Automated Creativity Evaluation for Large Language Models: A Reference-Based Approach

Ruizhe Li[1] **Chiwei Zhu**[1]* **Benfeng Xu**[1] **Xiaorui Wang**[2] **Zhendong Mao**[1]
[1]University of Science and Technology of China, Hefei, China
[2]MetastoneTechnology, Beijing, China
imlrz@mail.ustc.edu.cn tanz@mail.ustc.edu.cn

Abstract

Creative writing is a key capability of Large Language Models (LLMs), with potential applications in literature, storytelling, and various creative domains. However, evaluating the creativity of machine-generated texts remains a significant challenge, as existing methods either rely on costly manual annotations or fail to align closely with human assessments. In this paper, we propose an effective automated evaluation method based on the Torrance Test of Creative Writing (TTCW), which evaluates creativity as product. Our method employs a reference-based Likert-style approach, scoring generated creative texts relative to high-quality reference texts across various tests. Experimental results demonstrate that our method significantly improves the alignment between LLM evaluations and human assessments, achieving a pairwise accuracy of 0.75 (+15%).

1 Introduction

Creative writing is a key capability of Large Language Models (LLMs), with applications in literature, storytelling, and other creative domains (Orwig et al., 2024; Xie et al., 2023). However, studies have revealed a significant gap between the creative writing capabilities of LLMs and those of human experts (Ismayilzada et al., 2024; Chakrabarty et al., 2024). Bridging this gap requires further exploration and innovation, which in turn necessitates an effective and practical approach to evaluating the creativity of language models.

Although some studies (Stevenson et al., 2022; Summers-Stay et al., 2023; Guzik et al., 2023) have adapted creativity evaluation methods from traditional educational and psychological research—such as the Alternate Uses Task (AUT) (Guilford, 1967) and the Torrance Test of Creative Thinking (TTCT) (Torrance, 1966)—to assess LLMs, these approaches rely heavily on manual annotations. Furthermore, these methods typically evaluate creativity as a process by analyzing responses to open-ended questions designed to elicit creative thinking (Cramond, 2020), which are inherently difficult to assess automatically. Additionally, the limited number of predefined test questions introduces randomness and increases the likelihood of accidental outcomes(Zhao et al., 2024), potentially resulting in unreliable evaluations of LLM performance.

To address these challenges, evaluating creativity as a product rather than a process offers a promising alternative. For instance, Chakrabarty et al. (2024) introduced the Torrance Test of Creative Writing (TTCW), which assesses creativity based on candidates' textual outputs. This approach enhances scalability by allowing the number of test cases to increase continuously while adding the generated texts, thereby reducing randomness through averaging over larger samples. Moreover, automated evaluation of generated texts is more practical compared to subjective judgments of open-ended tasks. However, when applied with LLMs as evaluators, TTCW has not achieved satisfactory results, as reported by Chakrabarty et al. (2024).

In this paper, we aim to develop an effective automated evaluation method for assessing the creativity of LLMs using TTCW. We draw inspiration from reference-based evaluation methods commonly used in human assessments and automatic evaluations in other fields (Zhang et al., 2020; Yuan et al., 2021), and propose an approach which assigns a relative score to the generated texts compared to high-quality reference texts. Additionally, we adopt Likert-style scoring, a widely used method in psychological assessments, to rate subjective qualities like creativity (Roy, 2020).

Our main contributions are:

- Proposing a novel reference-based, Likert-scale comparative evaluation method tailored for creative text assessment;

* Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21475–21488

- Empirically demonstrating that this approach improves alignment with expert human judgments, achieving a pairwise accuracy of 0.75;
- Showing that the proposed framework is robust across different datasets and model families without requiring additional fine-tuning;
- Demonstrating that LLMs can effectively perform creativity evaluation when combined with appropriate reference-based prompting.

2 Related Work

2.1 Creativity Evaluation

In prior work, divergent thinking is widely recognized as a fundamental indicator of creativity in both research and educational settings (Baer, 1993). It is typically assessed through open-ended tasks that prompt individuals to generate creative responses. Most widely used methods for evaluating creativity are based on divergent thinking. For example, the Alternate Uses Task (AUT) (Guilford, 1967) asks participants to generate as many novel and unconventional uses as possible for a common object (e.g., a box) within a constrained time period. Similarly, the Torrance Test of Creative Thinking (TTCT)(Torrance, 1966) assesses creativity through responses to novel and unusual scenarios, relying on divergent thinking principles. Our research follows this tradition by grounding creativity evaluation in divergent thinking. Specifically, we adopt the Torrance Test of Creative Writing (TTCW) (Chakrabarty et al., 2024), a variant of TTCT, to evaluate the creativity of LLM-generated texts.

