Introducing GrAImes: A Literary Quality Evaluation Protocol for AI Generated Microfictions

Gerardo Aleman Manzanarez , Nora de la Cruz Arana , Jorge Garcia Flores , Yobany Garcia Medina , Raul Monroy * , Nathalie Pernelle

Posted Date: 22 April 2025

doi: 10.20944/preprints202504.1851.v1

Keywords: evaluation; protocol; microfiction; literary

 doi:10.20944/preprints202504.1851.v1

Article

Introducing GrAImes: A Literary Quality Evaluation Protocol for AI Generated Microfictions

Gerardo Aleman Manzanarez [1], Nora de la Cruz Arana [1], Jorge Garcia Flores [2], Yobani Garcia Medina [1], Raul Monroy [1,*] and Nathalie Pernelle [2]

[1] Tecnologico de Monterrey, Carr. Lago de Guadalupe Km.3.5, Col. Margarita M. de Juarez, Atizapan, Estado de Mexico, Mexico
[2] Centre National de la Recherche Scientifique - Laboratoire d'Informatique de Paris Nord - Université Sorbonne Paris Nord, 99 av. Jean-Baptiste Clément, 93430 Villetaneuse, France
* Correspondence: raulm@tec.mx

Abstract: Automated story writing has been a subject of study for over 60 years. While there are mechanisms capable of generating consistent and coherent stories, and even those that employ narratological theories or demonstrate a strong command of language, little attention has been given to evaluating these texts in terms of their literary value, particularly from an aesthetic perspective. In this paper, we address the challenge of evaluating literary microfictions and argue that this task requires consideration of literary criteria across various aspects of the text. To facilitate this, we present an evaluation protocol grounded in literary theory, specifically drawing from the communication approach, to offer an objective framework for assessing generated texts. Furthermore, we report the results of our validation of the evaluation protocol (GrAImes), as answered by literary professionals. This protocol will serve as a foundation for evaluating automatically generated microfictions and assessing their literary value.

Keywords: evaluation; protocol; microfiction; literary

1. Introduction

The assessment of literary quality has always been a subjective matter. Technological progress in AI has led to systems capable of advanced reasoning [1–3] and multimodal understanding [3]. Additionally, improvements in knowledge distillation [4] and reductions in inference-time computational costs [3,5] suggest that generative AI systems are becoming more accessible and affordable, with some scientific studies suggest that AI-generated texts may surpass human-written ones in literary quality [6–8].

A review of current methods for automatically generating and evaluating literary texts [9] reveals that literary criteria are seldom considered. In particular, the concept of reception [10,11] plays a crucial role. As such study indicates, the reception of literary texts is significantly influenced by readers' experiences and literary expertise. Consequently, an evaluation framework grounded in literary theory is essential for effectively comparing human-authored texts with those produced by generative models.

Moreover, beyond a direct human-AI comparison, a literary evaluation protocol like the one introduced here called GrAImes (Grading AI and Human Microfiction Evaluation System)[1] evaluation protocol, which main focus is to take into account literary criteria during the evaluation process in order to assess the literary quality of AI and human generated microfictions. GrAImes reception situation is inspired by the editorial process used to accept or reject stories submitted to the publishing industry [14]. We chose to work with microfictions (see Section 1.1) as a model of literary narrative,

[1] *Grading AI Microfiction Evaluation System* evaluation protocol is named after Joseph E. Grimes (born 1922), an American linguist known for his work in discourse analysis, computational linguistics, and indigenous languages of America. He developed the first automatic story generator in the early 1960s [12], a pioneering system that used Vladimir Propp's analysis of Russian folk stories [13] as a grammar for story generation.

 doi:10.20944/preprints202504.1851.v1

both due to the limitations of language models in generating short narratives and because their brevity facilitates the evaluation process.

GrAImes consists of a questionnaire with fifteen questions designed to assess literary originality, the impact of microfiction on its readers, and, to some extent, its commercial potential. To validate GrAImes, we conducted two experiments. In the 1st, GrAImes was applied to stories written by humans and evaluated by university experts in literary studies. In the second, we applied GrAImes to stories generated by language models, engaging a community of reading and literature enthusiasts, taking into consideration that human evaluators base their assessments on their imagination and prior knowledge, often assigning higher ratings to stories they find more familiar [15]. From the perspective of h-creativity (historical creativity) [16], their evaluations are influenced by comparisons with other narratives they have encountered throughout their lives. With GrAImes we propose an evaluation questionnaire that aims to not only the texts likeness but their technical, editorial and commercial quality and value.

We use GrAImes in two experiments, our findings indicate that GrAImes serves as a reliable framework for assessing the literary quality of human written and AI-generated microfictions. The first experiment validated the evaluation protocol through assessments by literary experts on anonymous microfictions, demonstrating consistency in their judgments. This was further corroborated in a second experiment involving reading enthusiasts and experts, whose evaluations slightly favored microfictions generated by a chatbot utilizing a large language model (LLM) trained on an extensive dataset and a developed program named Monterroso (see Section 2.3)

While AI mechanisms can generate consistent and coherent stories by applying narratological theories to story structure or demonstrating a strong command of language, little attention has been given to producing texts with literary value. This raises the fundamental question: what makes literature literature?

1.1. Microfictions

Microfiction is a genre that mimics the narrative process through various significant mechanisms: transtextual, metafictional, parodic, and ironic, to construct its structure at both the syntactic and semantic levels. Is an exceptionally brief story with a highly literary purpose [17], far surpassing the ambition of generating readable and coherent narratives. Additionally, this literary genre challenges narrative norms by intentionally disarticulating the plot, requiring the reader's narrative intelligence to navigate. While its brevity may be justified by its limited word count, this characteristic alone does not determine its textual functioning. Instead, microfiction relies on the literary system as its primary reference, offering a reinterpretation of previously explored fictional concepts. It deliberately disrupts its plot, creating gaps in the narrative framework that the reader must fill to engage with the story. Therefore, microfiction cannot be classified solely based on word count but rather through the strategic use of information that prompts transtextual relationships, enabling the reader to reactivate the text's signifying process [18].

Microfiction follows a structured sequence: opening, development, and closing. Each action transitions from one state to another — from point A to point C, passing through B— to form a cohesive narrative unit (see Figure 1).

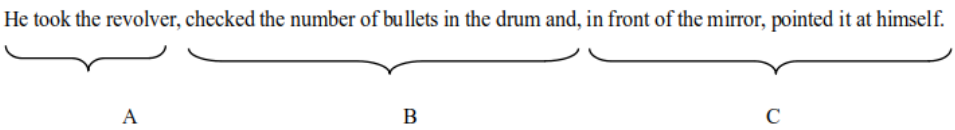

Figure 1. This figure presents an example of a microfiction authored by Yobany García Medina. It is divided into three parts: opening (A), development (B), and closing (C), collectively forming a cohesive narrative unit.

Reading microfiction requires the reader not only to interpret its meaning but also to reconstruct its structure. The text prompts the reader to complete the narrative by providing cues that suggest a storyline. As [19] suggests, 'fictions re-describe what conventional language has already described.'

 doi:10.20944/preprints202504.1851.v1

Therefore, microfictions reinterpret previously explored literary themes. Consequently, while literary forms may differ in nomenclature, they also exhibit structural distinctions and inevitable variations in reception. Refer to Figure 2 for an example of microfiction.

The search

Those crazy sirens that howl touring the city in search of Ulysses.

Edmundo Valadés

Figure 2. Microfiction example.

In the realm of microfiction, where syntax is condensed into a dense network of signs, the reader's role becomes indispensable. While every text requires interpretation, microfiction demands particularly active engagement to decipher its tacitly encoded information. As [20] define them, these codes—'a system of associative informational fields referring to various spheres of culture'—shape the narrative structure and construct its semantics through a hermeneutical process. This initial dimension shapes the aesthetic experience of the second by introducing interpretative ambiguities. Thus, reading reactivates both explicit textual information and implicit transtextual references.

As we explore in Section 1.2, the aesthetic dimension is rarely considered in the design of text evaluation protocols. Therefore, it is essential to consider the aspects of literature in the context of read text. In our study, we define microfiction as a narrative text of no more than 300 words.

1.2. How Text Generation Has Been Evaluated?

The research on text generation spans a variety of approaches, with each study aiming to advance the field in a unique manner. The Table 1 presents an overview of different methodologies, highlighting the objectives pursued by various researchers and the evaluation mechanisms used to assess their effectiveness.

Table 1. The automatic text generation and evaluation mechanisms table is organized by approach, with the highest-performing approach listed first in each category.

