Evaluating the Creativity of LLMs in Persian Literary Text Generation

Armin Tourajmehr*[1] Mohammad Reza Modarres*[1] Yadollah Yaghoobzadeh[2,1]
[1]Tehran Institute for Advanced Studies, Khatam University, Iran
[2]School of Electrical and Computer Engineering,
College of Engineering, University of Tehran, Tehran, Iran
a.tourajmehr@teias.institute mr.modarres@teias.institute
y.yaghoobzadeh@ut.ac.ir

Abstract

Large language models (LLMs) have demonstrated notable creative abilities in generating literary texts, including poetry and short stories. However, prior research has primarily centered on English, with limited exploration of non-English literary traditions and without standardized methods for assessing creativity. In this paper, we evaluate the capacity of LLMs to generate Persian literary text enriched with culturally relevant expressions. We build a dataset of user-generated Persian literary spanning 20 diverse topics and assess model outputs along four creativity dimensions—originality, fluency, flexibility, and elaboration—by adapting the Torrance Tests of Creative Thinking. To reduce evaluation costs, we adopt an LLM as a judge for automated scoring and validate its reliability against human judgments using intraclass correlation coefficients, observing strong agreement. In addition, we analyze the models' ability to understand and employ four core literary devices: simile, metaphor, hyperbole, and antithesis. Our results highlight both the strengths and limitations of LLMs in Persian literary text generation, underscoring the need for further refinement.[1]

1 Introduction

As LLMs continue to evolve and gain widespread use, there has been growing interest in their potential to perform tasks that require creativity. A prominent application is creative writing, where LLMs are increasingly employed to generate stories, poetry, and other literary forms. Yet debate persists over whether these models can genuinely emulate or replace human writers in producing creative text (Gervais and Shariff, 2024). A common criticism is that LLMs struggle with creativity, particularly in generating original, high-quality, and culturally nuanced outputs (Boussioux et al., 2024; Chakrabarty et al., 2023; Gómez-Rodríguez and Williams, 2023).

Most existing research on creative text generation has focused on English and English-speaking contexts. Consequently, the creative capabilities of LLMs in other languages—especially low-resource ones such as Persian—remain largely underexplored. Despite the significance of literary text generation as a distinct form of creative expression, to our knowledge no study has systematically evaluated LLM-generated literary texts in Persian. Moreover, prior work—whether in Persian or English—rarely considers the challenge of evaluating culturally grounded literary texts produced by native speakers, beyond limited domains such as story writing and poetry. This gap underscores the need for evaluating how well models align with the cultural and literary practices of human communities.

Evaluating creativity in LLMs presents unique challenges due to their distinct reasoning processes, the subjective nature of creativity, and the limitations of manual evaluation. A widely used framework for assessing human creativity is the Torrance Tests of Creative Thinking (TTCT) (Torrance, 1966), which evaluate four core dimensions: originality, fluency, flexibility, and elaboration. Current benchmarks such as the Alternative Uses Task (AUT) (Stevenson et al., 2022; Summers-Stay et al., 2023) capture divergent thinking but fall short in addressing the cultural depth and stylistic richness required for literary creativity.

Building on the work of Zhao et al. (2024), who adapted the TTCT for evaluating general-purpose creativity in LLMs, we extend this approach to Persian literary text generation—a domain that poses unique linguistic and cultural challenges. Unlike prior studies that focus on open-ended prompts, our framework emphasizes the generation of stylistically rich and culturally grounded Persian sentences. To this end, we introduce a culturally

*Equal contribution
[1]The dataset, code, and evaluation guide are available at github

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14762–14774
November 4-9, 2025

adapted evaluation framework based on the four TTCT dimensions. Moreover, in the absence of suitable resources, we compile and release *CPers* (Creativity in Persian)—the first dataset designed for this purpose—which provides a foundation for systematic creativity evaluation in low- and mid-resource languages such as Persian.

Evaluating six state-of-the-art LLMs—including GPT-3.5, GPT-4.1, DeepSeek-V3, DeepSeek-R1, Qwen2.5, and Gemma—we conduct one of the first systematic analyses of Persian literary creativity in LLM outputs. To ensure reliable scoring, we combine human annotations with an LLM-as-a-judge framework using Claude 3.7 Sonnet, which demonstrates strong alignment with human evaluators. Beyond creativity evaluation, we conduct two complementary studies: first, we examine word usage across topics to assess each model's cultural alignment with native Persian speakers and values; second, we analyze the presence of four literary devices frequently used in Persian literature—simile, metaphor, antithesis, and hyperbole—to explore how stylistic elements relate to creativity.

Our work makes the following key contributions:

- We present the first systematic evaluation of LLMs on Persian literary text generation, adapting the Torrance Tests of Creative Thinking (TTCT) to assess originality, fluency, flexibility, and elaboration in a culturally grounded context.
- We introduce *CPers*, a novel dataset of 4,371 Persian literary texts spanning 20 emotionally and culturally diverse topics, authored by native speakers, along with a human-annotated subset of 200 texts that includes creativity scores and labels for rhetorical devices. This dataset provides a benchmark for evaluating creativity in low-resource languages.
- We conduct a comprehensive evaluation of six state-of-the-art LLMs, combining human annotations and automated scoring via Claude 3.7 Sonnet, and analyze their performance across multiple dimensions of creativity.
- We investigate the use of key literary devices—simile, metaphor, antithesis, and hyperbole—as well as lexical patterns, revealing how the balance and nuanced deployment of these devices influence perceived creativity and cultural alignment.
- Our analysis provides insights into model design and training strategies, showing, for example, that reasoning-oriented models produce more elaborated and flexible literary texts, and highlighting the importance of multi-dimensional, culturally aware evaluation for creative text generation.

