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

Anonymous ACL submission

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 (+17%).

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 assign 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). Experimental results show that our method significantly improves the alignment between LLM evaluations and human assessments, achieving a pairwise accuracy of 0.75 (+17%).

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. The Remote Associates Test (RAT) (Mednick and Halpern, 1968) measures creativity by evaluating individuals' ability to identify associative links between unrelated words. 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

In this section, we present the evaluation framework designed for assessing the creativity of large language models (LLMs). The framework focuses on defining the problem setting and establishing a reference-based evaluation method for generated texts. Additionally, we discuss the prompt strategies employed in our experiments, which enhance the effectiveness of the evaluation process.

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

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.2 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 as "True") if the average score across two assessments is non-negative. 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:

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

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

$$\text{Score}_i^k = \sum_j I[(\text{L}_{i,j}^{k,+} - \text{L}_{i,j}^{k,-}) \geq 0]$$

Method	Model	AVG Spearman	AVG Kendall's Tau	Pairwise Accuracy
Baseline	Claude V1.3	0.15	0.16	0.53
	Claude V2	-0.36	-0.36	0.36
	Claude V2.1	-0.34	-0.33	0.36
	Claude 3-Opus	**0.25**	**0.22**	0.56
	Claude 3.5	0.14	0.13	0.50
	ChatGPT	-0.40	-0.38	0.31
	GPT-4	-0.04	-0.04	0.42
	GPT-4o	0.16	0.14	**0.64**
	Gemini-Pro	-0.31	-0.30	0.33
	Qwen-2-72B-Chat	-0.05	-0.07	0.42
Ours	Claude 3.5	**0.50(+0.36)**	**0.44(+0.31)**	**0.69(+0.19)**
	GPT-4o	0.25(+0.09)	0.22(+0.08)	0.56(-0.08)
	Qwen-2-72B-Chat	0.05(+0.10)	0.03(+0.10)	0.44(+0.02)

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. Results are reported for various models, including Claude V1.3, Claude 3.5, GPT-4, and others. 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.

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.

3.3 Prompt Strategy

Previous research has demonstrated that the analyze-rate strategy can improve performance in certain automated evaluation tasks when applied with GPT models (Chiang and yi Lee, 2023). This strategy involves prompting the model to first analyze the sample according to evaluation criteria before assigning a rating (in our framework, a label). In our experiments with different models, we observe a similar effect: the analyze-rate strategy significantly enhances output accuracy. Therefore, we adopt this strategy as the primary prompt methodology in our final prompt framework, which is detailed in Appendix A.1.

4 Experiment

4.1 Dataset

This study utilizes the dataset provided by (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 models such as GPT-3.5, GPT-4, and Claude V1.3. These human evaluations serve as a benchmark for assessing the accuracy of our automated evaluation methods in measuring creativity.

4.2 Baselines

For baseline comparisons, we adopt the original prompting method introduced by (Chakrabarty et al., 2024) with ten models: Claude 3.5, Claude 3-Opus, Claude V1.3, Qwen-2-72B-Chat(Team, 2024), Claude V2.1, Claude V2, GPT-4(OpenAI, 2023), GPT-4o, Gemini-Pro, and ChatGPT.

4.3 Main Result

In our experiments, we evaluate the effectiveness of the proposed method by its ability to correctly assess the relative capabilities of different models. Specifically, for stories generated from the same plot, we calculate their total scores and derive rankings, which are then compared to rankings provided by human expert evaluators. Higher ranking similarity indicates a more effective evaluation method. The ranking similarity is quantified using three metrics: Spearman's correlation(Spearman, 1904), Kendall's tau(Kendall, 1938), and pairwise accuracy, calculated as the proportion of correctly aligned pairwise comparisons between model rankings and human rankings

As shown in Table 1, our method significantly improves performance, particularly for Claude 3.5 and Qwen-2-72B-Chat. For Claude 3.5, Spearman's correlation increases from 0.14 to 0.50, Kendall's tau from 0.13 to 0.44, and pairwise