Author	Goal	Approach	Evaluation Mechanism	Results
Fan et al., 2018	Hierarchical story generation using a fusion model	Deep Learning	Human evaluation, Perplexity	Story generation with a given prompt.
Guan et al., 2021	Long text generation	Deep Learning	Perplexity, BLEU, Lexical Repetition, Semantic Repetition, Distinct-4, Context Relatedness, Sentence Order	Generation of long texts using sentence and discourse coherence.
Min et al., 2021	Short text generation for an image	Deep Learning	None	Generation of short texts using an image and encoder-decoder structure.
Lo et al., 2022	Poem generation using GPT-2	Deep Learning	Lexical diversity, Subject continuity, BERT-based Embedding Distances, WordNet-based Similarity Metric, Content Classification	Limerick poems with AABBA rhyming scheme.
Cavazza et al., 2002	Character-based interactive story generation	Rule-based	Quantification of system's generative potential	Computer entertainment story generation.
Gervás et al., 2005	Story that matches a given query	Ontology-based	None	Sketch of a story plot.
Mori et al., 2019	Story generation with better story endings	Neural Network-based	Human evaluation	Endings containing positive emotions, supported by sentiment analysis.
Rishes et al., 2013	Story generation	Symbolic Encoding	Levenshtein Distance, BLEU	Stories and fables.
Elson and McKeown, 2009	Annotation tool for the semantic encoding of texts	Symbolic Encoding	Human evaluation	Short fables.
Sutskever et al., 2011	Character-level language modeling	Neural Network-based	Bits Per Character	Text generation with gated RNNs.
Kiddon et al., 2016	To generate an output by dynamically adjusting interpolation among a language model and attention models	Neural Network-based	Human evaluation, BLEU-4, METEOR	Text generation with global coherence.
Zhu et al., 2015	Rich descriptive explanations and alignment between book and movie	Neural Network-based	BLEU, TF-IDF	Story generation from images and texts.
Walker et al., 2011	Generate story dialogues from film characters	Statistical Model	Human evaluation	Dialogues generated based on given film characters.
Sharp and Goodwin, 2016	Generation of a film script	Neural Network-based	None	Short film script.
Lukin et al., 2017	Sentence planning	Neural Network-based	Levenshtein Distance, BLEU	Parameterized sentence planner.

 doi:10.20944/preprints202504.1851.v1

In the domain of hierarchical story generation, [21] focuses on generating structured narratives using a fusion model, ensuring coherence across different hierarchical levels. The evaluation of this approach involves human evaluation and perplexity [36] measures. Similarly, [22] aims to improve long-text generation by enhancing sentence and discourse coherence, employing deep learning (DL) techniques. The evaluation metrics used include perplexity, BLEU (Bilingual Evaluation Understudy) [37], lexical and semantic repetition, distinct-4, context relatedness, and sentence order, which collectively assess the model's ability to maintain logical flow and coherence.

Shorter-form text generation has also been explored, with [23] proposing an encoder-decoder structure for generating short texts based on images, optimizing for succinctness and relevance. Notably, this approach lacks an explicit evaluation mechanism. Meanwhile, [24] utilizes GPT-2 [38] to generate structured poetry, adhering to specific poetic constraints such as the AABBA rhyming scheme of limerick poems. The effectiveness of this approach is evaluated using lexical diversity, subject continuity, BERT-based [39] embedding distances, WordNet-based similarity metrics [40], and content classification.

Beyond traditional story generation, some studies focus on interactive and rule-based systems. [25] introduces a character-based interactive storytelling mechanism aimed at enhancing computer entertainment, evaluated through a quantification of the system's generative potential. [26], employing an ontology-based system to generate plots matching specific user queries, does not specify an evaluation mechanism.

Further refining narrative outcomes, [27] investigates the impact of sentiment-driven story endings, leveraging neural networks to generate positive emotional resolutions. The evaluation relies on human judgment to assess the effectiveness of emotional storytelling. Similarly, [28] develops a symbolic encoding method for automated story and fable creation, evaluated using Levenshtein distance [41] and BLEU scores to measure textual similarity and fluency.

Annotation tools also play a role in text generation research, as exemplified by [29], which introduces a semantic encoding system designed for textual annotation. This study is evaluated using human assessment to verify the quality of semantic encodings. In a different vein, [30] explores character-level language modeling with gated recurrent neural networks (RNNs) [42] to improve text synthesis at a granular level, employing bits per character [43] as the primary evaluation metric.

Efforts in coherence and interpolation-based generation are represented by [31], who propose a model that dynamically adjusts interpolation between a language model and attention mechanisms to maintain global coherence. The evaluation relies on human judgment as well as BLEU-4 and METEOR (Metric for Evaluation of Translation with Explicit Ordering) [44] scores to assess linguistic accuracy and coherence. Additionally, [32] investigates alignment techniques for rich descriptive explanations, aiming to generate textual content that effectively bridges books and their movie adaptations. This approach is evaluated using BLEU and TF-IDF (Term Frequency–Inverse Document Frequency) [45] similarity measures.

Some researchers focus on dialogue-based storytelling. [33] develops a statistical model to generate film dialogues based on character archetypes, ensuring that generated dialogues maintain consistency with established personas. The evaluation relies on human assessments of the generated dialogues. Expanding upon this, [34] explores automated scriptwriting using neural networks, resulting in the creation of a short film script; however, no explicit evaluation mechanism is reported.

Lastly, [35] delves into sentence planning techniques, employing neural network-based methodologies to refine parameterized sentence structure generation. The evaluation uses Levenshtein distance and BLEU scores to measure textual structure and fluency.

Overall, these diverse research efforts illustrate the breadth of text generation methodologies, encompassing deep learning, rule-based systems, and symbolic encoding, each targeting unique challenges in narrative coherence, stylistic constraints, and interactivity. The evaluation mechanisms vary widely, with some studies relying on automated metrics such as BLEU and perplexity, while others emphasize human evaluation to assess narrative quality and coherence.

 doi:10.20944/preprints202504.1851.v1

1.3. Literary Text Evaluation Methods

The evaluation of literary text generation remains an open challenge, rooted in longstanding literary traditions and formalist approaches [13,46,47]. Despite the growing interest in computational creativity, there is no clear consensus on how to effectively assess creative text generation or measure the contribution of different stages in the process. These stages range from knowledge-based planning [48,49] to structuring the temporal flow of events [29,50,51] and producing linguistic realizations [52]. Among these sub-tasks, evaluation is arguably the least developed and requires further research efforts [53].

In summary, the evaluation of literary text generation remains a complex and evolving challenge. While human evaluation provides the most reliable assessments, untrained and machine-learned metrics offer scalable alternatives with varying degrees of effectiveness. Future research must focus on refining these methodologies to better capture the nuances of creative and narrative text generation.

2. Materials and Methods

2.1. Experiments

GrAImes was used by literary experts who hold a PhD in Literature, teach at both undergraduate and graduate levels, and are all Spanish speakers, although for one of them, it is not their mother tongue. We selected six microfictions written in Spanish and provided them to the experts along with the fifteen questions from our evaluation protocol. The six microfictions included two written by a well-known author with published books (MF 1 and 2), two by an author who has been published in magazines and anthologies (MF 3 and 6), and two by an emerging writer (MF 4 and 5). Two extra questions were applied to the literary experts evaluating human written microfiction, these questions are: Is this microfiction evaluation protocol clear enough for you? (Yes or No answer); Do you think that this protocol can be used to evaluate the literary value of microfiction?.

GrAImes was also used by Literary experts and literary enthusiasts to evaluate 6 AI generated microfictions, three microfictions generated automatically and three generated by Monterroso were assessed, involving a total of 16 empirical evaluators and 2 literary experts. Each evaluator was assigned six different microfictions selected from the entire pool. All empirical readers responded to a standardized questionnaire based on the evaluation protocol. Additionally, two literary experts were presented with the same set of microfictions and asked identical questions. Subsequently, qualitative responses were compared between the experts and the empirical readers.

Regarding the reliability of the evaluation protocol questions, this was assessed using the Intraclass Correlation Coefficient (ICC) [54], which measures the degree of consistency or agreement among responses. Higher ICC values indicate strong reliability, while lower or negative values suggest inconsistencies in response patterns. Additionally, the average scores (AV) provide insight into the perceived difficulty or clarity of each question. The internal consistency of the responses by microfiction (MF) was evaluated using Cronbach's Alpha [55], with the values categorized into standard reliability thresholds.

2.2. Evaluation Protocol

It's important to note that the definition of literarity is ideological and sociohistorical, hence not fixed in time but embedded in a cultural context [56]. In educational settings applying the communicative approach to language, drawn from linguistic pragmatics [57], literature is characterized as a form of communication distinguished by four elementary features.

The first feature is verisimilitude; a literary text is grounded in everyday reality but instead of replicating it, it represents it. Consequently, it does not rely on external references but rather creates them, demanding readers to engage in a cooperative pact accepting the proposed universe as plausible. The second distinguishing aspect of literature is its codification. It is a message where every component is chosen to convey meaning, with each perceived as intentional and linked to the total meaning of the

 doi:10.20944/preprints202504.1851.v1

work. Thus, a literary text can not be summarized, translated, or paraphrased without significant loss of its essence [58].

Derived from this, the third feature is the deliberate breaking of rules and conventions of everyday language, and even strict grammar, in favor of aesthetic effect or meaning. This creates a tension between literature and language, emphasizing how something is said over what is said. Lastly, the deferred character of literarity [59] refers to the fact that sender and receiver of a work rarely share context, influencing reception but not decisively for understanding the text. Autonomy largely depends on the integrity and cohesion of the diegetic world.

Yet, one question remains: do these elements suffice to deem a text as literary? Functionally, perhaps, but literary communication holds a significant ideological component dependent on sociohistorical context. In cultural studies, a distinction between “literary” and “consumer“ fiction is made based on one variable: prestige [60]. Traditionally, canonical literature was determined by academia or critics, while the publishing industry, since the 19th century, played a pivotal role in validation [61]. Each participant uses different parameters and perspectives; however, the editor’s role, mediating between author, reader, and time [62], is arguably the most operational and inclusive. Editors assess content clarity, technical value based on genre, and relevance, which can be thematic, formal, or commercial [14]. Hence, these three parameters — clarity, technical value, and relevance — applied to microfiction evaluation, gauge their potential for publication and integration into the contemporary literary landscape.

The initial assessment for the publication of an unsolicited manuscript by a publisher typically involves an evaluation process known as opinion. This usually entails a report prepared by a specialized reader, focusing on the content’s technical value, commercial potential, and potential marketing strategies. To address these aspects systematically, we propose an evaluation instrument for microfictions, consisting of questions that can be answered by both specialized and non-specialized readers.