2.2 Evaluating creativity of large language models

In recent years, efforts have been made to evaluate the creativity of LLMs. (Stevenson et al., 2022) and (Guzik et al., 2023) directly apply the Alternate Uses Task (AUT) and the Torrance Test of Creative Thinking (TTCT), respectively. However, both approaches rely heavily on manual annotations, which limit scalability and consistency. Other studies have investigated automated evaluation methods. For example, (Beaty and Johnson, 2021) demonstrated that latent semantic distance is a reliable and strong predictor of human creativity ratings in the AUT. (Zhao et al., 2024) utilizes GPT-4 to generate TTCT-inspired datasets and employs the model itself to evaluate responses. (Chakrabarty et al., 2024) proposes the Torrance Test of Creative Writing (TTCW) and applies it with LLMs as judges though did not yield satisfactory outcomes.

3 Methodology

3.1 Problem Setting

The task of evaluating the creativity of language models is defined as assessing the quality of their generated texts in response to specific prompts. Specifically, plots extracted from human-authored reference stories are used as prompts for the models to generate corresponding stories. The dataset used in this study adopts stories from The New Yorker as the references (Chakrabarty et al., 2024). The process can be denoted as:

$$\text{plot}_i = \text{LLM}_{\text{extract}}(\text{reference}_i)$$

$$\text{candidate}_i^k = \text{LLM}_k(\text{plot}_i)$$

where the reference is a high-quality human-authored story, and LLM_k represents the model being evaluated.

3.2 Reference-based Evaluation

Traditional automated evaluation exhibits two shortcomings: (i) LLM judges display sycophancy, inflating scores and reducing discriminative power (Perez et al., 2023; Sharma et al., 2023; Chen et al., 2024); and (ii) in reference-free settings, ratings are tied to model-specific scales rather than shared standards. We therefore employ a reference-based protocol in which the LLM compares each candidate with high-quality references while the candidate's identity is blinded. This design attenuates uniformly high scoring and anchors judgments to concrete quality targets, improving calibration and discriminative power.

In this evaluation framework, we adopt the Torrance Test of Creative Writing (TTCW) (Chakrabarty et al., 2024), which includes 14 binary tests designed to assess creativity across four dimensions: Fluency, Flexibility, Originality, and Elaboration (see A.3 for details). For each test, the LLM compares the candidate text against the reference text using a Likert scale with five levels: "significantly better" (+2), "slightly better" (+1), "the same" (0), "slightly worse" (-1), and "significantly worse" (-2). To minimize positional bias, the sequence of the candidate and reference texts is alternated, and each test is conducted twice. A test is considered passed (i.e., the test is labeled

Method	Model	AVG Spearman	AVG Kendall's Tau	Pairwise Accuracy
Baseline	claude35	0.14	0.13	**0.64**
	gpt-4o	0.16	0.14	**0.64**
	qwen2-72b-chat	-0.12	-0.11	0.58
Ours	claude35(ours)	**0.49(+0.35)**	**0.44(+0.31)**	**0.75(+0.11)**
	gpt-4o(ours)	0.38(+0.22)	0.36(+0.22)	0.72(+0.08)
	qwen2-72b-chat(ours)	0.22(+0.34)	0.16(+0.27)	0.61(+0.03)

Table 1: Comparison of Baseline and Proposed Methods Across Different Models. The table presents the performance of baseline and proposed methods on three metrics: AVG Spearman, AVG Kendall's Tau, and Pairwise Accuracy. The bolded values in the "Baseline" section represent the highest scores among baseline models. The "Ours" section highlights significant improvements achieved by the proposed method, with changes relative to the baseline shown in parentheses.

as "True") if the average score across two assessments is higher than the cutoff score. The overall creativity score of a candidate text is calculated as the total number of tests passed out of the 14 binary tests.

The process is formally represented as:

$$\mathrm{L}_{i,j}^{k,+} = \mathrm{LLM}_{\text{evaluator}}(\text{test}_j, \text{reference}_i, \text{candidate}_i^k)$$

$$\mathrm{L}_{i,j}^{k,-} = \mathrm{LLM}_{\text{evaluator}}(\text{test}_j, \text{candidate}_i^k, \text{reference}_i)$$

$$\text{Score}_i^k = \sum_j I[(\mathrm{L}_{i,j}^{k,+} - \mathrm{L}_{i,j}^{k,-}) \geq \text{score}_{cutoff}]$$

where $\mathrm{L}_{i,k}^{k,+}$ is the label reflecting the extent to which the candidate$_i^k$ is better than the reference$_i$, and $\mathrm{L}_{i,k}^{k,-}$ represents the opposite.The $score_{cutoff}$ is a threshold used to convert Likert-scale scores into binary labels, determining whether a candidate passes a given test. A detailed discussion on how the $score_{cutoff}$ is determined and optimized can be found in Discussion Section 5.1.