Model	**Fluency**			**Flexibility**			**Originality**			**Elaboration**		
	ρ	τ	ACC	ρ	τ	ACC	ρ	τ	ACC	ρ	τ	ACC
claude35	0.20	0.19	0.47	-0.08	-0.08	0.56	0.36	0.36	0.81	0.00	0.00	0.44
claude35_ours	0.36	0.35	0.61	-0.04	-0.04	0.61	0.08	0.08	0.58	0.33	0.33	0.64
qwen2-72b-chat	0.00	0.00	0.31	-0.04	-0.04	0.72	0.20	0.19	0.69	0.00	0.00	0.50
qwen2-72b-chat_ours	0.23	0.22	0.44	-0.02	-0.03	0.75	-0.04	-0.04	0.64	0.17	0.17	0.47
Increment (claude35)	0.16	0.16	0.14	0.04	0.04	0.05	-0.28	-0.28	-0.23	0.33	0.33	0.20
Increment (qwen2)	0.23	0.22	0.13	0.02	0.01	0.03	-0.24	-0.23	-0.05	0.17	0.17	-0.03
AVG Increment	0.20	0.19	0.14	0.03	0.03	0.04	-0.26	-0.26	-0.14	0.25	0.25	0.09

Table 2: Performance Comparison Across Creativity Dimensions. The table reports the performance of different models on four creativity dimensions: Fluency, Flexibility, Originality, and Elaboration. Metrics include Spearman's rank correlation coefficient (ρ), Kendall's tau (τ), and Pairwise Accuracy (ACC). Results are presented for baseline models (e.g., claude35, qwen2-72b-chat) and their enhanced versions using the proposed method (e.g., claude35_ours). The Increment rows represent performance improvements for individual models, while AVG Increment shows the average gains in metrics across models.

accuracy from 0.50 to 0.69. Complete results for all evaluation metrics across individual stories and models, including Spearman's correlation, Kendall's tau, and pairwise accuracy, are provided in A.3. These figures offer a detailed breakdown of results across each story, demonstrating consistent improvements in alignment with human evaluations.

These findings establish a new benchmark for automated creativity assessment, achieving the highest recorded accuracy to date. The proposed method demonstrates robust alignment with human evaluations, offering a practical and scalable framework for assessing creative writing capabilities in language models.

5 Discussion

Table 2 compares our method with baseline models across four key creativity dimensions: Fluency, Flexibility, Originality, and Elaboration. The results are reported using three metrics: Spearman's rank correlation (ρ), Kendall's tau (τ), and pairwise accuracy (ACC). For Fluency and Elaboration, our method demonstrates significant improvements over the baseline, reflecting a stronger alignment with human assessments of text fluency and elaboration.

However, the Originality dimension shows a different trend. While Claude 3.5 improves in Fluency and Elaboration, it experiences a decrease in ρ (-28%) and pairwise accuracy (-28%) for Originality, suggesting that our reference-based method may not be effective for evaluating originality. One possible explanation is that LLMs already have enough capacity to assess originality on their own, and the redundant references may introduce bias, distorting the evaluation. Therefore, we propose continuing to use the original baseline method for Originality evaluation, while applying our method to the other three dimensions.

By combining the results from the baseline method for Originality and our method for Fluency, Flexibility, and Elaboration, we establish a new, hybrid benchmark for automated creativity evaluation. The detailed results of this method can be found in A.3 under "ours_opt". This combined approach achieves a pairwise accuracy of 75%, effectively leveraging the strengths of both methods to improve overall performance.