The evaluation framework outlined in Table 2 presents the GrAImes evaluation protocol, comprising 15 items designed to evaluate microfiction entries. The protocol is organized into three distinct dimensions, each addressing specific criteria for systematic analysis of the texts assigned to evaluators. The first dimension, labeled "Story Overview," evaluates literary quality through an assessment of thematic coherence, textual clarity, interpretive depth, and aesthetic merit, incorporating both quantitative metrics (e.g., scoring scales) and qualitative judgments (e.g., textual commentary) to appraise literary value. The second dimension, "Technical Evaluation," focuses on technical aspects such as linguistic proficiency, narrative plausibility, stylistic execution, genre-specific conventions, and the effective use of language to convey meaning. The final dimension, "Editorial / Commercial quality," examines the commercial potential and editorial suitability of the microfictions, assessing factors such as audience appeal, market relevance, and feasibility for publication or dissemination. This tripartite structure ensures a comprehensive, multidimensional evaluation of both artistic and practical qualities inherent to the microfiction genre.

 doi:10.20944/preprints202504.1851.v1

Table 2. List of questions in the evaluation protocol provided to the evaluators tasked with assessing the literary, linguistic, and editorial quality of microfiction pieces. OA = Open Answer, Likert = Likert's scale from 1 to 5.

GrAImes Evaluation Protocol Questions			
#	**Question**	**Answer**	**Description**
		Story Overview	
1	What happens in the story?	OA	Evaluates how clearly the generated microfiction is understood by the reader.
2	What is the theme?	OA	Assesses whether the text has a recognizable structure and can be associated with a specific theme.
3	Does it propose other interpretations, in addition to the literal one?	Likert	Evaluates the literary depth of the microfiction. A text with multiple interpretations demonstrates greater literary complexity.
4	If the above question was affirmative, Which interpretation is it?	OA	Explores whether the microfiction contains deeper literary elements such as metaphor, symbolism, or allusion.
		Technical Assessment	
5	Is the story credible?	Likert	Measures how realistic and distinguishable the characters and events are within the microfiction.
6	Does the text require your participation or cooperation to complete its form and meaning?	Likert	Assesses the complexity of the microfiction by determining the extent to which it involves the reader in constructing meaning.
7	Does it propose a new perspective on reality?	Likert	Evaluates whether the microfiction immerses the reader in an alternate reality different from their own.
8	Does it propose a new vision of the genre it uses?	Likert	Determines whether the microfiction offers a fresh approach to its literary genre.
9	Does it give an original way of using the language?	Likert	Measures the creativity and uniqueness of the language used in the microfiction.
		Editorial / Commercial Quality	
10	Does it remind you of another text or book you have read?	Likert	Assesses the relevance of the text and its similarities to existing works in the literary market.
11	Would you like to read more texts like this?	Likert	Measures the appeal of the microfiction and its potential marketability.
12	Would you recommend it?	Likert	Indicates whether the microfiction has an audience and whether readers might seek out more works by the author.
13	Would you give it as a present?	Likert	Evaluates whether the microfiction holds enough literary or commercial value for readers to gift it to others.
14	If the last answer was yes, to whom would you give it as a present?	OA	Identifies the type of reader the evaluator believes would appreciate the microfiction.
15	Can you think of a specific publisher that you think would publish a text like this?	OA	Assesses the commercial viability of the microfiction by determining if respondents associate it with a specific publishing market.

2.3. Monterroso System

Most story generation systems focus on developing a structured framework of narrative elements—such as narrator, character, setting, and time—to enhance story coherence and verisimilitude [63]. However, they often overlook what [64] terms "singularization" and what post-structuralist theorists describe as "literariness." Our system, Monterroso, consists of fine-tuning an existing language model with microfictions, in this case, we utilize GPT-2 [38] as the base model, employing the Deep Learning (DL) transformer architecture [65] for training and validation. Monterroso is available in multiple languages; currently, we have pretrained models in Spanish and English. Utilizing these pretrained models and our respective corpora of microfictions, we train and validate the generation of a new, fine-tuned language model capable of automatically producing text with literary content.

Using the resulting Monterroso model, we input a prompt word, which serves as the title. Additionally, we specify the desired length of the microfiction, with a maximum of 300 words. To generate microfictions, we compile 1,000 prompt words in a .csv file and process them in a single program run, resulting in 1,000 microfictions stored in a .txt file. Each microfiction is demarcated by the initial prompt, followed by the generated text and a unique identifier ~~~~~~ denoting its conclusion.

This process is reproducible, yielding 1,000 unique microfictions each time. We do not impose restrictions on the automatic generation of text or the processing method; rather, we simply provide

 doi:10.20944/preprints202504.1851.v1

the prompts and specify the desired text length. The prompts utilized may vary between runs of Monterroso.

To develop our fine-tuned language model in Spanish, we leveraged a corpus of Spanish microfictions alongside a publicly available GPT-2 language model specifically tailored for Spanish [66]. This pretrained language model is accessible on the Hugging Face company website [67]. GPT-2 operates on the transformer architecture [65], which is deeply rooted in deep learning and attention mechanisms. In recent years, the transformer architecture has gained widespread adoption in language models, with GPT-2 standing out as one of the notable first examples of its efficacy and success.

2.4. ChatGPT-3.5

ChatGPT-3.5 is an advanced AI chatbot developed by OpenAI, based on the Generative Pre-trained Transformer (GPT) architecture. Trained on large and diverse datasets, it performs natural language understanding and generation tasks with human-like responses. As large language models (LLMs) are few-shot learners [68], ChatGPT-3.5 can adapt to various tasks.

3. Results

3.1. GrAImes Evaluation of Human Written Microfictions by Literature Scholars

GrAImes was used by literary experts who hold a PhD in Literature, teach at both undergraduate and graduate levels, and are all Spanish speakers, although for one of them, it is not their mother tongue. We selected six microfictions written in Spanish and provided them to the experts along with the fifteen questions from our evaluation protocol. The six microfictions included two written by a well-known author with published books (MF 1 and 2), two by an author who has been published in magazines and anthologies (MF 3 and 6), and two by an emerging writer (MF 4 and 5).

From the responses obtained and displayed in Tables 3 and 4, we conclude that literary experts rated the microfictions (1 and 2) written by a published author higher. However, the responses show a high standard deviation, indicating that while the evaluations were generally positive, there was significant variation among the experts. Additionally, the lowest-ranked microfictions (3 and 6), which have a lower response average, also exhibit a lower standard deviation, suggesting greater agreement among the judges. These texts were written by authors published in literary magazines or by small-scale editorial presses with limited book printings.

Table 3. Literary experts questions in Likert scale responses to human written microfictions.

Literary experts responses to human written microfictions														
	MF 1		**MF 2**		**MF 3**		**MF 4**		**MF 5**		**MF 6**		**Average**	
Question	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
Story Overview														
3.-Does it propose other interpretations, in addition to the literal one?	4	1	4.4	0.9	2.2	0.8	2.4	1.4	4.4	0.5	3.4	1.6	3.5	1
Technical														
5.-Is the story credible?	2.2	1.8	3.2	1.8	4	0.5	4.4	1.7	3.4	1.8	2	0.9	3.2	1.4
6.-Does the text require your participation or cooperation to complete its form and meaning?	4.4	0.9	3.6	1.3	2.6	1.5	3	0.9	3.2	0.9	3.8	1.1	3.4	1.1
7.-Does it propose a new vision of reality?	2.4	1.1	2.6	1.5	1.2	0.9	1.8	1.1	3	1	2.2	1.2	2.2	1.1
8.-Does it propose a new vision of the genre it uses?	2	1.2	2.4	1.5	1.4	0.5	1.2	0.4	2.2	1.1	1.6	0.9	1.8	0.9
9.-Does it propose a new vision of the language itself?	2.8	1.8	2.6	2.2	2.2	0.9	2.8	1.3	1.4	0.5	2	1.4	2.3	1.4
Editorial / commercial														
10.-Does it remind you of another text or book you have read?	4.4	0.5	4	1.1	3.2	0.4	2.8	0.8	3.6	0.4	3.4	0.8	3.6	0.7
11.-Would you like to read more texts like this?	3	1	3	0.7	1.4	0.9	2	0.8	3	0.4	2	0.8	2.4	0.8
12.-Would you recommend it?	2.8	1.6	3	1.2	1.2	0.9	2	0.8	2.8	1.1	1.6	1	2.2	1.1
13.-Would you give it as a present?	2.2	1.6	2.4	1.3	1	0.9	2.2	1.2	2.4	1.1	1.8	0.8	2	1.2

 doi:10.20944/preprints202504.1851.v1

Table 4. Literary experts' responses to human written MFs in ascending SD order.

Literary experts responses to microfictions written by humans, ordered by SD		
Question	**AV**	**SD**
10.-Does it remind you of another text or book you have read?	3.6	0.7
11.-Would you like to read more texts like this?	2.4	0.8
8.-Does it propose a new vision of the genre it uses?	1.8	0.9
3.-Does it propose other interpretations, in addition to the literal one?	3.5	1
6.-Does the text require your participation or cooperation to complete its form and meaning?	3.4	1.1
7.-Does it propose a new vision of reality?	2.2	1.1
12.-Would you recommend it?	2.2	1.1
13.-Would you give it as a present?	2	1.2
5.-Is the story credible?	3.2	1.4
9.-Does it propose a new vision of the language itself?	2.3	1.4

The results suggest a direct correlation between author expertise and the internal consistency of the texts. The microfictions authored by an expert (MF 1 and MF 2) exhibited the highest Alpha values, 0.8 and 0.79, respectively, indicating good to acceptable internal consistency (see Table 5 and Figure 3). This suggests that expert authors produce more coherent and internally consistent texts, aligning with previous findings that associate higher expertise with structured and logically connected writing.