2 Related Work

Creative writing is a cognitively complex and performative language task that requires linguistic fluency, cultural and literary competence, narrative coherence, and the capacity for originality and imagination. Recent work has increasingly explored the use of LLMs in creative domains, including humor generation (Zhong et al., 2023), comedy creation (Mirowski et al., 2024), and psychological creativity assessments (Bellemare-Pepin et al., 2024). Studies show that LLMs can produce poetic and narrative content of high quality (Franceschelli and Musolesi, 2023), and that human judges often struggle to distinguish between human-written and model-generated stories (Clark et al., 2021).

Creativity, however, remains difficult to evaluate due to its subjective nature. Common assessment methods include the Divergent Association Task (DAT) (Olson et al., 2021), the Remote Associates Test (RAT) (Mednick, 1962), and the widely used Torrance Tests of Creative Thinking (TTCT) (Torrance, 1966) in psychometric studies.

Stevenson et al. (2022) show that human creative outputs outperform those of GPT-3 on the Alternative Uses Task. Summers-Stay et al. (2023) further demonstrate that while GPT-3 could generate original ideas, it often failed to filter out impractical ones. Naeini et al. (2023) introduce the OnlyConnect Wall dataset to simulate RAT-like tasks for evaluating creative problem solving in LLMs. Their findings reveal that red herrings reduce model performance, though their analysis does not incorporate advanced prompting or retrieval-augmented methods, leaving room for further exploration. Similarly, Atmakuru et al. (2024) propose the CS4 benchmark to assess creativity under varying prompt specificity, promoting originality over memorization. Unlike CS4's focus on general storytelling, our work specifically targets literary creativity, with emphasis on style, emotion, and cultural depth.

Recent narrative-level analyses (Tian et al., 2024) show that LLMs systematically generate

Criteria	Questions
Originality	- Is the sentence creative and far from common clichés? - Is the sentence similar to famous sentences, poems, or Persian proverbs? - Does the sentence contain at least one of the literary devices of simile, metaphor, antithesis, or hyperbole?
Fluency	- Is the sentence grammatically correct and understandable? - Does the sentence seem fluent and natural to a Persian reader? - Can the sentence be used in a literary text or everyday conversation?
Flexibility	- Does the sentence use multiple ideas to express the intended topic? - Does the sentence look at the topic from a new perspective? - Does the sentence use different styles (e.g., ironic, humorous, philosophical)?
Elaboration	- Does the sentence go into detail and use a variety of vocabulary? - Does the sentence create a clear mental image in the reader? - Does the sentence convey a specific feeling (e.g., love, sadness, hope) well?

Table 1: A creativity assessment framework for Persian texts, grounded in originality, fluency, flexibility, and elaboration.

stories that are more predictable and positive, while struggling with managing climaxes and emotional arcs. Structural approaches such as marking turning points can improve narrative quality, but a substantial gap with human writing remains. Other studies indicate that, although LLMs perform strongly in linguistic fluency, they still lag behind humans in novelty, diversity, and surprise (Ismayilzada et al., 2025). Collaborative generation with multiple models can enhance diversity and creativity, but often at the cost of coherence (Venkatraman et al., 2025).

Zhao et al. (2024) present a scalable benchmark for evaluating LLM creativity using a modified version of the TTCT and automated GPT-4 scoring on general creative tasks. While their work demonstrates the feasibility of large-scale creativity testing in English, it does not account for cultural or linguistic differences that are central to literary creativity.

In contrast, our study is the first to evaluate LLM-generated literary creativity in Persian, a culturally rich yet underrepresented language. We adapt the TTCT to reflect stylistic, emotional, and metaphorical aspects characteristic of Persian literature. This not only fills an important gap in cross-lingual creativity evaluation but also offers a framework for studying literary creativity in diverse cultural traditions, with potential extensions to other languages.

3 Methodology

We propose a framework for evaluating the creativity of LLMs in Persian literary text generation. Our evaluation builds on the Torrance Tests of Creative Thinking (TTCT) by focusing on four dimensions: originality, fluency, flexibility, and elaboration. We construct a dataset of Persian literary texts and employ both human annotators and LLM-based reviewers.

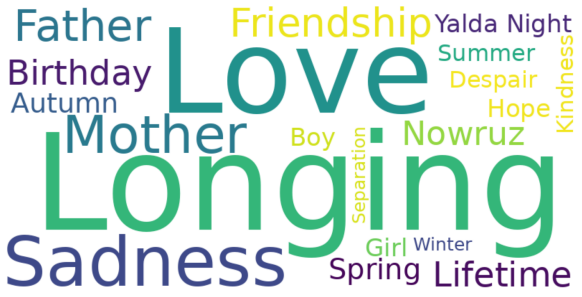


Figure 1: Word cloud illustrating the distribution of themes in the CPers dataset. The most frequent theme appears 670 times, while the least frequent occurs 50 times.

3.1 CPers Dataset

To conduct this study, access to a dataset specifically tailored to Persian literary texts is essential. However, no publicly available dataset rooted in Persian-speaking culture exists. We therefore create a new resource, *CPers*, by collecting texts from various online sources. The final dataset comprises 4,371 texts spanning 20 distinct topics. Authored by everyday people, these writings capture a wide range of human emotions and relationships. Each text averages approximately 26 words in length. The distribution of topics is balanced, with no category exceeding 15% or falling below 1% of the corpus (see Figure 1). This balance ensures diverse coverage of cultural and emotional themes. Additional details on dataset construction and a full list of topics are provided in Appendix A. Representative samples are shown in Figure 2.

The topics include universal themes such as love, kindness, hope, disappointment, friendship, and

Topic	Sentence
عشق Love	از بس که دوستت دارم فکر می‌کنم؛ دیگر هیچ دوست داشتنی همرنگ دوست داشتن های من نیست. I love you so much that I believe no other love resembles the way I love you.
دلتنگی Longing	هیچگاه خط دلت را مشغول نکن، شاید دلتنگی پشت خط باشه، دلتنگم. Never keep your heart line busy, maybe longing is calling, I'm missing you.
پدر Father	پدر یعنی جاده‌های زندگی را با شجاعت هموار کردن. Father means paving the roads of life with courage.
نوروز New Year (Nowruz)	بهار یک نقطه دارد نقطه آغاز بهار زندگیتان بی‌انتها باد سال نو مبارک. Spring begins with a single point, the point of a new beginning. May the spring of your life be endless, Happy New Year.