3.3 Prompt Strategy

Previous research has demonstrated that the analyze-rate strategy can improve performance in evaluation tasks when applied with GPT models (Chiang and yi Lee, 2023). This strategy is similar to zero-shot Chain-of-Thought (CoT) reasoning, but specifically adapted for evaluation tasks. Instead of directly assigning a rating, the model is first prompted to analyze the sample according to the evaluation criteria before providing a final score. In our experiments with different models, We observe the same improvement. Therefore, we adopt this strategy in our final prompt framework, which is detailed in Appendix A.1.

4 Experiment

4.1 Datasets and Setup

We evaluate our proposed framework on two datasets to assess both its effectiveness and generalizability.

TTCW Dataset We first conduct experiments on the dataset from Chakrabarty et al. (2024), which includes human annotations assessing the creative quality of 12 original stories from The New Yorker alongside corresponding LLM-generated stories produced by GPT-3.5, GPT-4, and Claude V1.3. This dataset serves as our main benchmark, where we tune hyperparameters and evaluate alignment with expert preferences. Dataset statistics and details are provided in Appendix A.2.

GRW Dataset To test the robustness and generalization of our method, we further evaluate it on the dataset from Gómez-Rodríguez and Williams (2023), which similarly provides expert assessments of stories generated by different language models from shared storylines. We reuse the same hyperparameters from the TTCW experiments without any retuning, allowing us to assess whether our framework generalizes effectively beyond the original setting. It is worth noting that the dimensions used to evaluate creativity in this dataset differ from those in the TTCW rubric. Accordingly, our evaluations on this dataset are conducted using its native set of dimensions. A complete list of these dimensions is provided in Appendix A.11. In addition, since the dataset does not include human-expert-authored reference stories, we designate the GPT-4-generated texts—which consistently receive the highest expert ratings—as reference texts in our evaluation.

Model Configuration For all text generation tasks, we used consistent hyperparameters across all models: temperature=0.8 and top_p=0.8. Detailed model versions and additional technical specifications are provided in Appendix A.6.

4.2 Evaluation Protocol

For both datasets, we compare model-generated story rankings—produced by our method—with human rankings. We use the following metrics to quantify alignment:Spearman's correlation(Spearman, 1904), Kendall's tau(Kendall, 1938), and pairwise accuracy, which is calculated as the proportion of correctly aligned pairwise comparisons between model rankings and human rankings.

As a baseline, we evaluate each story by directly prompting the evaluator LLM to determine whether it satisfies the specific requirements of each TTCW test item. If the model affirms that the story meets a given criterion, the test is considered passed. The final score is likewise computed as the total number of passed tests.

4.3 Main Results

As shown in Table 1, our method improves ranking accuracy across various models on the TTCW dataset, achieving a highest pairwise accuracy of 0.75 (+15%) and yielding substantial gains in both Spearman and Kendall's tau correlations. Detailed results for the TTCW dataset are provided in Appendix A.7.

On the GRW dataset, our method maintains strong performance without any hyperparameter adjustment, further demonstrating its robustness and generalizability across domains and evaluation settings (see Table 2). It achieves consistent improvements across all three metrics, with a pairwise accuracy of 0.78 and notable increases in correlation metrics as well.

Metric	Baseline	Ours
AVG Spearman	0.53	0.68 (+0.15)
AVG Kendall's Tau	0.43	0.57 (+0.14)
Pairwise Accuracy	0.72	0.78 (+0.06)

Table 2: Performance comparison between the baseline and our proposed method on the GRW dataset, evaluated using qwen2-72b-chat. The table reports average values for three metrics: Spearman's correlation, Kendall's tau, and pairwise accuracy. Improvements over the baseline are shown in parentheses.

5 Analysis and Ablation Studies

5.1 Cutoff Score

To determine the optimal cutoff score, we conducted a hyperparameter search on the TTCW dataset, as detailed in Appendix A.8. The results indicate that setting the cutoff at -2 yields the best ranking similarity, which aligns with our expectations. A cutoff of -2 corresponds to cases where the average performance across two trials is slightly worse than the reference. Given that reference texts generally exceed the minimum passing standard by a significant margin, we consider candidates who perform only slightly worse than the reference to have still met the test's criteria.