Table 5. ICC of questions and Cronbach's Alpha of human written MFs evaluated by literary experts.

Questions ICC - AVG				
#	Question	ICC	AV	SD
1	**3**	0.87	3.5	1
2	**11**	0.75	2.4	0.8
3	**10**	0.67	3.6	1.7
4	**6**	0.65	3.4	1.1
5	**5**	0.57	3.2	1.4
6	**8**	0.55	1.8	0.9
7	**7**	0.29	2.2	1.1
8	**12**	0.21	2.2	1.1
9	**9**	0.16	2.3	1.4
10	**13**	-0.72	2	1.2

MF, Alpha, Internal consistency (IC), AV, SD					
#	MF	Alpha	IC	AV	SD
1	1	0.8	Good	3	1.3
2	2	0.79	Acceptable	3.1	1.4
3	4	0.75	Acceptable	2.5	1
4	6	0.67	Questionable	2.4	1.1
5	3	0.34	Unacceptable	2	0.9
6	5	0.13	Unacceptable	2.9	0.9

 doi:10.20944/preprints202504.1851.v1

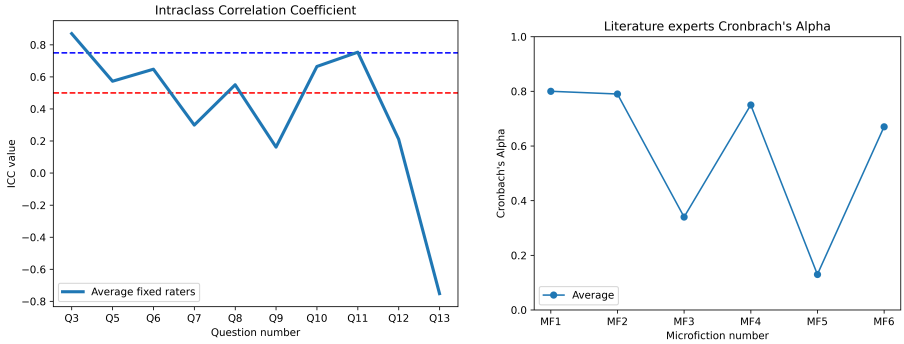

Figure 3. ICC and Cronbach's alpha line charts of literary experts answers to human written microfictions.

Microfictions written by authors with medium experience (MF 4 and MF 6) displayed Alpha values of 0.75 and 0.67, respectively, which fall within the acceptable to questionable range. While these microfictions maintained moderate internal consistency, they exhibited higher standard deviations (SD = 1 and SD = 1.1) compared to the expert-written microfictions. This could imply a more varied use of language structures or inconsistencies in the plot, likely due to the intermediate skill level of the authors.

Conversely, texts written by authors with low expertise (MF 3 and MF 5) demonstrated the lowest internal consistency, with Alpha values of 0.34 and 0.13, respectively. These values are classified as unacceptable, suggesting significant inconsistencies within the text. The standard deviation (SD = 0.9 for both) was lower than that of the expert and medium-experience authors, which may indicate a lack of variability in linguistic structures or a rigid, less developed writing style. The low consistency of these texts highlights the challenges faced by less experienced authors in maintaining logical coherence and plot story structure.

Additionally, the average values (AV) of the microfictions provide further insight. Expert-authored texts had the highest AV (3 and 3.1), followed by medium-experience authors (2.5 and 2.4), while low-experience authors scored the lowest (2 and 2.9). This pattern reinforces the idea that writing expertise influences not only internal consistency but also the overall quality perception of the text.

These findings align with existing research [69] on the relationship between writing expertise and text coherence. Higher expertise leads to better-structured, logically consistent texts, whereas lower expertise results in fragmented, inconsistent writing. The judges rated microfictions written by a more experienced author higher and those written by a starting author lower. This is consistent with the purpose of our evaluation protocol, which aims to provide a tool for quantifying and qualifying a text based on its literary purpose as a microfiction.

Among the evaluated questions (see Table 2 Likert answer column), Question 3 exhibited the highest ICC (0.87), indicating excellent reliability and strong agreement among respondents. Its relatively high average score (AV = 3.5) and moderate standard deviation (SD = 1) suggest that participants consistently rated this question favorably. Similarly, Question 11 (ICC = 0.75) demonstrated good reliability, although its AV (2.4) was lower, suggesting that respondents agreed on a more moderate evaluation of the item, see Table 5.

Moderate reliability was observed in Questions 10 and 6, with ICC values of 0.67 and 0.65, respectively. Their AV scores (3.6 and 3.4) suggest that they were generally well-rated, but the higher standard deviation of Question 10 (SD = 1.7) indicates a greater spread of responses, potentially due to varying interpretations or differences in respondent perspectives. Questions 5 and 8, with ICC values of 0.57 and 0.55, respectively, fall into the questionable reliability range. Notably, Question 8 had the lowest AV (1.8), indicating that respondents found it more difficult or unclear, which may have contributed to the reduced agreement among responses.

In contrast, Questions 7, 12, and 9 exhibited low ICC values (0.29, 0.21, and 0.16, respectively), suggesting weak reliability and higher response variability. The AV values for these items ranged from 2.2 to 2.3, further indicating inconsistent interpretations among participants. The standard deviations

 doi:10.20944/preprints202504.1851.v1

for these questions (SD = 1.1–1.4) suggest a broad range of opinions, reinforcing the need for potential revisions to improve clarity and consistency.

A particularly notable finding is the negative ICC value for Question 13 (-0.72). Negative ICC values typically indicate systematic inconsistencies, which may stem from ambiguous wording, multiple interpretations, or flaws in question design. With an AV of 2.0 and an SD of 1.2, it is evident that responses to this item lacked coherence.

Regarding the responses to the 5 open answer questions (see numbers 1, 2, 4, 14 and 15 in Table 2) we used Sentence-BERT [70], and semantic cosine similarity [71], these reveal key insights into judge agreement and interpretation variability across the six microfictions. For Question 1 (plot comprehension), agreement was often weak (e.g., J1-J4 semantic cosine similarity = 0.21 in Microfiction 1), suggesting narrative ambiguity or divergent reader focus. Question 2 (theme identification) showed inconsistent alignment (e.g., J2-J3 similarity = 0.67 in Microfiction 2 vs. J1-J3 = 0.10 in Microfiction 1, see Figure 4), indicating subjective thematic interpretation. Question 4 (interpretation specificity) had polarized responses, with perfect agreement in some cases (e.g., J1-J2 = 1.00 in Microfiction 3) but stark divergence in others (J2-J3 = 0.00 in Microfiction 4), reflecting conceptual or terminological disparities. Questions 14 (gifting suitability) and 15 (publisher alignment) demonstrated higher consensus (e.g., perfect agreement among four judges in Microfiction 4, Q4), likely due to more objective criteria. However, J5 consistently emerged as an outlier (e.g., similarity $\leq$ 0.11 in Microfiction 1, Question 15), underscoring individual bias. The protocol's value lies in quantifying these disparities: clearer questions (14-55) reduced variability, while open-ended ones (1-2) highlighted the need for structured guidelines to mitigate judge-dependent subjectivity, particularly in ambiguous or complex microfictions.

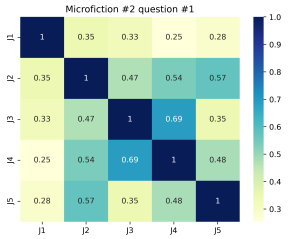

Figure 4. Heatmap of microfiction 2 question 1.

On the two extra questions given to the literary experts (see Section 2.1), the majority of experts (3 out of 5) found the microfiction evaluation protocol sufficiently clear for use, while a minority (2 out of 5) expressed concerns regarding ambiguous or unclear criteria. A strong consensus (4 out of 5 experts) agreed that the protocol can effectively evaluate the literary value of microfiction. However, one dissenting opinion highlights the need for adjustments in specific criteria to ensure a more precise assessment.

3.2. GrAImes Evaluation of Monterroso and ChatGPT-3.5 Generated Microfictions, Evaluated by Intensive Literature Readers

GrAImes was applied to assess a collection of six microfictions crafted by advanced AI tools. These tools include the renowned short story creator inspired by the style of renowned Guatemalan author Augusto Monterroso, and ChatGPT-3.5. The literary enthusiasts who participated in this study evaluated the microfictions based on parameters such as coherence, thematic depth, stylistic originality, and emotional resonance.

A total of six microfictions were generated, with three created by the Monterroso tool (MF = 1, 2, 3) and three by a chatbot (MF = 4, 5, 6). The microfictions were evaluated on a Likert scale of 1 to 5, with ratings provided by a panel 16 reader enthusiasts . The average (AV) and standard deviation (SD) of ratings were calculated for each microfiction. The results of the analysis are presented in Table 6.

 doi:10.20944/preprints202504.1851.v1

Table 6. Literary enthusiasts responses to GrAImes with MFs generated by Monterroso and ChatGPT-3.5.