6 Conclusion

In this paper, we proposed an effective automated evaluation method for assessing the creativity of LLM-generated texts, based on the TTCW. Our method employs a reference-based Likert-style approach, scoring generated texts relative to high-quality human-authored references across various creativity dimensions. Experimental results show that our method significantly improves the alignment between LLM evaluations and human assessments, particularly in Fluency and Elaboration, achieving a pairwise accuracy of 0.69 (+8%). However, we observed a decrease in performance for the Originality dimension. Therefore, we suggest continuing to use the original baseline method for assessing Originality, while applying ours for the other dimensions. This hybrid approach establishes a new benchmark for automated creativity evaluation, with a pairwise accuracy of 0.75(+17%).

7 Limitation

One limitation of our method is its reliance on reference stories, which may restrict its scalability for unrestricted article-level evaluations. 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.

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.

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.

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.

Martha T Mednick and Sharon Halpern. 1968. Remote associates test. *Psychological Review*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

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.

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.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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 some background knowledge on specific aspects of creative writing. Carefully read both stories and, using the provided background knowledge, 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]]".
>
> Stories and Question...
>
> Remember, you must end your answer with one of these: [[A»B]], [[A>B]], [[A=B]], [[B>A]], [[B»A]]

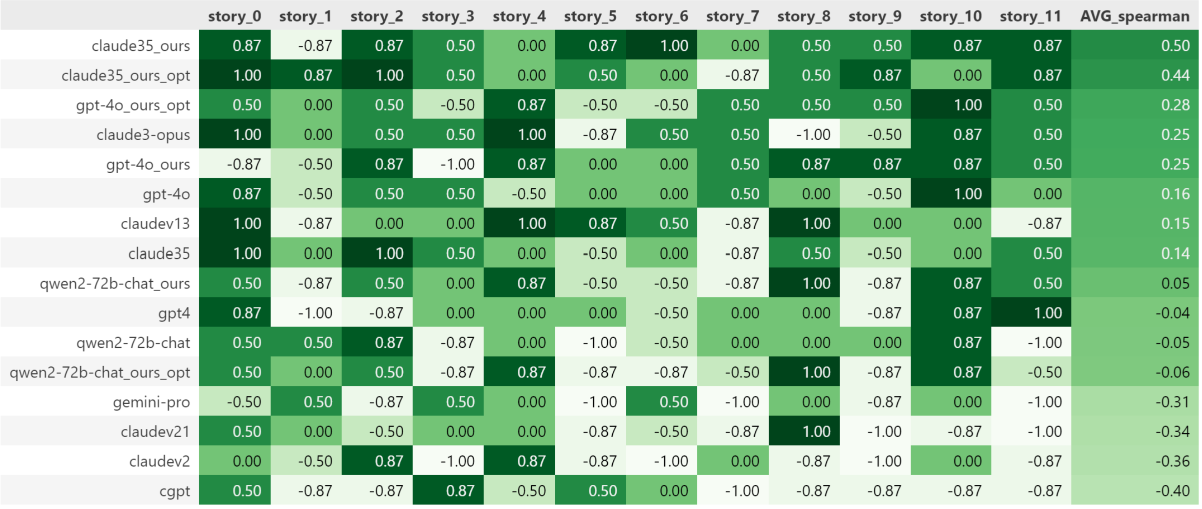
A.2 TTCW Test

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 3: TTCW Dimensions and Questions