Figure 2: Sample entries from the *CPers* dataset, showing Persian literary sentences and their emotional or cultural topics, along with English translations. Note that Nowruz, the Persian New Year, coincides with the first day of spring.

sadness, as well as culturally significant occasions such as Nowruz (Persian New Year), Father's Day, and Mother's Day. Although most texts were produced by non-professional writers or inspired by classical literary figures, efforts were made to preserve literary richness in the majority of cases. During dataset construction, we prioritized texts that incorporated at least one rhetorical device—such as simile, metaphor, antithesis, or hyperbole—so that the collection would reflect not only everyday language use but also the stylistic depth characteristic of Persian literary expression.

3.2 Evaluation Metric

To assess the creativity of texts, whether human- or LLM-generated, we develop a new evaluation framework inspired by the TTCT and specifically adapted to the Persian-speaking cultural context. The framework is organized around four key dimensions—originality, fluency, flexibility, and elaboration. Each dimension is assessed through three culturally tailored questions, yielding a total of 12 questions (see Table 1). Responses are rated on a five-point scale, where 1 indicates the lowest and 5 the highest score. This structure enables a systematic and culturally relevant assessment of creative text generation.

A central adaptation concerns the notion of *fluency*. In the original TTCT, fluency is often measured quantitatively as ideational fluency—the number of distinct ideas produced. This metric, however, is not directly applicable to short Persian texts, particularly those generated by LLMs. Creative ideas in Persian are frequently conveyed implicitly or metaphorically within a single sentence, making idea-counting both ambiguous and culturally biased. To address this, we redefine fluency to evaluate the grammatical accuracy of the text, its naturalness for native speakers, and its appropriateness for either literary or conversational contexts. This redefinition aligns more closely with how fluency is perceived in Persian creative writing. The questions are developed through iterative refinement, informed by pilot annotation sessions and error analyses of both human-written and model-generated Persian texts.

Text type	Originality	Fluency	Flexibility	Elaboration
Model text	0.70	0.78	0.76	0.69
Human text	0.67	0.42	0.45	0.69

Table 2: ICC scores of Human-1 with Human-2 on *Model texts* and *Human texts* (p-value $<< 0.05$ for all dimensions).

The decision to adapt the TTCT for Persian arises from both linguistic and cultural considerations. These adaptations are necessary to ensure construct validity and cultural fairness in evaluating creativity. Although our implementation is tailored for Persian, the overall structure and methodology are general and can be extended to other languages and cultural contexts with appropriate modifications.

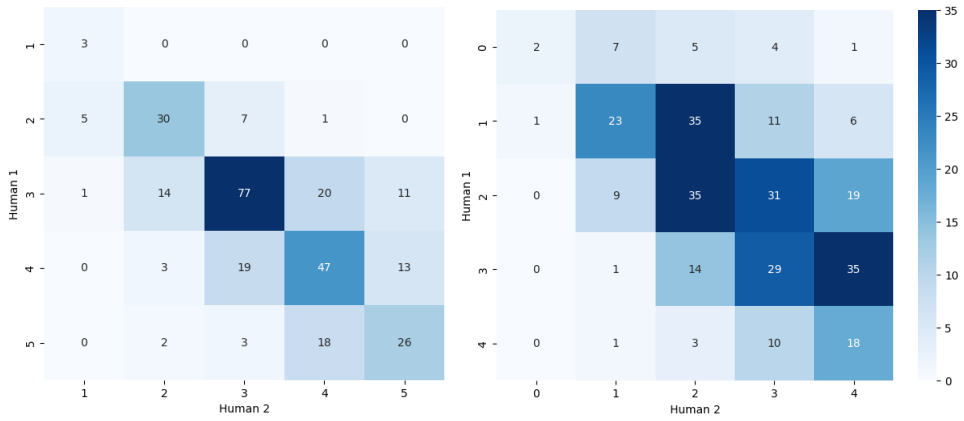
3.3 Human Annotated Dataset

One hundred instances are selected from the *CPers* dataset, referred to as *Human texts*. These texts cover five topics—*love*, *longing*, *friendship*, *hope*, and *despair*—representing a balanced range of human emotions. Using GPT-3.5 (OpenAI, 2023) as the base model, an additional 100 literary texts are generated across the same five topics via zero-shot prompting (as described in 4.1), referred to as *Model texts*, ensuring alignment with the human-written topics.

To establish a ground truth for creativity evaluation, two human annotators assessed both 100 *Human texts* and 100 *Model texts* using the proposed 12-question framework covering the four key dimensions. For each text, an overall creativity score was computed by averaging the scores across these dimensions.

To ensure consistency among annotators, calibration meetings were conducted using a demo dataset prior to the main annotation task. Inter-rater reliability was assessed by examining the variation in scores assigned by different annotators. The confusion matrix illustrating agreement on the originality criterion is shown in Figure 3[2]. As observed, in

[2]For results on other criteria, see Appendix B.



(a) (b)

Figure 3: Confusion matrices showing inter-rater agreement between two human annotators for the originality criterion on: (a) Model texts, and (b) Human texts, within the proposed framework.

Model	Originality	Fluency	Flexibility	Elaboration
GPT-4o	0.57	0.63	0.67	0.30
Claude 3.7 Sonnet	0.46	0.69	0.55	0.54

(a) ICC of the average scores of Human-1 and Human-2 on *Model texts* (p-value $<<$ 0.05 for all dimensions).

Model	Originality	Fluency	Flexibility	Elaboration
GPT-4o	0.64	-0.26	0.61	0.58
Claude 3.7 Sonnet	0.65	0.39	0.46	0.59

(b) ICC of the average scores of Human-1 and Human-2 on *Human texts* (p-value $<<$ 0.05 for all dimensions).