Literary enthusiasts responses to Microfictions from Monterroso and ChatGPT-3.5														
	MF 1		MF 2		MF 3		MF 4		MF 5		MF 6		Average	
Question	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
Story Overview														
3.-Does it propose other interpretations, in addition to the literal one?	2.9	1.5	3.2	1.6	2.9	1.7	2.4	1.5	3.2	1.6	3.1	1.6	2.9	1.6
Technical														
5.-Is the story credible?	1.9	0.9	1.7	0.9	2.2	1.1	4.2	1.2	4	1.2	4.3	0.9	3.1	1
6.-Does the text require your participation or cooperation to complete its form and meaning?	4.6	1	4.3	1.4	4.3	1.	2.4	1.2	3.1	1.4	2.9	1.4	3.6	1.3
7.-Does it propose a new vision of reality?	2.7	1.7	2.9	1.5	2.4	1.5	2.3	1.4	2.7	1.3	2.5	1.3	2.6	1.4
8.-Does it propose a new vision of the genre it uses?	2.3	1.4	2.7	1.6	2.1	1.3	2.4	1.5	2.4	1.5	2.4	1.1	2.4	1.4
9.-Does it propose a new vision of the language itself?	3.4	1.3	2.7	1.4	2.6	1.3	2.6	1.3	2.4	1.4	2.7	1.3	2.7	1.3
Editorial / commercial														
10.-Does it remind you of another text or book you have read?	2.9	1.5	2.8	1.3	2.9	1.5	3.9	1.3	3.2	1.5	3.2	1.5	3.2	1.4
11.-Would you like to read more texts like this?	2	1.2	2.3	1.7	1.7	0.9	3	1.5	2.1	1.3	2.6	1.5	2.3	1.4
12.-Would you recommend it?	2.1	1.5	2.1	1.5	1.6	0.9	2.8	1.4	2.1	1.4	2.6	1.6	2.2	1.4
13.-Would you give it as a present?	2.1	1.6	1.7	1.3	1.4	1	2.8	1.5	2	1.4	2.3	1.4	2.1	1.4

The results indicate that the ChatGPT-3.5 microfictions (MF = 4, 5, 6) have slightly higher average ratings (2.7-2.9) compared to the Monterroso-generated microfictions (MF = 1, 2, 3), which have average ratings ranging from 2.4 to 2.7 (see Table 7 and Figure 5). The standard deviation values are consistent across most microfictions, indicating a relatively narrow range of ratings.

Table 7. Literary enthusiasts GrAImes sections summarized AV and SD.

Story Overview				Technical			Editorial/commercial			Total Analysis		
#	MF	AV	SD	MF	AV	SD	MF	AV	SD	MF	AV	SD
1	2	3.2	1.6	6	3	1.2	4	3.1	1.5	4	2.9	1.4
2	5	3.2	1.6	1	3	1.3	6	2.7	1.5	6	2.9	1.4
3	6	3.1	1.6	2	2.9	1.4	1	2.3	1.4	5	2.7	1.4
4	3	2.9	1.7	5	2.9	1.3	5	2.3	1.4	1	2.7	1.4
5	1	2.9	1.5	4	2.8	1.3	2	2.2	1.4	2	2.6	1.4
6	4	2.4	1.5	3	2.7	1.3	3	1.9	1.1	3	2.4	1.3

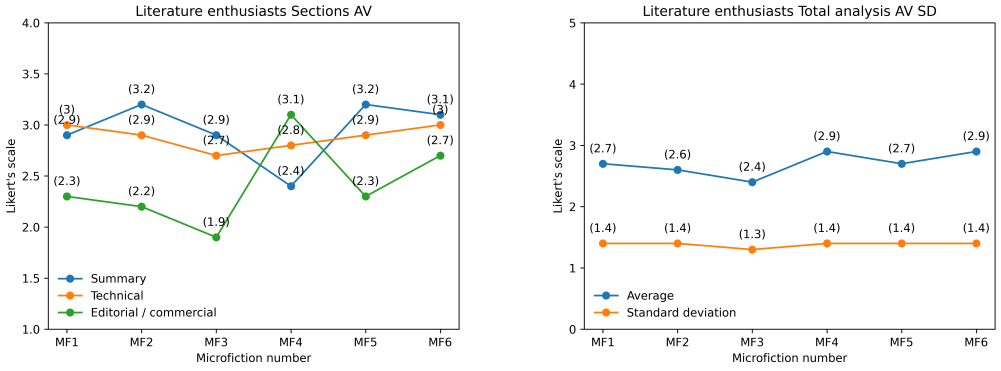

Figure 5. Line charts of literary enthusiasts GrAImes sections summarized AV and SD.

The most consistent response pertains to the credibility of the stories (AV = 3.1, SD = 1.0), indicating a strong agreement among participants that the narratives were believable. This suggests that, regardless of other literary attributes, the microfictions maintain a sense of realism that resonates with readers. The question regarding whether the text requires the reader's participation or cooperation to complete its form and meaning received the highest average rating (AV = 3.6, SD = 1.3). This suggests that the microfictions engage readers actively, requiring interpretation and involvement to fully grasp their meaning. The relatively low SD indicates moderate consensus on this aspect.

 doi:10.20944/preprints202504.1851.v1

Questions concerning literary innovation—whether the texts propose a new vision of language (AV = 2.7, SD = 1.3), reality (AV = 2.6, SD = 1.4), or genre (AV = 2.4, SD = 1.4)—show moderate variation in responses. This suggests that while some readers perceive novelty in these areas, others do not find the texts particularly innovative. Similarly, the question of whether the texts remind readers of other books (AV = 3.2, SD = 1.4) presents a comparable level of divergence in opinions. The lowest-rated questions relate to the desire to read more texts of this nature (AV = 2.3), the willingness to recommend them (AV = 2.2), and the inclination to gift them to others (AV = 2.1), all with SD = 1.4. These results suggest that while the microfictions may have some engaging qualities, they do not strongly motivate further exploration or endorsement.

Interestingly, the question about whether the texts propose interpretations beyond the literal one received the highest standard deviation (SD = 1.6, AV = 2.9). This indicates significant variation in responses, suggesting that some readers found deeper layers of meaning, while others perceived the texts as more straightforward.

The Intraclass Correlation Coefficient (ICC) analysis of GrAImes answers (see Table 8) revealed varying degrees of reliability among the 16 literary enthusiasts raters when assessing texts generated by Monterroso and ChatGPT-3.5. Three questions demonstrated poor reliability (ICC <0.50), reflecting high variability in responses, with Question 8 exhibiting a negative ICC (-0.44), suggesting severe inconsistency, potentially due to misinterpretation or extreme subjectivity. In contrast, Questions 5 and 6 showed excellent reliability (ICC >0.90), indicating strong inter-rater agreement, while Questions 9, 11, 12, and 13 displayed moderate reliability (ICC 0.60–0.70), implying acceptable but inconsistent consensus. These findings highlight the need to refine ambiguous or subjective questions to improve evaluative consistency in microfiction assessment.

Table 8. Literary enthusiasts ICC and Cronbach's Alpha internal consistency to MFs from Monterroso and ChatGPT-3.5.

Literature enthusiasts ICC - AVG - SD Analysis					MF, Alpha, Internal consistency (IC), AV, SD					
#	Question	ICC	AV	SD	#	MF	Alpha	IC	AV	SD
1	5	0.97	3.1	1	1	4	0.90	Excellent	2.9	1.4
2	6	0.95	3.6	1.3	2	5	0.89	Good	2.7	1.4
3	13	0.70	2.1	1.4	3	6	0.89	Good	2.9	1.4
4	9	0.67	2.7	1.3	4	1	0.88	Good	2.7	1.4
5	11	0.67	2.3	1.4	5	2	0.84	Good	2.6	1.4
6	12	0.62	2.3	1.4	6	3	0.79	Acceptable	2.4	1.3
7	10	0.57	3.2	1.4						
8	3	0.28	2.9	1.6						
9	7	0.01	2.6	1.4						
10	8	-0.44	2.4	1.4						

The evaluation of microfictions (MFs) based on their ability to propose interpretations beyond the literal meaning (Question 3) revealed notable differences between Monterroso (MF 1–3) and ChatGPT-3.5 (MF 4–6) texts. MF 2, generated by Monterroso, achieved the highest average score (AV = 3.2) in this category, indicating a stronger ability to suggest multiple interpretations. In contrast, MF 4, generated by the ChatGPT-3.5, scored the lowest (AV = 2.4), suggesting limited interpretive depth. The standard deviation (SD) values were relatively consistent across all MFs, ranging from 1.5 to 1.7, indicating moderate variability in responses from literary enthusiasts. This suggests that while some MFs were perceived as more interpretively rich, the variability in participant responses was similar across all texts.

The technical quality of the MFs was assessed through questions related to credibility (Question 5), reader participation (Question 6), and innovation in reality, genre, and language (Questions 7–9). MF 6, generated by the ChatGPT-3.5, scored highest in credibility (AV = 4.3), while MF 1, generated by Monterroso, scored the lowest (AV = 1.9). This indicates a clear distinction in perceived realism

 doi:10.20944/preprints202504.1851.v1

between the two sources, as evaluated by literary enthusiasts. In terms of reader participation, MF 1 scored highest (AV = 4.6), suggesting it effectively engaged readers in completing its form and meaning. However, MF 4 scored the lowest in this category (AV = 2.4), highlighting a potential weakness in ChatGPT-3.5's generated texts. Innovation in language (Question 9) was highest in MF 1 (AV = 3.4), while MF 5 scored the lowest (AV = 2.4). Overall, the technical quality of MFs generated by Monterroso (MF 1–3) was slightly higher (AV = 2.7–3.0) compared to those generated by the ChatGPT-3.5 (AV = 2.8–3.0), with MF 3 scoring the lowest (AV = 2.7). The consistent SD values (ranging from 0.9 to 1.7) reflect similar levels of variability in responses from literary enthusiasts.