5.2 Likert Scale Granularity

To further investigate the impact of Likert scale granularity on experimental results, we conducted an additional study using qwen2-72b-chat to explore different rating scales. Specifically, we evaluated the performance of 3-point, 5-point, and 7-point Likert scales to determine the optimal level of granularity for our evaluation framework. The Spearman's correlation for the 5-point scale is 0.22, outperforming both the 3-point scale (-0.07) and the 7-point scale (0.01). These findings suggest that the 5-point Likert scale is a more effective choice for our evaluation framework.

5.3 Ablation Study

To evaluate the impact of the Reference-Based Approach and the Analyze-Rate Strategy on the evaluation framework, we conducted ablation experiments by separately removing each component. In the ablation of the Reference-Based Approach, we removed the reference-based comparison, instructing the LLM to assess the candidate text solely based on its content and generate a binary label at the end of its response. In the ablation of the Analyze-Rate Strategy, we removed the analyze-rate prompting method and prompted the LLM to assign a label directly, without an explicit instruction to analyze the sample before rating.

The results, detailed in the A.9, indicate that both the Reference-Based Approach and the Analyze-Rate Strategy significantly enhance evaluation performance. Removing either component led to a decrease in ranking similarity and evaluation stability.

5.4 Reference Quality Impact

Because our method relies on reference texts, an important question is whether reference choice influences evaluation outcomes. We therefore ran experiments using different references as the evaluation standard. The results show that higher-quality references generally perform better overall, and that references whose quality is close to that of the candidates deliver the strongest discrimination. Based on these findings, we propose qualitative guidelines for selecting references to optimize evaluation: (i) use high-quality references; and (ii) prefer references whose quality is reasonably close to that of the candidates, avoiding references that are much higher or much lower in quality. A detailed analysis of reference dependency appears in Appendix A.10.

6 Conclusion

We proposed an automated evaluation framework for assessing the creativity of large language models (LLMs) using the Torrance Test of Creative Writing (TTCW). By adopting a reference-based Likert-style evaluation and an analyze-rate prompting strategy, our method improves alignment with human assessments, achieving a pairwise accuracy of 0.75 (+15%). Ablation studies highlight the complementary roles of the reference-based approach and analyze-rate prompting, while additional GRW dataset demonstrate its robustness and generalizability. These results establish a new benchmark result for automated creativity evaluation, offering a scalable alternative to manual annotation.

7 Limitations

One limitation of our method is its reliance on reference stories, which may restrict its scalability for unrestricted article-level evaluations. Additionally, our method may not be suitable when all candidate texts are far inferior to the reference, as this could result in all labels being assigned as significantly worse, making it impossible to distinguish relative rankings among candidates. Nonetheless, this approach serves as a robust framework for comparing the creative capabilities of different models, providing valuable insights into their relative performance.

8 Potential Risks

The proposed evaluation framework, while promising, carries potential risks that may impact its broader application and outcomes. One concern is amplifying biases in reference texts, which could favor certain styles or cultural norms while disadvantaging unconventional outputs. Additionally, automating creativity evaluation risks reducing human oversight, potentially overlooking nuanced, subjective aspects of creativity that machines cannot fully capture. Addressing these challenges requires careful reference selection and maintaining a balance between automated and human evaluations.

9 Acknowledgement

This research is supported by Artificial Intelligence-National Science and Technology Major Project 2023ZD0121200 and National Natural Science Foundation of China under Grant 62222212.

References

John Baer. 1993. Creativity and divergent thinking: A task-specific approach.

Roger E. Beaty and Dan R. Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2):757–780.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models. *Preprint*, arXiv:2310.05657.

B Cramond. 2020. Choosing a creativity assessment that is fit for purpose. *Assessing Creativity: A palette of possibilities*, pages 58–63.

J.P. Guilford. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14.

Erik E. Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065.

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

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *Preprint*, arXiv:2411.02316.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

William Orwig, Emma R. Edenbaum, Joshua D. Greene, and Daniel L. Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior*, 58(1):128–136.

Ethan Perez, Sam Ringer, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*.

A. Roy. 2020. *A Comprehensive Guide for Design, Collection, Analysis and Presentation of Likert and Other Rating Scale Data: Analysis of Likert Scale Data*. Amazon Digital Services LLC - KDP Print US.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Charles Spearman. 1904. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

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

Douglas Summers-Stay, Stephanie M. Lukin, and Clare R. Voss. 2023. Brainstorm, then select: a generative language model improves its creativity score.

E Paul Torrance. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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. *Preprint*, arXiv:2401.12491.

A Appendix

A.1 Prompt

In this section, we provide the prompt used to generate the evaluation results.

Please act as an experienced and impartial literary critic to evaluate the creativity of two stories. You will be provided with two stories, Story A and Story B. You will then be given a specific aspect of creative writing. Carefully read both stories and, based on the given aspect, critically analyze them for their creativity.