Table 3: ICC of the average scores of Human-1 and Human-2 with models across four TTCT-based dimensions.

most cases, ratings differed by no more than one point.

To quantify agreement, we employ the Intraclass Correlation Coefficient (ICC), a statistical measure that evaluates the consistency of observations within groups. Unlike Pearson correlation, which only captures linear relationships, ICC assesses the closeness of scores, making it particularly suitable in our context where most inter-rater differences fall within a single point on the 1–5 scale. ICC therefore provides a more appropriate measure of agreement for two raters independently scoring each sentence. The ICC values for *Human texts* and *Model texts* are presented in Table 2. These results indicate strong and consistent agreement across both text types and all creativity dimensions, reflecting the inherently subjective nature of evaluating creative writing (Gómez-Rodríguez and Williams, 2023). Notably, as shown in Table 2, annotators exhibit slightly higher agreement on model-generated texts across all evaluation criteria.

3.4 LLM as Judge

To establish a robust and consistent framework for evaluating the creativity of LLM-generated texts, we compare the judgment behavior of two general-purpose language models—GPT-4o (OpenAI, 2024) and Claude 3.7 Sonnet (Anthropic, 2025). The primary criterion for selecting a model as a judge is its alignment with human evaluators, since creativity—particularly in literary and culturally nuanced contexts—is inherently subjective and challenging to assess automatically.

As reported in Table 2, ICC scores between the two human annotators are consistently high across all four TTCT dimensions, indicating strong inter-annotator reliability. We therefore use the average of their scores as the human gold standard and compute ICC values between these averages and the ratings produced by each LLM.

Quantitative results indicate that Claude 3.7 Sonnet exhibits stronger alignment with human annotations compared to GPT-4o. As shown in Tables 3a and 3b, Claude's creativity ratings across the four TTCT dimensions—originality, fluency, flexibility, and elaboration—closely match those of human annotators for both *Human texts* and *Model texts*. Claude achieves high average ICC scores with both human raters across all criteria and text types, whereas GPT-4o shows lower average correlations and even a negative correlation in one instance.

We also observe that GPT-4o tends to assign higher scores to *Model texts* than to *Human texts*, introducing a bias that complicates the assessment of literary creativity and limits deeper analysis of stylistic qualities.

Given its stronger alignment with human judg-

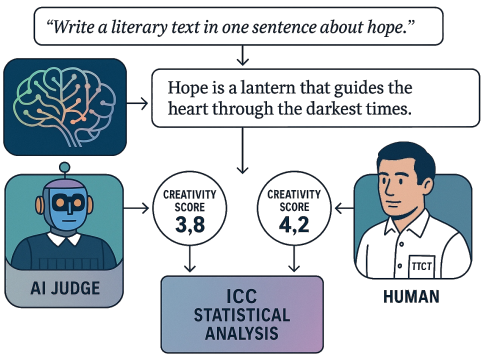


Figure 4: The overall flew of the evaluation framework and selection of the judge model.

ments and its ability to capture culturally relevant stylistic nuances, we select Claude 3.7 Sonnet as the final LLM judge. This choice ensures that the automated evaluation framework remains consistent with human intuition and culturally grounded criteria. The overall evaluation process is summarized in Figure 4.

3.5 Prompting Strategy

To generate model outputs for evaluation, we adopt a consistent prompting strategy across all systems. Each model is instructed to produce 100 single-sentence literary texts for each of five themes—*love*, *longing*, *friendship*, *hope*, and *despair*—corresponding to the topics selected for human annotations (Section 3.3). To simulate spontaneous human-like text production, we use a zero-shot prompt in Persian of the form: *"Write a literary text in one sentence about {Topic}"*.

This prompt reflects how native speakers intuitively generate literary expressions: concise, topic-driven, and stylistically rich. All outputs are generated with a temperature setting of 1 to encourage creativity and variability while maintaining coherence. The resulting texts are then evaluated using the framework described in Section 3.4.

Also in rare cases where a model produces duplicate sentences for different prompts (observed for DeepSeek-R1), we replace the repeated outputs with alternative, non-repetitive generations from the same model to preserve diversity in the evaluation.

4 Experiments

Here, we evaluate the creativity of Persian literary texts generated by various LLMs. We compare model performance, analyze the use of key rhetorical devices—specifically simile, metaphor, antithesis, and hyperbole—and assess the extent to which generated texts conform to Persian literary norms.

Model	Originality	Fluency	Flexibility	Elaboration	Creativity
Gemma 3	0.035	0.006	0.044	0.023	0.010
Deepseek V3	0.050	0.030	0.044	0.012	0.015
GPT-4.1	0.031	0.006	0.017	0.015	0.006
Qwen2.5	0.060	0.053	0.047	0.046	0.012
GPT-3.5	0.021	0.065	0.078	0.062	0.021
Deepseek-R1	0.095	0.020	0.139	0.046	0.072

Table 4: Standard deviations of scores for different models across evaluation criteria.

4.1 Setup

We use six LLMs to generate creative Persian literary texts: GPT-3.5-Turbo (OpenAI, 2023), GPT-4.1 (OpenAI, 2025), DeepSeek-V3-0324-671B (DeepSeek, 2025b), Gemma-3-27B-Instruct (Gemma, 2025), Qwen2.5-VL-32B-Instruct (Qwen, 2025), and Deepseek-R1-671B (DeepSeek, 2025a). These models were selected for their strong performance in generative and instruction-following tasks, representing a mix of proprietary and open-source systems with varying capabilities in multilingual and creative text generation.

4.2 Comparing LLMs

To evaluate the creative potential of language models in a culturally grounded context, we assess their ability to generate Persian literary texts. The evaluation focuses on four core creativity dimensions applied to the texts (100 per model) generated in response to Persian literary prompts.