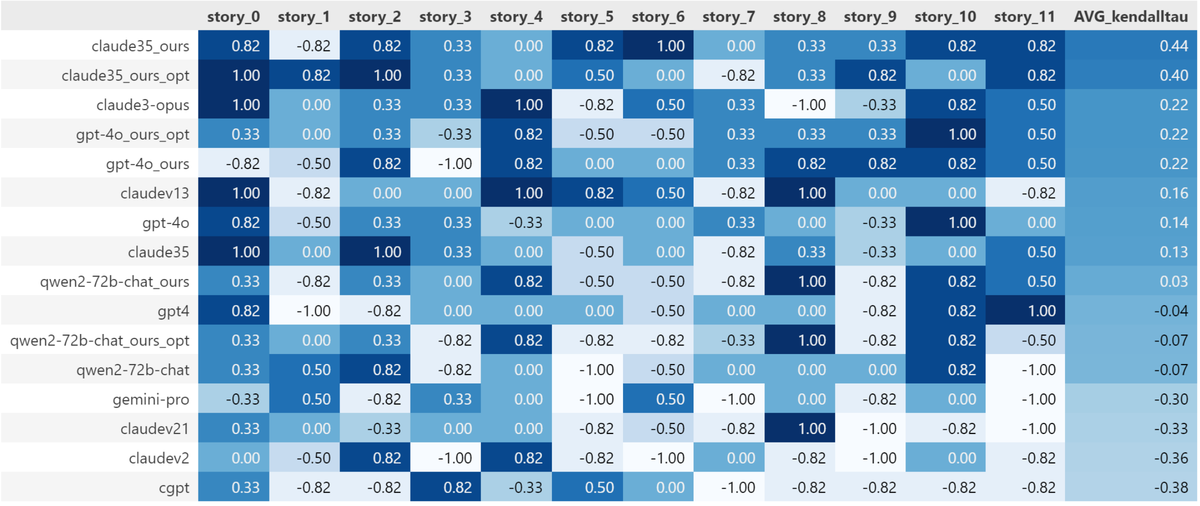
Dimension	**Question**
Fluency	Does the end of the story feel natural and earned, as opposed to arbitrary or abrupt?
Fluency	Do the different elements of the story work together to form a unified, engaging, and satisfying whole?
Fluency	Does the story have an appropriate balance between scene and summary/exposition, 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.3 Full Result