Think step by step, and describe your thought process using concise phrases. After providing your analysis, you must conclude by outputting only one of the following choices as your final verdict with a label:

1. Story A is significantly better: [[A»B]]
2. Story A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Story B is slightly better: [[B>A]]
5. Story B is significantly better: [[B»A]]

Example output: "A: narrative ending, ... B: poor character development, ... Therefore: [[A>B]]".

==========
Story A:

[STORYA]

==========
Story B:

[STORYB]

==========
Aspect:

[BACKGROUND]

==========

Remember, you must end your answer with one of these: [[A»B]], [[A>B]], [[A=B]], [[B>A]], [[B»A]]

A.2 Dataset Statistics

A.2.1 Word Counts for Different Models

To provide further insights into the dataset, Table 3 presents the word counts of generated stories across different models. While differences in verbosity and writing style exist, stories generated from the same storyline tend to have similar word counts, reducing the potential impact of length variations on evaluation scores.

Table 3: Word counts of generated stories for different models. The New Yorker column represents the original human-written reference texts.

Story Name	Claude	GPT-3.5	GPT-4	New Yorker
A Triangle	831	1126	1074	959
Barbara, Detroit, 1996	1245	1452	1460	1432
Beyond Nature	1245	1628	1326	1476
Certain European Movies	1304	1623	1480	1584
Keys	1370	1630	1297	1433
Listening For the Click	1463	1623	1612	1467
Maintenance, Hvidovre	1270	1992	1911	2066
Returns	1519	1726	1765	1715
The Facade Renovation That's Going Well	1332	1544	1477	1501
The Kingdom That Failed	1344	1344	1356	1525
The Last Dance with my Dad	1406	2455	1932	2233
Trash	1541	2215	2398	2350

A.2.2 TTCW Score Distribution

Figure 1 presents the distribution of TTCW test scores across different models. Each histogram represents the number of stories that passed a given number of tests, along with the corresponding average score. The dashed lines indicate the average number of tests passed by each model. This data is directly reproduced from the original work by Chakrabarty et al. (2024).

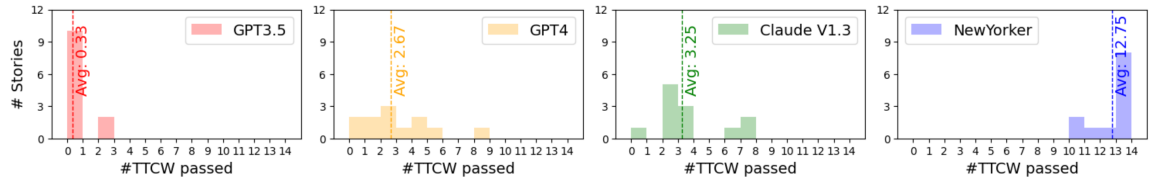

Figure 1: Distribution of TTCW test scores across different models. The dashed lines indicate the average number of tests passed by each model.

A.3 TTCW Test

This section presents the TTCW test, which outlines the dimensions and guiding questions for evaluating creativity in stories. The test includes four key dimensions: fluency, flexibility, originality, and elaboration, each accompanied by detailed background knowledge to facilitate a structured analysis. The Torrance Test of Creative Writing (TTCW) is distributed under the BSD-3-Clause license.

Table 4: TTCW Dimensions and Questions

Dimension	**Question**
Fluency	Does the end of the story feel natural and earned, as opposed to arbitrary or abrupt?
Fluency	Do the different elements of the story work together to form a unified, engaging, and satisfying whole?
Fluency	Does the story have an appropriate balance between scene and summary/exposition, or does it rely too heavily on one element?
Fluency	Does the manipulation of time (compression or stretching) feel appropriate and balanced?
Fluency	Does the story make sophisticated use of idiom, metaphor, or literary allusion?
Flexibility	Does the story achieve a good balance between interiority and exteriority, in a way that feels emotionally flexible?
Flexibility	Does the story contain turns that are both surprising and appropriate?
Flexibility	Does the story provide diverse perspectives, and if there are unlikeable characters, are their perspectives presented convincingly and accurately?
Originality	Is the story an original piece of writing without any clichés?
Originality	Does the story show originality in its form and/or structure?
Originality	Will an average reader of this story obtain a unique and original idea from reading it?
Elaboration	Are there passages in the story that involve subtext, and if so, does the subtext enrich the setting or feel forced?
Elaboration	Does the writer make the fictional world believable at the sensory level?
Elaboration	Does each character feel developed with appropriate complexity, ensuring no character exists solely for plot convenience?