Each experiment is repeated three times to ensure reliability, and the reported values correspond to the average scores across runs. In addition, we compute the standard deviation of scores across the three runs, which provides a measure of stability for each model's performance (see Table 4). The generally low standard deviations indicate that the evaluation is consistent and robust across repeated trials.

Ratings are provided by the Claude 3.7 Sonnet model on a scale of 1 to 5 per dimension, and averages are calculated, including an overall creativity score (mean of all four dimensions), as shown in Figure 5.

DeepSeek-R1 achieves the highest overall creativity score among the evaluated models. Its

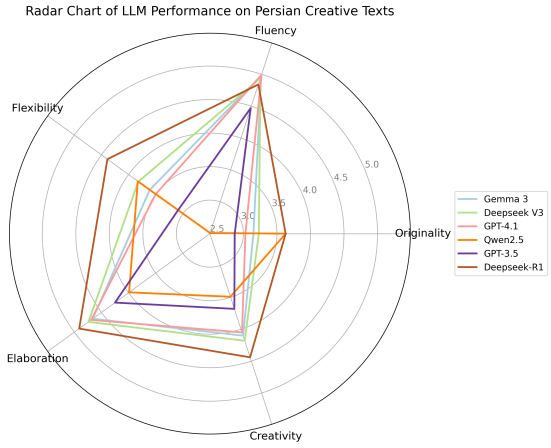


Figure 5: This figure compares creativity assessment scores across models. DeepSeek-R1 achieves the highest overall creativity score (4.44) and leads in flexibility (39) and elaboration (4.91). GPT-4.1 scores highest in fluency (4.99), while Qwen2.5-VL-32B-Instruct shows strong originality (3.63) but records the lowest fluency score (2.51), reflecting distinct performance patterns across different creativity dimensions.

strong performance in elaboration and flexibility indicates the model's ability to express diverse ideas, adopt multiple perspectives, expand on details, and evoke vivid sensory imagery. However, DeepSeek-R1 occasionally generates repetitive outputs. For instance, the sentence "Love is the silent song of two hearts that, at the distance of a glance, breathe eternity in one breath" appears three times on different prompts. To maintain diversity, we replace repeated outputs with other non-repetitive generations from the model.

One possible explanation for DeepSeek-R1's strong performance, particularly in elaboration and flexibility, is its reasoning-oriented training. The model is trained via reinforcement learning with objectives that promote structured thinking, including a "think first, answer later" approach (DeepSeek, 2025a), which may implicitly support more detailed and diverse content generation.

DeepSeek-V3-0324 also demonstrates strong performance, particularly in elaboration and flexibility, suggesting that models in the DeepSeek family are capable of producing stylistically rich and imaginative literary texts with varied perspectives. Its low standard deviations across dimensions indicate stable and reliable performance.

GPT-4.1 outperforms its predecessor, GPT-3.5, particularly in fluency and readability. Its outputs are grammatically correct and natural for Persian readers. However, it receives relatively low scores

Model	Sentence
Qwen-2.5-VL-32b-instruct	نارنج‌های یکسره و خورشیدی که از افق درآمد، امیدی بود که ناپدید می‌شد. The constant oranges and the sun that rose from the horizon were a hope that was vanishing.
Deepseek- R1	ناامیدی، سایه ی سنگین غروبی است که در سکوت بادهای خزان، آواز آخرین برگ های امید را در خاک روح به خاکستر می کشاند. Despair is the heavy shadow of a sunset that, in the silence of autumn winds, turns the song of the last leaves of hope into ashes within the soil of the soul.
GPT4.1	پدر یعنی جاده های زندگی را با شجاعت هموار کردن. Father means paving the roads of life with courage.

Figure 6: Sample sentences on despair from Qwen-2.5, DeepSeek-R1, and GPT-4.1, along with their English translations.

in originality, indicating that while its texts are coherent, they often rely on conventional expressions and lack inventive use of literary devices—though it still shows notable improvement over GPT-3.5. Compared to DeepSeek models, GPT-4.1 exhibits more consistent behavior, with smaller variance across runs.

In contrast, Qwen2.5-VL-32B-Instruct achieves higher originality scores but performs poorly in fluency. Its outputs are more novel and less clichéd, yet occasionally lack clarity and readability for Persian speakers. Additionally, the model scores lower in elaboration; while it introduces unique ideas, it struggles to create vivid imagery or convey emotional depth (e.g., love, sadness). The relatively larger standard deviations for Qwen2.5 in originality and fluency confirm this variability in creative performance.

These findings underscore the importance of evaluating creative text generation across multiple dimensions. Selecting models based on specific creativity criteria is essential for literary applications that require both stylistic authenticity and cultural nuance.

5 Analysis

5.1 Word Frequency Analysis

While Table 5 does not reveal strong stylistic distinctions across all models, it suggests possible lexical similarities within model families. GPT-3.5 and GPT-4.1, for example, often rely on similar metaphorical terms such as hope, light, darkness, and heart. This may indicate that, despite architectural improvements, GPT-4.1 inherits certain lexical tendencies from GPT-3.5. The frequent use of binary oppositions like light/dark could reflect a preference for familiar, easily retrievable metaphors. Gemma-3-27B-IT exhibits a similar pattern, frequently reusing common symbolic contrasts, which may suggest a shared limitation in

Text Source	1st Word	2nd Word	3rd Word	4th Word	5th Word
Human	Hope (14)	Having (5)	Life (5)	Sky (3)	Gaze (3)
GPT-3.5	Hope (24)	Heart (16)	Light (7)	Darkness (5)	Bright (4)
GPT-4.1	Hope (20)	Night (15)	Heart (11)	Dark (8)	Bright (6)
Gemma-3-27B-Instruct	Hope (20)	Night (19)	Dark (17)	Light (17)	Star (12)
Qwen2.5-VL-32B-Instruct	Hope (23)	Light (9)	Heart (8)	Darkness (6)	Life (5)
QwQ-32B	Hope (20)	Darkness (7)	Light (6)	Black (4)	Sky (4)
DeepSeek-V3-0324	Hope (20)	Darkness (12)	Night (11)	Bird (9)	Sound (6)
DeepSeek-R1	Hope (19)	Dark (11)	Night (9)	Sunrise (8)	Sound (7)

Table 5: Top 5 most frequent words generated on the theme of hope across different text sources (frequency in parentheses).

stylistic exploration.