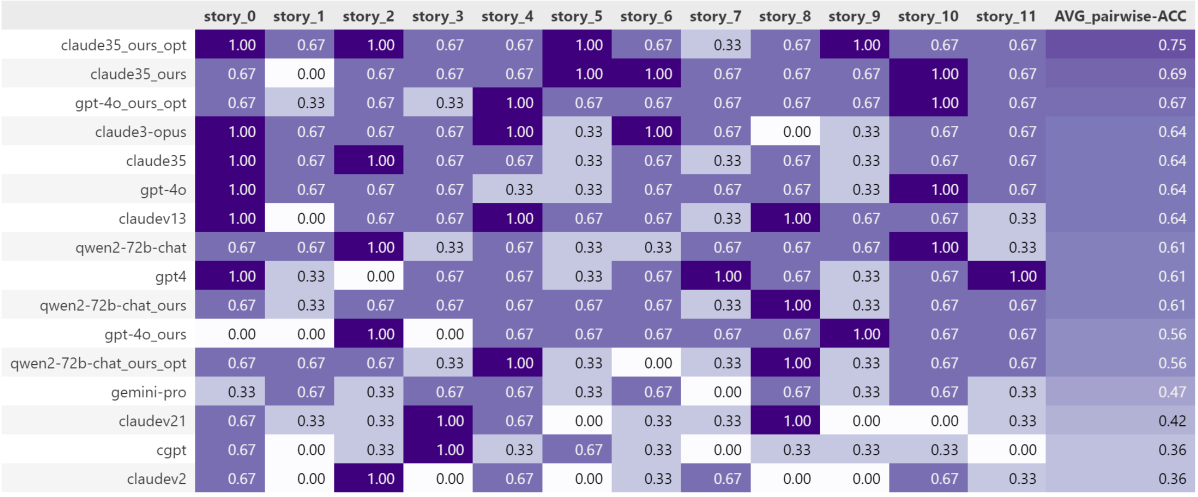
	story_0	story_1	story_2	story_3	story_4	story_5	story_6	story_7	story_8	story_9	story_10	story_11	AVG_spearman
claude35_ours	0.87	-0.87	0.87	0.50	0.00	0.87	1.00	0.00	0.50	0.50	0.87	0.87	0.50
claude35_ours_opt	1.00	0.87	1.00	0.50	0.00	0.50	0.00	-0.87	0.50	0.87	0.00	0.87	0.44
gpt-4o_ours_opt	0.50	0.00	0.50	-0.50	0.87	-0.50	-0.50	0.50	0.50	0.50	1.00	0.50	0.28
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
gpt-4o_ours	-0.87	-0.50	0.87	-1.00	0.87	0.00	0.00	0.50	0.87	0.87	0.87	0.50	0.25
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
qwen2-72b-chat_ours	0.50	-0.87	0.50	0.00	0.87	-0.50	-0.50	-0.87	1.00	-0.87	0.87	0.50	0.05
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	0.50	0.50	0.87	-0.87	0.00	-1.00	-0.50	0.00	0.00	0.00	0.87	-1.00	-0.05
qwen2-72b-chat_ours_opt	0.50	0.00	0.50	-0.87	0.87	-0.87	-0.87	-0.50	1.00	-0.87	0.87	-0.50	-0.06
gemini-pro	-0.50	0.50	-0.87	0.50	0.00	-1.00	0.50	-1.00	0.00	-0.87	0.00	-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	-1.00	0.00	-0.87	-1.00	0.00	-0.87	-0.36
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 1: 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	0.82	-0.82	0.82	0.33	0.00	0.82	1.00	0.00	0.33	0.33	0.82	0.82	0.44
claude35_ours_opt	1.00	0.82	1.00	0.33	0.00	0.50	0.00	-0.82	0.33	0.82	0.00	0.82	0.40
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
gpt-4o_ours_opt	0.33	0.00	0.33	-0.33	0.82	-0.50	-0.50	0.33	0.33	0.33	1.00	0.50	0.22
gpt-4o_ours	-0.82	-0.50	0.82	-1.00	0.82	0.00	0.00	0.33	0.82	0.82	0.82	0.50	0.22
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
qwen2-72b-chat_ours	0.33	-0.82	0.33	0.00	0.82	-0.50	-0.50	-0.82	1.00	-0.82	0.82	0.50	0.03
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_ours_opt	0.33	0.00	0.33	-0.82	0.82	-0.82	-0.82	-0.33	1.00	-0.82	0.82	-0.50	-0.07
qwen2-72b-chat	0.33	0.50	0.82	-0.82	0.00	-1.00	-0.50	0.00	0.00	0.00	0.82	-1.00	-0.07
gemini-pro	-0.33	0.50	-0.82	0.33	0.00	-1.00	0.50	-1.00	0.00	-0.82	0.00	-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	-1.00	0.00	-0.82	-1.00	0.00	-0.82	-0.36
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 2: 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_opt	1.00	0.67	1.00	0.67	0.67	1.00	0.67	0.33	0.67	1.00	0.67	0.67	0.75
claude35_ours	0.67	0.00	0.67	0.67	0.67	1.00	1.00	0.67	0.67	0.67	1.00	0.67	0.69
gpt-4o_ours_opt	0.67	0.33	0.67	0.33	1.00	0.67	0.67	0.67	0.67	0.67	1.00	0.67	0.67
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	0.67	0.67	1.00	0.33	0.67	0.33	0.33	0.67	0.67	0.67	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_ours	0.67	0.33	0.67	0.67	0.67	0.67	0.67	0.33	1.00	0.33	0.67	0.67	0.61
gpt-4o_ours	0.00	0.00	1.00	0.00	0.67	0.67	0.67	0.67	0.67	1.00	0.67	0.67	0.56
qwen2-72b-chat_ours_opt	0.67	0.67	0.67	0.33	1.00	0.33	0.00	0.33	1.00	0.33	0.67	0.67	0.56
gemini-pro	0.33	0.67	0.33	0.67	0.67	0.33	0.67	0.00	0.67	0.33	0.67	0.33	0.47
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	0.33	0.67	0.00	0.00	0.67	0.33	0.36

Figure 3: 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.