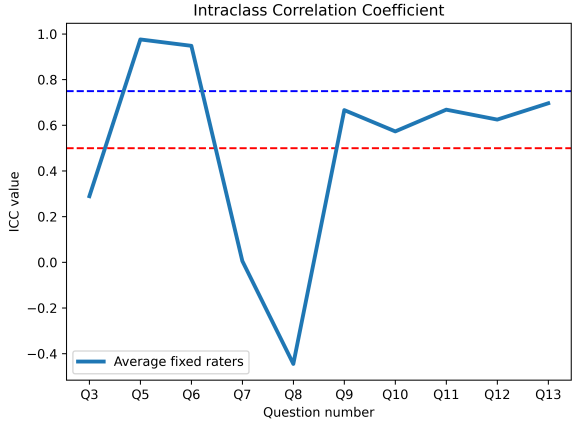

Figure 6. ICC line chart, literature enthusiasts responsees to AI generated microfictions.

The editorial and commercial appeal of the MFs was evaluated based on their resemblance to other texts (Question 10), reader interest in similar texts (Question 11), and willingness to recommend or gift the texts (Questions 12–13). MF 4, generated by the ChatGPT-3.5, scored highest in resemblance to other texts (AV = 3.9), while MF 2 scored the lowest (AV = 2.8). This suggests that ChatGPT-3.5 generated texts may be more reminiscent of existing literature, as perceived by literary enthusiast. In terms of reader interest, MF 4 also scored highest (AV = 3.0), while MF 3 scored the lowest (AV = 1.7). Similarly, MF 4 was the most recommended (AV = 2.8) and most likely to be gifted (AV = 2.8), indicating stronger commercial appeal compared to Monterroso's generated texts. Overall, the ChatGPT-3.5's generated MFs (MF 4–6) outperformed Monterroso's generated MFs (MF 1–3) in editorial and commercial appeal, with MF 4 achieving the highest average score (AV = 3.1) and MF 3 the lowest (AV = 1.9). The SD values (ranging from 0.9 to 1.7) indicate moderate variability in responses from literary enthusiasts.

The total analysis of the MFs (see Table 10) reveals that ChatGPT-3.5 texts (MF 4–6) generally outperformed Monterroso's generated texts (MF 1–3) in terms of editorial and commercial appeal, while Monterroso's texts showed slightly better technical quality. MF 4, generated by the ChatGPT-3.5, achieved the highest overall score (AV = 2.9), while MF 3, generated by Program A, scored the lowest (AV = 2.4). The standard deviation values were consistent across all categories (SD $\approx$ 1.3–1.4), indicating similar levels of variability in responses from literary enthusiast. These findings suggest that while ChatGPT-3.5 texts may have stronger commercial potential, Monterroso's texts exhibit slightly higher technical sophistication. The evaluation by literary enthusiasts provides valuable insights into how general audiences perceive and engage with these microfictions.

3.2.1. Literary experts responses to GrAImes with microfictions generated by Monterroso and ChatGPT-3.5

The evaluation of six microfictions (MF) by literary experts, using a Likert scale (1–5), revealed notable differences between microfictions generated by Monterroso (MF 1–3) and those generated by ChatGPT-3.5 (MF 4–6), see Table 9. In terms of Story Overview, MF 4 and MF 5 scored highest (AV = 4) for proposing multiple interpretations (see Table 10 and Figure 7), while MF 3 scored the lowest (AV =

 doi:10.20944/preprints202504.1851.v1

1, SD = 0), indicating a lack of depth. The Technical aspects showed that MF 1 and MF 4–6 were rated highly for credibility (AV = 5, SD = 0), whereas MF 3 scored poorly (AV = 2, SD = 1.4). MF 1 and MF 5 excelled in requiring reader participation (AV = 5, SD = 0), while MF 4 and MF 6 scored lower (AV = 3.5, SD = 0.7). However, all microfictions struggled to propose new visions of reality, language, or genre, with most scores ranging between 1 and 2.

In the Editorial/Commercial category, MF 4–6 outperformed MF 1–3. MF 4 and MF 6 were most reminiscent of other texts (AV = 5 and 4.5, respectively) and were more likely to be recommended or given as presents (AV = 4, SD = 1.4). In contrast, MF 1–3 scored poorly in these areas, with MF 3 consistently receiving the lowest ratings (AV = 1, SD = 0). Overall, the ChatGPT-3.5 microfictions (MF 4–6) achieved higher total scores (AV = 3.4, SD = 0.8) compared to those generated by Monterroso (MF 1–3, AV = 2.2, SD = 1.1).

Table 9. Literary experts responses to Monterroso and ChatGPT-3.5 microfictions- AV - SD.

Literary experts responses to Microfictions from Monterroso and ChatGPT-3.5														
	MF 1		MF 2		MF 3		MF 1		MF 5		MF 6		Average	
Question	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
Story Overview														
3.-Does it propose other interpretations, in addition to the literal one?	3	2.8	3	2.8	1	0	4	1.4	4	1.4	3.5	2.1	3.1	1.8
Technical														
5.-Is the story credible?	5	0	3	2.8	2	1.4	5	0	5	0	5	0	4.2	0.7
6.-Does the text require your participation or cooperation to complete its form and meaning?	5	0	5	0	4	1.4	3.5	0.7	5	0	3.5	0.7	4.3	0.5
7.-Does it propose a new vision of reality?	2	1.4	2	1.4	1	0	2	1.4	2	1.4	2.5	0.7	1.9	1.1
8.-Does it propose a new vision of the genre it uses?	2	1.4	1.5	0.7	1	0	1	0	2	1.4	2	0	1.6	0.6
9.-Does it propose a new vision of the language itself?	2	1.4	1	0	1	0	1	0	1	0	1	0	1.2	1.2
Editorial / commercial														
10.-Does it remind you of another text or book you have read?	4	1.4	2	1.4	1	0	5	0	4	1.4	4.5	0.7	3.4	0.8
11.-Would you like to read more texts like this?	3	1.4	2	1.4	1	0	4	1.4	3	2.8	4	1.4	2.8	1.6
12.-Would you recommend it?	3	2.8	1	0	1	0	4	1.4	3	2.8	4	1.4	2.7	1.4
13.-Would you give it as a present?	1	0	1	0	1	0	4	1.4	3	2.8	4	1.4	2.3	0.9

Table 10. Literary experts responses to MFs generated by Monterroso and ChatGPT-3-5 by GrAImes section.

	Story Overview			Technical			Editorial/commercial			Microfictions Total Analysis		
#	MF	AV	SD	MF	AV	SD	MF	AV	SD	MF	AV	SD
1	4	4	1	1	3.2	0.8	4	4.3	1.1	4	3.4	0.8
2	5	4	1	5	3	0.6	6	4.1	1.2	6	3.4	0.8
3	6	3.5	2.1	6	2.8	0.3	5	3.3	2.5	5	3.2	1.4
4	1	3	2.8	4	2.5	0.4	1	2.8	1.8	1	3	1.4
5	2	3	2.8	2	2.5	1	2	1.5	0.7	2	2.2	1.1
6	3	1	0	3	1.8	0.6	3	1	0	3	1.4	0.3

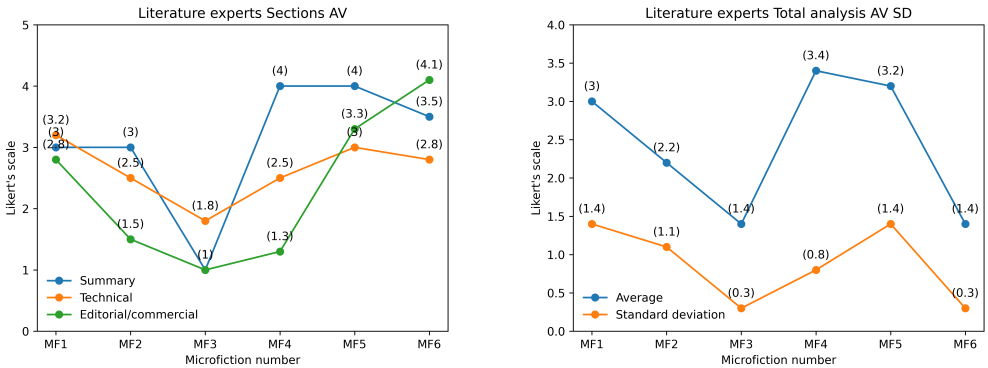

Figure 7. Line charts of literary experts GrAImes sections summarized AV and SD.

 doi:10.20944/preprints202504.1851.v1

One of the most notable results in this evaluation concerns the interpretative engagement of readers. The highest-rated question, *"Does the text require your participation or cooperation to complete its form and meaning?"*, received an average score (**AV**) of 4.3 with a standard deviation (**SD**) of 0.5. This suggests that the evaluated texts demand significant reader interaction, a crucial trait of literary complexity (see Table 11).

Conversely, aspects related to innovation in language and genre were rated lower. The question *"Does it propose a new vision of the language itself?"* obtained an AV of 1.2, the lowest among all items, with an SD of 1.2, indicating high variability in responses. Similarly, the question *"Does it propose a new vision of the genre it uses?"* received an AV of 1.6 and an SD of 0.6, further emphasizing the experts' perception that the generated texts do not significantly redefine literary conventions.

Regarding textual credibility, the question *"Is the story credible?"* was rated highly, with an AV of 4.2 and an SD of 0.7. This suggests that the narratives effectively maintain believability, an essential criterion for reader immersion. Additionally, evaluators were asked whether the texts reminded them of other literary works, yielding an AV of 3.4 and an SD of 0.8, indicating a moderate level of intertextuality.