A comparable trend is observed between DeepSeek-R1 and DeepSeek-V3-0324, which often use overlapping terms such as dark, night, and sound. These similarities may arise from shared training data, decoding strategies, or model architecture. While these terms are not inherently uncreative, their repeated use suggests that both models draw from a similar pool of literary expressions. In contrast, human-written texts display more varied and grounded imagery—e.g., sky and gaze—reflecting a more intuitive and emotionally nuanced approach to expression. These observations suggest that human creativity, even in short texts, tends to involve subtler and more diverse lexical choices than current LLMs typically produce.

In this context, reasoning-oriented LLMs appear to positively influence the outputs. Specifically, the frequency of common and repetitive words decreases, while more creative and human-like terms emerge. For instance, in DeepSeek-R1, the word "sunrise" appears more frequently—a term that is more vivid and imaginative compared to simpler descriptors like "dark" or "light".

To further explore this effect, we also evaluated another reasoning-enhanced model, QWQ-32B, an improved version of Qwen. This model similarly shows an increased frequency of the word "sky", a term commonly found in human-generated texts. In both cases, the frequency of highly repeated words, such as "hope" and "dark", decreases relative to their corresponding base models (i.e., models from the same family without reasoning enhancements). Detailed analyses of word relationships, text similarity,and creativity metrics are left for future work.

5.2 Figure of Speech Analysis

To evaluate the stylistic richness and creative capacity of generated texts, we analyzed the use of four common figures of speech—simile, metaphor, antithesis, and hyperbole—comparing human-written texts with outputs from the six LLMs, as shown in Figure 7.

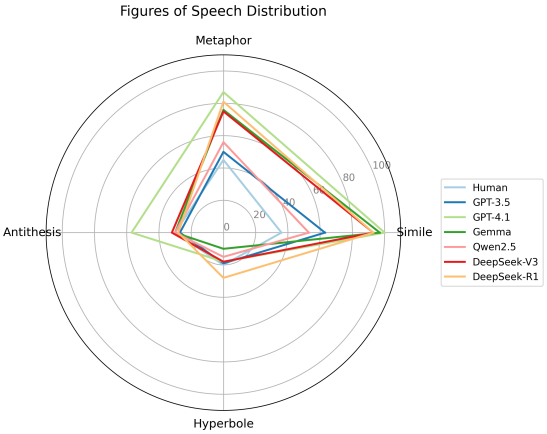


Figure 7: Count of figures of speech (simile, metaphor, antithesis, hyperbole) used in human-written and model-generated texts across different models.

Two human annotators independently labeled the rhetorical devices, achieving inter-annotator agreement rates of 80% for *Human texts* and 84% for *Model texts*. We initially intended to use LLMs as judges in the same manner as for text creativity. However, both Claude and GPT-4o exhibited inconsistencies in identifying figures of speech, particularly similes and metaphors. These models often confused the two devices and did not demonstrate a clear understanding of their distinctions. Consequently, we relied on human annotators to accurately identify the rhetorical devices in each text.

Human-written texts display a balanced use of literary devices—primarily metaphors and similes—reflecting natural stylistic coherence. Among the models, GPT-4.1 produces the highest number of rhetorical devices overall, with similes appearing in all 100 of its generated texts. This overreliance may partly explain its lower originality score. In contrast, Qwen2.5-VL-32B-Instruct demonstrates a more selective and varied use of rhetorical devices, aligning with its higher originality rating. This suggests that a deliberate and inventive deployment of literary elements can better enhance perceived creativity.

DeepSeek-R1, while employing metaphors and similes extensively, also shows a more balanced application of antithesis and hyperbole. Its effective use of hyperbole, in particular, likely contributes to its strong elaboration scores by enhancing vivid imagery and mental representation.

Overall, our analysis indicates that the sheer quantity of stylistic devices alone does not guarantee creativity. Rather, the type, balance, and nuanced application of figures of speech play a crucial role in shaping the originality, fluency, flexibility, and elaboration of generated texts.

6 Conclusion

In this work, we introduce a novel framework for evaluating creativity in LLMs within the context of Persian literary text generation. Building on the TTCT, we propose a culturally adapted evaluation scheme that captures four core dimensions—originality, fluency, flexibility, and elaboration—which can be applied to any language. Our analysis shows that no single model performs well across all dimensions of creativity. Instead, different models exhibit varying strengths in different aspects: for example, DeepSeek-R1 and GPT-4.1 demonstrate high expressive richness, whereas others, such as Qwen2.5, generate more concise yet culturally resonant outputs. This highlights that creative ability in LLMs is distributed unevenly across dimensions rather than concentrated in a single model.

A key outcome of our analysis is that models tend to follow learned patterns rather than demonstrating genuinely diverse creativity. Their use of rhetorical devices skews heavily toward simile and metaphor, with limited balance across other devices such as antithesis and hyperbole. Human texts, in contrast, often integrate multiple figures of speech within a single sentence, resulting in richer and more creative expression. Moreover, when we attempt to use LLMs as judges for labeling rhetorical devices, they struggle—particularly in distinguishing between metaphor and simile—highlighting current limitations in nuanced literary understanding.

We also curate a unique dataset, CPers, comprising 4,371 single-sentence literary texts spanning diverse topics and emotions. This dataset is the first of its kind in Persian, and, to the best of our knowledge, no comparable dataset exists in English. We also release our human-annotated subset, including 100 human-written and 600 model-generated texts, with annotations covering both creativity scores and the figures of speech employed in each text.