A.4 Plot and Story Example

A.4.1 Example Plot Summary

A woman experiences a disorienting night in a maternity ward where she encounters other similarly disoriented new mothers, leading to an uncanny mix-up where she leaves the hospital with a baby that she realizes is not her own, yet accepts the situation with an inexplicable sense of happiness.

A.4.2 Story Excerpt (Generated by GPT-4)

There is no sound in the world quite like a baby crying. Late at night, on the chilly outskirts of Prague, in the maternity ward of an unfamiliar hospital, its howl rang through her head like the warble of a siren...

...

... In her heart, she knew that her Baby Autolycus, the product of an inexplicable and enigmatic night, was the embodiment of the ultimate connection between herself and the unknown mother, an unbreakable bond woven into their very souls.

A.5 An example of our framework

Question	**Does the story contain turns that are both surprising and appropriate?**	**Score**
High Flexibility	*She walked into the study and saw a stack of letters on the desk, the envelopes in her mother's handwriting, dated years after her mother had died...*	+2
Low Flexibility	*She walked into the study and suddenly turned into a dinosaur. Then she flew away.*	–2

Table 5: Example illustrating how the framework distinguishes high vs. low creativity on the *Flexibility* dimension. The high-flexibility text demonstrates a surprising yet appropriate turn, while the low-flexibility text introduces an arbitrary and incoherent change.

A.6 Model Specifications

For transparency and reproducibility, we report the specific model versions used in our experiments:

- **GPT-4:** gpt-4-0613 checkpoint provided by OpenAI
- **GPT-4o:** gpt-4o-20240513 version
- **Claude 3.5:** claude-3.5-sonnet-20240620 version

A.7 Full Result

	story_0	story_1	story_2	story_3	story_4	story_5	story_6	story_7	story_8	story_9	story_10	story_11	AVG_spearman
claude35_ours	0.50	0.50	1.00	0.50	1.00	-0.87	0.00	0.50	0.00	0.87	1.00	0.87	0.49
gpt-4o_ours	1.00	0.50	1.00	0.50	0.87	-0.87	0.00	0.50	-0.50	0.00	1.00	0.50	0.38
claude3-opus	1.00	0.00	0.50	0.50	1.00	-0.87	0.50	0.50	-1.00	-0.50	0.87	0.50	0.25
qwen2-72b-chat_v3_ours	0.50	0.87	0.87	0.50	0.00	-1.00	-1.00	0.87	-0.87	0.50	0.87	0.50	0.22
gpt-4o	0.87	-0.50	0.50	0.50	-0.50	0.00	0.00	0.50	0.00	-0.50	1.00	0.00	0.16
claudev13	1.00	-0.87	0.00	0.00	1.00	0.87	0.50	-0.87	1.00	0.00	0.00	-0.87	0.15
claude35	1.00	0.00	1.00	0.50	0.00	-0.50	0.00	-0.87	0.50	-0.50	0.00	0.50	0.14
gpt4	0.87	-1.00	-0.87	0.00	0.00	0.00	-0.50	0.00	0.00	-0.87	0.87	1.00	-0.04
qwen2-72b-chat	0.00	1.00	0.87	-0.87	0.00	-1.00	-1.00	0.00	-0.87	-0.50	0.87	0.00	-0.12
gemini-pro	-0.87	0.00	-0.87	-0.87	0.00	-1.00	0.00	-0.87	0.87	0.00	0.87	-1.00	-0.31
claudev21	0.50	0.00	-0.50	0.00	0.00	-0.87	-0.50	-0.87	1.00	-1.00	-0.87	-1.00	-0.34
claudev2	0.00	-0.50	0.87	-1.00	0.87	-0.87	-0.87	0.00	-0.87	-1.00	0.00	-0.87	-0.35
cgpt	0.50	-0.87	-0.87	0.87	-0.50	0.50	0.00	-1.00	-0.87	-0.87	-0.87	-0.87	-0.40

Figure 2: Complete Spearman correlation results across individual stories and models. Models labeled 'ours' indicate performance using our proposed method. The results are sorted in descending order of the average values.