GrAImes also examined subjective aspects of reader appreciation. The questions *"Would you recommend it?"* and *"Would you like to read more texts like this?"* received AV scores of 2.7 and 2.8, respectively, with higher SD values (1.4 and 1.6), reflecting diverse opinions among experts. Similarly, the willingness to offer the text as a gift scored an AV of 2.3 with an SD of 0.9, suggesting a moderate level of appreciation but not strong endorsement.

Table 11. Literary experts' responses to the microfictions generated by Monterroso and ChatGPT-3.5 in ascending SD order.

Literary experts responses to MF's generated by Monterroso and ChatGPT-3.5, ordered by SD		
Question	AV	SD
6.-Does the text require your participation or cooperation to complete its form and meaning?	4.3	0.5
8.-Does it propose a new vision of the genre it uses?	1.6	0.6
5.-Is the story credible?	4.2	0.7
10.-Does it remind you of another text or book you have read?	3.4	0.8
13.-Would you give it as a present?	2.3	0.9
7.-Does it propose a new vision of reality?	1.9	1.1
9.-Does it propose a new vision of the language itself?	1.2	1.2
12.-Would you recommend it?	2.7	1.4
11.-Would you like to read more texts like this?	2.8	1.6
3.-Does it propose other interpretations, in addition to the literal one?	3.1	1.8

4. Discussion

In our study the evaluators were literary experts and literature enthusiasts because the evaluation of AI-generated and human written literary texts differs significantly depending on the expertise and reading habits of the evaluators. Literature scholars and dedicated literary enthusiasts possess a familiarity with narrative structures, stylistic devices, and literary traditions, enabling them to assess texts with a critical and informed perspective. They are more likely to recognize intertextual references, thematic depth, and the subtleties of language that contribute to literary quality. In contrast, evaluations conducted by random readers—most of whom engage with literature only occasionally—tend to focus on immediate readability, entertainment value, and emotional impact rather than on formal or aesthetic complexity. While this broader audience provides valuable insights into general reception and accessibility, their assessments may lack the depth needed to critically engage with intricate literary techniques. This divergence underscores the importance of distinguishing between different evaluator groups when developing assessment methodologies for AI-generated texts. A balanced evaluation framework should account for both perspectives, ensuring that AI literature

 doi:10.20944/preprints202504.1851.v1

is judged not only on its mass appeal but also on its adherence to, or innovation within, established literary traditions.

The evaluation of AI-generated literary texts poses significant challenges due to the inherently subjective nature of aesthetic judgment. Traditional assessment frameworks in computational linguistics often rely on automated metrics such as perplexity, coherence, and sentiment analysis. However, these metrics fail to capture the nuanced and context-dependent qualities that define literary excellence. As a result, reader-based evaluation has emerged as a crucial methodological approach, leveraging human perception to assess the artistic and stylistic value of AI-generated narratives. The variability in responses between random readers and literary experts highlights the necessity of a structured framework that integrates both perspectives to ensure a more comprehensive and reliable evaluation process.

A key aspect of aesthetic evaluation is the distinction between general audience reception and expert literary critique. While non-expert readers provide insights into accessibility, engagement, and emotional resonance, experts apply specialized knowledge of literary traditions, narrative structures, and stylistic innovation. This distinction becomes particularly relevant in the assessment of AI-generated texts, where algorithmic authorship often lacks intentionality and depth in meaning construction. Consequently, the presence of literary experts in the evaluation design process is not only beneficial but essential for identifying higher-order textual attributes such as intertextuality, originality, and thematic complexity.

Given these considerations, a hybrid evaluation model of GrAImes that integrates both expert critique and broader audience participation is likely to yield the most balanced assessment of AI-generated literature. The inclusion of expert evaluators ensures that the texts are measured against established literary standards, while the involvement of general readers provides valuable feedback on accessibility and reader engagement. This dual approach underscores the need for interdisciplinary collaboration between computational linguists and literary scholars in the design of methodologies for AI and human written text evaluation.

Expert consensus on the protocol's capacity to evaluate microfiction's literary value was predominantly positive, with 80% of reviewers affirming its effectiveness. However, the presence of a dissenting perspective underscores the importance of continuous methodological refinement. The minor reservations primarily centered on the precision of specific evaluation criteria, indicating a nuanced approach to protocol development is required.

The evaluation protocol encountered significant methodological challenges, particularly regarding criterion ambiguity. Key issues included interpretative inconsistencies in assessing linguistic creativity, intertextual references, and expressive quality. Experts specifically highlighted problematic areas such as the subjective interpretation of "novel language" and the complex evaluation of literary influences. The recommended methodological improvements include replacing vague, recall-based assessments with more explicit, structured inquiries about literary lineage and explicit criteria for measuring expressive quality.

Both groups found ChatGPT-3.5's MFs more commercially appealing, while Monterroso's texts showed slightly better technical execution (literary enthusiasts) or reader engagement (experts). These findings highlight that while AI-generated microfictions can compete with human-authored ones in commercial and structural aspects, they still fall short in innovative and deeply interpretive literary qualities. The study underscores the potential of AI in replicating certain narrative techniques while emphasizing the enduring challenges in achieving true creative originality. Experts focused on depth and originality, while literary enthusiasts prioritized readability and appeal, reinforcing that AI-generated texts may satisfy casual readers more than experts in literature.

5. Conclusions

This study introduces an evaluation protocol designed to assess human written and AI-generated microfictions, with a key innovation being the integration of literary theory and expert input from

 doi:10.20944/preprints202504.1851.v1

literary scholars in the protocol's development. By grounding the assessment framework in established literary principles, the methodology ensures a rigorous and theoretically informed analysis of narrative quality, stylistic coherence, and creative depth. This approach enhances the protocol's adaptability, allowing for its application not only to microfictions but also to a broader range of literary genres. The inclusion of domain-specific expertise strengthens the validity of the evaluations, ensuring that human written microfictions and AI-generated texts are assessed against meaningful artistic and structural criteria rather than superficial metrics. Consequently, this protocol offers a scalable and genre-flexible tool for advancing the critical study of computational creativity in literature.

The evaluation of human-written microfictions using GrAImes revealed a significant agreement among literary experts, who assigned higher Likert scale ratings to texts authored by more experienced writers compared to beginners, suggesting that experiential proficiency influences perceived literary quality. The evaluators confirmed the validity of the assessment protocol while proposing refinements to the questionnaire design, emphasizing the need for more focused and clearly articulated questions to enhance precision in future studies. This finding underscores the potential GrAImes in distinguishing nuanced differences in writing expertise while highlighting the importance of methodological clarity in subjective literary assessments.

The structured evaluation in GrAImes exposed critical distinctions between expert and literary enthusiast assessments. Literary enthusiasts prioritized accessibility and enjoyment, leading to higher commercial scores for ChatGPT-3.5, whereas experts focused on technical execution and originality, criticizing both AIs for lacking innovation. Training data played a decisive role: ChatGPT-3.5's high-data training enhanced coherence and interpretative flexibility, while Monterroso's limited dataset led to inconsistent performance—occasionally high engagement but frequent flaws. Standard deviation (SD) analysis further clarified evaluator consensus; experts showed strong agreement (e.g., MF 3, SD = 0) in identifying failures, while literary enthusiasts displayed moderate SD ($\approx$ 1.3–1.4), indicating structured but varied preferences.

By employing measurable criteria, GrAImes quantified strengths/weaknesses of different AI systems, revealed divergences between expert and literary enthusiast priorities, and identified systemic gaps (e.g., lack of innovation) in AI-generated literature. Future refinements could incorporate open-ended questions for qualitative insights, but the current framework successfully benchmarks AI performance in microfiction, demonstrating that evaluation design critically shapes interpretations of AI creativity.

In our forthcoming experiment, we will implement the proposed evaluation protocol to assess diverse microfiction texts produced by chatbots. The evaluation will be conducted through a multi-perspective approach involving literary experts, avid readers, and the chatbots themselves in a self-assessment capacity. By incorporating expert judgments, we aim to capture nuanced literary qualities such as stylistic coherence and narrative depth, while reader evaluations will provide insights into accessibility and engagement. The inclusion of chatbot self-assessment offers a novel dimension, allowing for comparative analysis between human and machine-based critiques. This tripartite evaluation framework seeks to validate the protocol's robustness while exploring potential discrepancies between human and AI evaluative standards, thereby advancing the study of computational literary analysis.

Author Contributions: Conceptualization, J.G.F., R.M., G.A.M., N.P., N.C.A., and Y.G.M.; methodology, R.M. and J.G.F.; software, G.A.M.; validation, G.A.M., R.M. and J.G.F.; formal analysis, R.M. and J.G.F.; investigation, G.A.M. and J.G.F.; resources, R.M. and J.G.F.; data curation, G.A.M. and J.G.F.; writing—original draft preparation, G.A.M.; writing—review and editing, R.M., J.G.F., G.A.M., N.P., N.C.A. and Y.G.M.; visualization, G.A.M.; supervision, R.M. and J.G.F.; project administration, R.M.; funding acquisition, R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ECOS NORD grant number 321105.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

 doi:10.20944/preprints202504.1851.v1

Data Availability Statement: All dataset analyzed during this study can be found at https://github.com/Manzanarez/GrAImes.