Taken together, our findings indicate that LLMs serve as useful tools for Persian literary text generation, but expectations remain modest: their creativity does not yet parallel human-level diversity and literary nuance. These results emphasize the need for culturally grounded evaluation in multilingual NLP, particularly for low-resource, high-context languages.

Future work explores increasing the number of annotators, improving the judge model, examining more diverse topics, testing diverse prompting strategies, and conducting cross-lingual comparisons to assess the adaptability of LLM creativity across cultures.

Limitations

While our framework provides a structured and culturally grounded approach to evaluating creativity in Persian literary text generation, it is not without limitations. The evaluation questions were designed and scored by the authors, who—while fluent in the language and familiar with literary conventions—are not formally trained in psychology or Persian literary studies. Future work could benefit from interdisciplinary collaboration to refine both the criteria and the evaluation process.

Our analysis focused on 100 samples and five themes, offering a practical but narrow window into the broader dataset. Creativity, however, often unfolds more vividly across longer narratives and diverse emotional contexts. Evaluating paragraph-level or multi-sentence outputs may uncover richer stylistic patterns and deeper coherence that go unnoticed at the sentence level.

We also restricted our prompting to a zero-shot setup. Exploring other prompting strategies—such as few-shot, chain-of-thought, or instruction prompting—could help reveal how different models respond to varying task formulations, and whether prompt design can shape creativity in meaningful ways.

Moreover, while our focus on Persian fills a critical gap, it leaves open the question of how these models perform across languages. A cross-lingual comparison would shed light on whether the observed creative behaviors are language-dependent or model-intrinsic, and could further reveal how cultural and linguistic structure shape creative expression.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Anirudh Atmakuru, Jatin Nainani, Rohith Sid-

dhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *Preprint*, arXiv:2410.04197.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *Preprint*, arXiv:2405.13012.

Léonard Boussioux, Jacqueline N. Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani. 2024. The crowdless future? generative ai and creative problem-solving. *Organization Science*. Published online: 16 Aug 2024.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. arXiv preprint arXiv:2309.14556. To appear in ACM CHI 2024.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

DeepSeek. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://arxiv.org/abs/2501.12948. Accessed: 2025-05-09.

DeepSeek. 2025b. Deepseek-v3-0324.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Gemma. 2025. Gemma-3-27b-instruct.

Daniel Gervais and Shaheen Shariff. 2024. The creative agency of large language models: A philosophical inquiry. *AI and Ethics*.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. arXiv preprint arXiv:2310.08433.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC 2025)*. Accepted to ICCC 2025.

Sarnoff A. Mednick. 1962. The associative basis of the creative process. *Psychological Review*, 69(3):220–232.

Piotr Wojciech Mirowski, Juliette Love, Kory W. Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of llms' humour alignment with comedians. *Preprint*, arXiv:2405.20956.

Saeid Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *arXiv preprint arXiv:2306.11167*.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*.

OpenAI. 2023. Gpt-3.5 turbo.

OpenAI. 2024. Gpt-4o technical report. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-05-09.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Qwen. 2025. Qwen2.5-vl-32b-instruct.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*. Cite as: arXiv:2206.08932 [cs.AI].

Douglas Summers-Stay, Clare R. Voss, and Stephanie M. Lukin. 2023. Brainstorm, then select: A generative language model improves its creativity score. In *Proceedings of the AAAI-23 Workshop on Creative AI Across Modalities*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics. Outstanding Paper Award.

E. P. Torrance. 1966. *Torrance Tests of Creative Thinking: Directions Manual and Scoring Guide*. Personnel Press, Incorporated.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. Collabstory: Multi-llm collaborative story generation and authorship analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679, Albuquerque, New Mexico. Association for Computational Linguistics.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. 2024. Assessing and understanding creativity in large

language models. *arXiv preprint arXiv:2401.12491*. Cite as: arXiv:2401.12491 [cs.CL].

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2023. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. *Preprint*, arXiv:2312.02439.

A Dataset Construction

The *CPers* dataset, introduced in Section 3.1, contains 4,371 short literary-style Persian texts collected from a variety of online sources. These texts reflect a wide range of cultural and emotional themes and are primarily written by native Persian speakers.

Topics Covered

The dataset covers 20 culturally significant themes, including love, mother, father, longing, birthday, boy, girl, Yalda Night (an Iranian celebration marking the longest night of the year), friendship, Nowruz (the Iranian New Year), autumn, winter, spring, summer, despair, sorrow, life, separation, hope, and kindness.

Collection Process

The texts are gathered from publicly available website and blogs featuring Persian literary and emotional content. We focuse on collecting relatively short texts, typically one sentence or a few lines, suitable for sentence-level creativity evaluation. All data instances have been reviewed by humans to ensure they do not contain any personal information or offensive content.

Data Usage and Disclaimer

The data from online resources used to create the dataset is anonymized and publicly available. The *CPers* dataset is intended for research purposes.

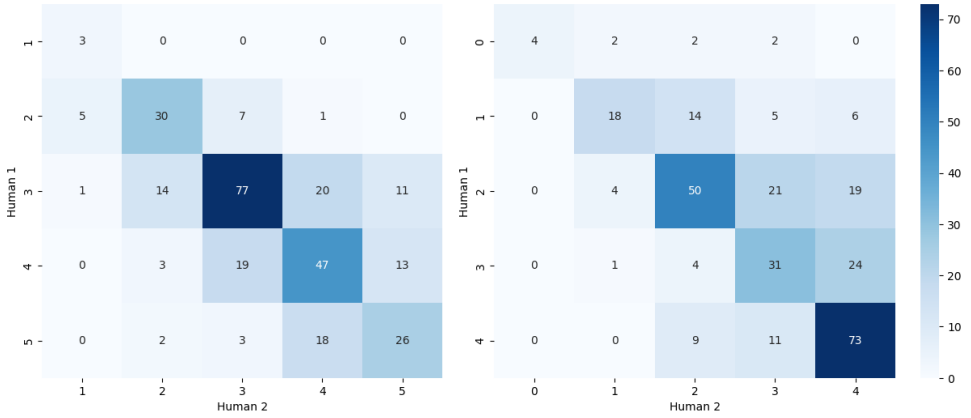
Source Attribution

Texts are sourced from a range of publicly available platforms.[3]

