R, Lapata M (2008) Modeling local coherence: An entity-based approach. Computational Linguistics 34(1): 1-34.

11. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2020) International Conference on Learning Representations.

12. Zadeh LA (1965) Fuzzy sets. Information and Control 8(3): 338-353.

Copyright© Noshokaty IE, et al.