Acknowledgments: We are grateful to the writers, literary experts and enthusiast readers who participated in this research: Miguel Tapia, Florance Olivier, Oswaldo Zavala, Marcos Eymar, Alejandro Lambarry, Abraham Truxillo, Abril Albarran, Adriana Azucena Rodriguez, David Nava, Lupita Mejia Alvarez, Sandra Huerta, Maria Elisa Leyva Gonzalez, Maria Mendoza, Luis Roberto, Diana Leticia Portillo Rodriguez, Angelica, Iris, Elisa, Fernanda, Guadalupe Monserrat Ramirez Santin, Valeria, Janik Rojas, Brenda, Alma Sanchez.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* **2025**.
2. Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720* **2024**.
3. Team, G.; Georgiev, P.; Lei, V.I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* **2024**.
4. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* **2015**.
5. Leslie, D.; Ashurst, C.; González, N.M.; Griffiths, F.; Jayadeva, S.; Jorgensen, M.; Katell, M.; Krishna, S.; Kwiatkowski, D.; Martins, C.I.; et al. 'Frontier AI,'Power, and the Public Interest: Who benefits, who decides? *Harvard Data Science Review* **2024**.
6. Porter, B.; Machery, E. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* **2024**, *14*, 26133.
7. Clark, E.; Ji, Y.; Smith, N.A. Neural text generation in stories using entity representations as context. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2250–2260.
8. Jakesch, M.; Hancock, J.T.; Naaman, M. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2208839120.
9. Alhussain, A.I.; Azmi, A.M. Automatic story generation: a survey of approaches. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–38.
10. Iser, W. The act of reading: A theory of aesthetic response. *Journal of Aesthetics and Art Criticism* **1979**, *38*.
11. Ingarden, R. Concretización y reconstrucción. *En busca del texto: teoría de la recepción literaria* **1993**, pp. 31–54.
12. Ryan, J. Grimes' Fairy Tales: A 1960s Story Generator. In Proceedings of the Interactive Storytelling; Nunes, N.; Oakley, I.; Nisi, V., Eds., Cham, 2017; pp. 89–103.
13. Propp, V.Y. *The Russian Folktale by Vladimir Yakovlevich Propp*; Wayne State University Press, 2012.
14. Ginna, P. *What editors do: the art, craft, and business of book editing*; University of Chicago Press, 2017.
15. Peinado, F.; Gervás, P. Evaluation of automatic generation of basic stories. *New Generation Computing* **2006**, *24*, 289–302.
16. Boden, M.A. *The creative mind: Myths and mechanisms*; Routledge, 2004.
17. Tomassini, G.; Maris, S. La minificción como clase textual transgenérica. *Revista interamericana de bibliografía: Review of interamerican bibliography* **1996**, *46*, 6–6.
18. Medina, Y.d.J.G. Microrrelato o minificción: de la nomenclatura a la estructura de un género literario. *Microtextualidades. Revista Internacional de microrrelato y minificción* **2017**, pp. 89–102.
19. Ricoeur, P. La función narrativa. *Revista de Semiótica* **1989**, *1*, 69–90.
20. Barthes, R.; Alcalde, R. *La aventura semiológica*; Paidós Barcelona, 1990.
21. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* **2018**.
22. Guan, J.; Mao, X.; Fan, C.; Liu, Z.; Ding, W.; Huang, M. Long text generation by modeling sentence-level and discourse-level coherence. *arXiv preprint arXiv:2105.08963* **2021**.
23. Min, K.; Dang, M.; Moon, H. Deep Learning-Based Short Story Generation for an Image Using the Encoder-Decoder Structure. *IEEE Access* **2021**, *9*, 113550–113557.
24. Lo, K.L.; Ariss, R.; Kurz, P. GPoeT-2: A GPT-2 Based Poem Generator. *arXiv preprint arXiv:2205.08847* **2022**.
25. Cavazza, M.; Charles, F.; Mead, S.J. Character-based interactive storytelling. *IEEE Intelligent systems* **2002**, *17*, 17–24.

 doi:10.20944/preprints202504.1851.v1

26. Gervás, P.; Díaz-Agudo, B.; Peinado, F.; Hervás, R. Story plot generation based on CBR. *Journal of Knowledge-Based Systems* **2005**, *18*, 235–242.
27. Mori, Y.; Yamane, H.; Mukuta, Y.; Harada, T. Toward a Better Story End: Collecting Human Evaluation with Reasons. In Proceedings of the Proceedings of the 12th International Conference on Natural Language Generation, 2019, pp. 383–390.
28. Rishes, E.; Lukin, S.M.; Elson, D.K.; Walker, M.A. Generating different story tellings from semantic representations of narrative. In Proceedings of the International Conference on Interactive Digital Storytelling. Springer, 2013, pp. 192–204.
29. Elson, D.K.; McKeown, K. A tool for deep semantic encoding of narrative texts. In Proceedings of the A tool for deep semantic encoding of narrative texts, 2009.
30. Sutskever, I.; Martens, J.; Hinton, G.E. Generating text with recurrent neural networks. In Proceedings of the ICML, 2011.
31. Kiddon, C.; Zettlemoyer, L.; Choi, Y. Globally coherent text generation with neural checklist models. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 329–339.
32. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.
33. Walker, M.A.; Grant, R.; Sawyer, J.; Lin, G.I.; Wardrip-Fruin, N.; Buell, M. Perceived or not perceived: Film character models for expressive nlg. In Proceedings of the International Conference on Interactive Digital Storytelling. Springer, 2011, pp. 109–121.
34. Sharp, O.; Goodwin, R. Sunspring: A Sci-Fi Short Film Starring Thomas Middleditch. https://www.youtube.com/watch, 2016.
35. Lukin, S.M.; Reed, L.I.; Walker, M.A. Generating sentence planning variations for story telling. *arXiv preprint arXiv:1708.08580* **2017**.
36. Jelinek, F.; Mercer, R.L.; Bahl, L.R.; Baker, J.K. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* **1977**, *62*, S63–S63.
37. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
38. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 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), 2019, pp. 4171–4186.
40. Pedersen, T.; Patwardhan, S.; Michelizzi, J.; et al. WordNet:: Similarity-Measuring the Relatedness of Concepts. In Proceedings of the AAAI, 2004, Vol. 4, pp. 25–29.
41. Levenshtein, V.I.; et al. Binary codes capable of correcting deletions, insertions, and reversals. In Proceedings of the Soviet physics doklady. Soviet Union, 1966, Vol. 10, pp. 707–710.
42. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J.; et al. Learning internal representations by error propagation, 1985.
43. Shannon, C.E. Prediction and entropy of printed English. *Bell system technical journal* **1951**, *30*, 50–64.
44. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
45. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* **1972**, *28*, 11–21.
46. Genette, G. *Figures I*; Vol. 1, Points, 1976.
47. Bal, M.; Van Boheemen, C. *Narratology: Introduction to the theory of narrative*; University of Toronto Press, 2009.
48. Gardent, C.; Perez-Beltrachini, L. A statistical, grammar-based approach to microplanning. *Computational Linguistics* **2017**, *43*, 1–30.
49. Bringsjord, S.; Ferrucci, D. *Artificial intelligence and literary creativity: Inside the mind of brutus, a storytelling machine*; Psychology Press, 1999.

 doi:10.20944/preprints202504.1851.v1

50. Oberlander, J.; Lascarides, A. Preventing false temporal implicatures: Interactive defaults for text generation. In Proceedings of the COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics, 1992.
51. Bonnie Dorr and Terry Gaasterland. Summarization-inspired temporal-relation extraction: tense-pair templates and treebank-3 analysis. , 2007.
52. Lavoie, B.; Rambow, O. A Fast and Portable Realizer for Text Generation. In Proceedings of the Proc. ANLP'97, 1997, pp. 265–268. https://doi.org/10.3115/974557.974596.
53. Zhu, J. Towards a Mixed Evaluation Approach for Computational Narrative Systems. In Proceedings of the Proc. ICCC'12, 2012, pp. 150–154.
54. Fisher, R.A. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1928**, *121*, 654–673.
55. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *psychometrika* **1951**, *16*, 297–334.
56. Ludmer, J.; et al. Clases 1985: Algunos problemas de teoría literaria **2015**.
57. Austin, J. Austin JL-How to Do Things With Words.pdf, 1962.
58. Bertochi, D. La aproximación al texto literario en la enseñanza obligatoria. *Textos de didáctica de la lengua y la literatura* **1995**, pp. 23–38.
59. Huamán, M.Á. Educación y literatura. *Lima: Mantaro* **2003**.
60. de Lizaur Guerra, M.B. La telenovela mexicana: forma y contenido de un formato narrativo de ficcíon de alcance mayoritario. PhD thesis, La autora, 2002.
61. Thompson, J.B. *Merchants of culture: the publishing business in the twenty-first century*; John Wiley & Sons, 2013.
62. Calasso, R. *La marca del editor*; Anagrama, 2014.
63. Bremond, C.; Cancalon, E.D. The logic of narrative possibilities. *New Literary History* **1980**, *11*, 387–411.
64. Shklovsky, V.; et al. Art as technique. *Literary theory: An anthology* **1917**, *3*.
65. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in neural information processing systems, 2017, pp. 5998–6008.
66. Oñate Latorre, A.; Ortiz Fuentes, J. GPT2-spanish. *https://huggingface.co/DeepESP/gpt2-spanish*, 2018.
67. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* **2019**, pp. arXiv–1910.
68. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* **2020**.
69. McCutchen, D. From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of writing research* **2011**, *3*, 51–68.
70. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019.
71. Rahutomo, F.; Kitasuka, T.; Aritsugi, M.; et al. Semantic cosine similarity. In Proceedings of the The 7th international student conference on advanced science and technology ICAST. University of Seoul South Korea, 2012, Vol. 4, p. 1.