[3]Example sources include: https://digipostal.ir, https://www.beytoote.com, https://roozaneh.net, https://vista.ir, https://shereno.com, https://salamdonya.com, https://chishi.ir, https://www.delgarm.com, https://diamag.ir, https://fararu.com, http://www.coca.ir, https://www.alamto.com, https://setare.com, https://wikimatn.com, https://robinarose.com, https://www.tasvirezendegi.com, https://www.talab.org, https://delbaraneh.com, https://www.bishtarazyek.com, https://topnaz.com, https://magerta.ir, https://namnak.com

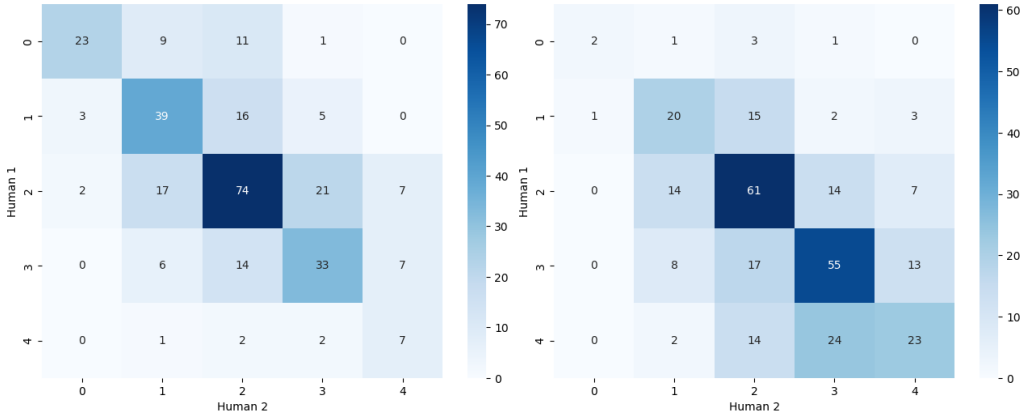
B Confusion Matrices

Confusion matrix to show agreement between human annotators across all criteria and text types are presented in Figures 8, 9, 10, and 11.



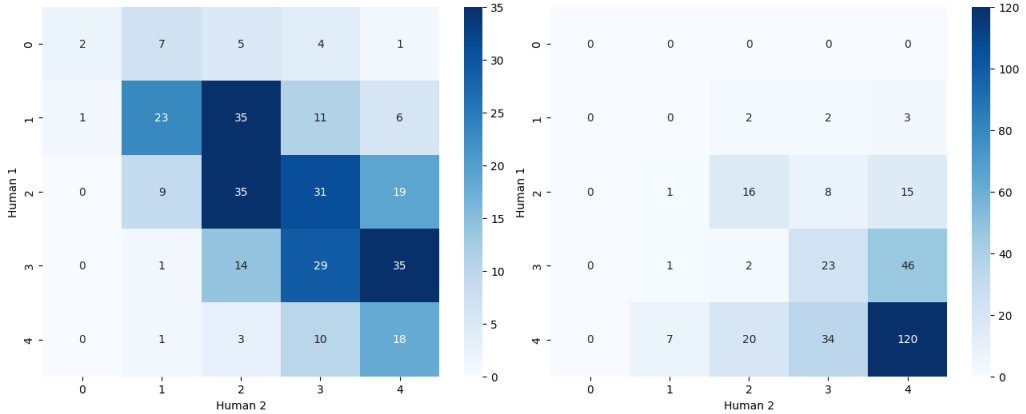
(a) (b)

Figure 8: Confusion matrices showing inter-rater agreement between humans on Model texts: (a) Originality criteria, and (b) Fluency criteria.



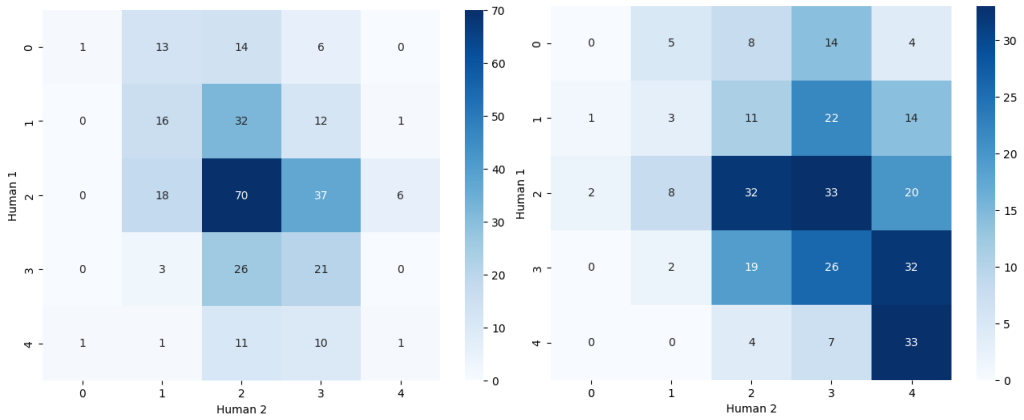
(a) (b)

Figure 9: Confusion matrices showing inter-rater agreement humans on Model texts: (a) Flexibility criteria, and (b) Elaboration criteria.



(a) (b)

Figure 10: Confusion matrices showing inter-rater agreement between humans on Human texts: (a) Originality criteria, and (b) Fluency criteria.



(a) (b)

Figure 11: Confusion matrices showing inter-rater agreement between humans on Human texts: (a) Flexibility criteria, and (b) Elaboration criteria.