	story_0	story_1	story_2	story_3	story_4	story_5	story_6	story_7	story_8	story_9	story_10	story_11	AVG_kendalltau
claude35_ours	0.33	0.50	1.00	0.33	1.00	-0.82	0.00	0.33	0.00	0.82	1.00	0.82	0.44
gpt-4o_ours	1.00	0.50	1.00	0.33	0.82	-0.82	0.00	0.33	-0.33	0.00	1.00	0.50	0.36
claude3-opus	1.00	0.00	0.33	0.33	1.00	-0.82	0.50	0.33	-1.00	-0.33	0.82	0.50	0.22
qwen2-72b-chat_v3_ours	0.33	0.82	0.82	0.33	0.00	-1.00	-1.00	0.82	-0.82	0.33	0.82	0.50	0.16
claudev13	1.00	-0.82	0.00	0.00	1.00	0.82	0.50	-0.82	1.00	0.00	0.00	-0.82	0.16
gpt-4o	0.82	-0.50	0.33	0.33	-0.33	0.00	0.00	0.33	0.00	-0.33	1.00	0.00	0.14
claude35	1.00	0.00	1.00	0.33	0.00	-0.50	0.00	-0.82	0.33	-0.33	0.00	0.50	0.13
gpt4	0.82	-1.00	-0.82	0.00	0.00	0.00	-0.50	0.00	0.00	-0.82	0.82	1.00	-0.04
qwen2-72b-chat	0.00	1.00	0.82	-0.82	0.00	-1.00	-1.00	0.00	-0.82	-0.33	0.82	0.00	-0.11
gemini-pro	-0.82	0.00	-0.82	-0.82	0.00	-1.00	0.00	-0.82	0.82	0.00	0.82	-1.00	-0.30
claudev21	0.33	0.00	-0.33	0.00	0.00	-0.82	-0.50	-0.82	1.00	-1.00	-0.82	-1.00	-0.33
claudev2	0.00	-0.50	0.82	-1.00	0.82	-0.82	-0.82	0.00	-0.82	-1.00	0.00	-0.82	-0.34
cgpt	0.33	-0.82	-0.82	0.82	-0.33	0.50	0.00	-1.00	-0.82	-0.82	-0.82	-0.82	-0.38

Figure 3: Complete Kendall's tau results across individual stories and models. Models labeled 'ours' indicate performance using our proposed method. The results are sorted in descending order of the average values.

	story_0	story_1	story_2	story_3	story_4	story_5	story_6	story_7	story_8	story_9	story_10	story_11	AVG_pairwise-ACC
claude35_ours	0.67	0.67	1.00	0.67	1.00	0.33	0.67	0.67	0.67	1.00	1.00	0.67	0.75
gpt-4o_ours	1.00	0.67	1.00	0.67	1.00	0.33	0.67	0.67	0.33	0.67	1.00	0.67	0.72
claude3-opus	1.00	0.67	0.67	0.67	1.00	0.33	1.00	0.67	0.00	0.33	0.67	0.67	0.64
claude35	1.00	0.67	1.00	0.67	0.67	0.33	0.67	0.33	0.67	0.33	0.67	0.67	0.64
gpt-4o	1.00	0.67	0.67	0.67	0.33	0.33	0.67	0.67	0.67	0.33	1.00	0.67	0.64
claudev13	1.00	0.00	0.67	0.67	1.00	0.67	0.67	0.33	1.00	0.67	0.67	0.33	0.64
qwen2-72b-chat_v3_ours	0.67	1.00	0.67	0.67	0.67	0.33	0.33	1.00	0.00	0.67	0.67	0.67	0.61
gemini-pro	0.33	0.67	0.33	0.33	0.67	0.33	1.00	0.33	1.00	1.00	1.00	0.33	0.61
gpt4	1.00	0.33	0.00	0.67	0.67	0.33	0.67	1.00	0.67	0.33	0.67	1.00	0.61
qwen2-72b-chat	0.67	1.00	1.00	0.33	0.67	0.33	0.33	0.67	0.00	0.33	1.00	0.67	0.58
claudev21	0.67	0.33	0.33	1.00	0.67	0.00	0.33	0.33	1.00	0.00	0.00	0.33	0.42
cgpt	0.67	0.00	0.33	1.00	0.33	0.67	0.33	0.00	0.33	0.33	0.33	0.00	0.36
claudev2	0.67	0.00	1.00	0.00	0.67	0.00	0.00	0.67	0.00	0.00	0.67	0.33	0.33

Figure 4: Complete Pairwise accuracy results across individual stories and models. Models labeled 'ours' indicate performance using our proposed method. The results are sorted in descending order of the average values.

A.8 Results Obtained with Different Cutoff Scores

Model	Cutoff = -3	Cutoff = -2	Cutoff = -1	Cutoff = 0
Qwen	-0.12	-0.12	-0.12	-0.12
Qwen-Ours	0.17	0.22	-0.05	-0.01
GPT-4o	0.16	0.16	0.16	0.16
GPT-4o-Ours	0.27	0.38	0.33	0.22
Claude 3.5	0.14	0.14	0.14	0.14
Claude 3.5-Ours	0.20	0.49	0.37	0.30

Table 6: Spearman correlation of different models under varying cutoff scores.

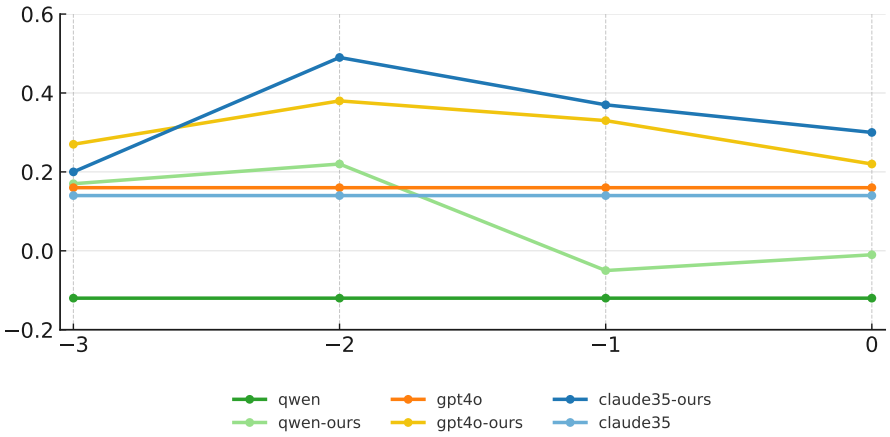

Figure 5: Spearman correlation performance under different cutoff scores.

A.9 Ablation Study

Method	Qwen2-72B-Chat	Claude 3.5
Ours	**0.22**	**0.49**
Reference-Based Approach Only	0.00	0.42
Analyze-Rate Prompting Only	0.16	0.45
Baseline	-0.12	0.14

Table 7: Ablation study results showing Spearman's correlation (ρ) for different evaluation strategies. The best performance for each model is in **bold**.

A.10 Reference Dependency Analysis

To investigate the impact of reference quality on evaluation performance, we conducted additional experiments using Qwen2-72B-Chat with different reference texts on the TTCW dataset. Table 8 shows the results when using human-expert authored references versus LLM-generated references of varying quality.

Reference	Human Expert Score	Spearman	Kendall	Pairwise Acc
Human-Expert	12.75	0.22	0.16	0.61
GPT-4	2.67	**0.24**	**0.19**	**0.69**
Claude	3.25	0.16	0.15	0.39
GPT-3.5	0.33	-0.41	-0.37	0.28

Table 8: Performance comparison using different reference texts with Qwen2-72B-Chat on TTCW dataset. Human Expert Score represents the average TTCW score for the reference text.

Overall, higher-quality references perform better, and references whose quality is close to that of the candidates provide the strongest discrimination. On TTCW with Qwen2-72B-Chat, using GPT-4 as the reference yields the best alignment (Spearman 0.24, Kendall's τ 0.19, pairwise accuracy 0.69), exceeding the human-expert reference (0.22/0.16/0.61), while a much weaker reference (GPT-3.5) substantially degrades alignment (pairwise accuracy 0.28). These results suggest two practical guidelines: (i) use high-quality references; and (ii) prefer references whose quality is reasonably close to the candidates'.

A.11 Evaluation Dimensions for GRW Dataset

Context: In our evaluation on the GRW dataset from Gómez-Rodríguez and Williams (2023), we adopted the set of evaluation dimensions originally defined by the dataset itself. Specifically, we modified the prompts used in our framework by replacing the `BACKGROUND` —originally containing the TTCW rubric—with the following dimensions. This replacement was applied consistently to both the baseline and our method.

Evaluation Dimensions:

- **Readability:** Overall cohesion and holistic flow of the story.
- **Narrative elements:** Use of narrative techniques such as vocabulary choice, imagery, setting, themes, dialogue, characterisation, and point of view.
- **Structural elements:** Control of structural aspects such as spelling, grammar, punctuation, paragraphing, and formatting.
- **Plot logic:** Coherence of narrative structure including hook, conflict, initial crisis, rising and falling action, and resolution (Freytag's pyramid).
- **Creativity:** Originality, innovation, research, credibility, and avoidance of clichés or derivative tropes.
- **JKT style:** Emulation of John Kennedy Toole's writing style using specified stylistic indicators.
- **Epic genre:** Appropriate use and understanding of the heroic/legendary adventure genre.
- **Combat:** Credibility and vividness of a single combat scene.
- **Acc. characters:** Accurate and effective inclusion of both Ignatius J. Reilly and a pterodactyl in action and description.
- **Humor:** Effective use of dark, characteristically humorous tone.

Model Setup: For the baseline, each model was prompted to self-assess story quality based on these dimensions. Our method (`ours`) integrates the same set of dimensions into a unified evaluation strategy combining the `Reference-Based Approach` and `Analyze-Rate Prompting`.