An Analysis of Scoring Methods for Reranking in Large Language Model Story Generation

Megan Deering
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
caomegan@cs.toronto.edu

Gerald Penn
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
gpenn@cs.toronto.edu

Abstract

Outline-conditioned story generation using Large Language Models (LLMs) offers a promising approach for automating narrative creation. Some outline-conditioned story generation methods use automatic scoring during the generation process in order to improve the story quality. However, current research has shown that automatic scoring is not ideal for assessing story quality. This paper evaluates three proposed automatic story-scoring methods to improve the reranking of outputs during the generation process. These scoring methods leverage different prompting strategies and fine-tuning techniques to enhance the accuracy and relevance of the assessments. By experimenting with these approaches within a beam search framework, we aim to identify the most effective methods for optimizing story-generation outcomes. While we have found no significant overall difference between these methods in terms of their agreement with human ratings during story generation, the overall story ratings by human evaluators are average. These findings motivate the need for improved automatic scoring techniques and datasets while also indicating that simpler, more easily implementable scoring methods for reranking perform comparably to more complex approaches.

1 Introduction

Recent advancements in AI, particularly in Large Language Models (LLMs), have highlighted how these tools can assist in writing. In creative writing in particular, there has been exploration into how these tools can be used to write fiction stories such as in Yang et al. (2023) and Chung et al. (2022). This issue is particularly challenging because fiction writing demands creativity, raising questions about how to assess the quality of outputs from LLMs in such a creative domain and how to encourage the generation of higher-quality, more engaging stories with these tools.

In the task of outline-conditioned story generation, as first proposed in Rashkin et al. (2020), the input is an outline consisting of key plot-points, characters and events, which are then used to generate a flowing narrative. Each point primes the generation of a passage of text from its corresponding plot-point. Subsequent passages are generated from the next plot-point and the context of the previously generated passages. In our paper, we use outlines consisting of character information, events, and settings in a shallow (two-level) hierarchical structure.

As stated by Rashkin et al. (2020), the difficulty of this task lies in the fact that a model must fluently connect the points given in the text, while still following the outline. Papers such as Yang et al. (2022) and Yang et al. (2023) employ automatic scoring techniques with reranking during the generation process in order to ensure that these stories remain coherent and relevant to the given outline.

Most recent research has suggested that automatic scoring, including scoring by LLMs, is ineffective at assessing story quality. Instead, most story generation papers use human raters as their primary scoring method (Yao et al., 2019; Rashkin et al., 2020; Yang et al., 2023). However, there are cases where automatic scoring is a necessary component in the story generation system. For example, Yang et al. (2023) use automatic scoring in a beam search to select the ideal sequence of passages in a story. This is referred to as reranking, the process of scoring multiple candidate outputs to select the one that best meets specific criteria (Haroutunian et al., 2023). In the context of story generation, reranking could be done automatically or with humans in the loop. Using human-in-the-loop in this context could be tedious and time consuming, however, making the task harder for the humans involved.

In this paper, we investigate how automatic scoring can be used in the story generation process to improve the story outputs. We look at which scor-

Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025), pages 109–120

ing techniques are the best for this specific task. Specifically, we look at three different techniques of automatic scoring using LLMs:

1. **Log-likelihood-based scoring**: Prompts a model with a yes or no scoring question and then uses the log probability of "yes" as a score.

2. **Simple prompt-based scoring**: Prompts a model with a numerical scoring question and uses the output as the score.

3. **Fine-tuning**: Uses a dataset of story scores and fine-tunes a pre-trained model on it.

We use each of these three methods in a reranking framework to generate stories. We then get human raters to evaluate which of these methods, when used for reranking, generates the best stories.

We found that there was no significant difference between these methods when used in reranking for story generation. This means that simple methods like prompt-based scoring perform just as well as other methods which may be more time consuming to implement. It unfortunately also means that fine-tuning a model does not seem to improve the story quality when used for scoring in reranking. Additionally, it further motivates the need for better automatic scoring techniques and datasets.

Our contributions are as follows:

1. We thoroughly evaluate three different automatic scoring techniques for reranking in story generation.

2. We provide code[1] which can be used by others to integrate these scoring methods into their systems for reranking.

2 Related Work

Several previous papers have explored outline-conditioned story generation, where outlines are used as input to generate a story (Yao et al., 2019; Rashkin et al., 2020; Wang et al., 2022).

The DOC framework (Yang et al., 2023) uses verbose outlines and a reranking system in their generation process to choose the best sequence of passages. This paper was inspired by an earlier paper called RE3 (Yang et al., 2022) which also used reranking to choose the best story continuations. They found this reranking component to be critical for plot coherence and premise relevance. Both Yang et al. (2023) and Yang et al. (2022) use a trained model for this reranking.

Zhu et al. (2023) introduces a system with the modularity of the original DOC framework, but which is able to be integrated with more modern LLMs. It also uses log-likelihood-based scoring rather than a trained model for the reranking component. However, the scoring method used for reranking in their system is never fully evaluated. Additionally much previous research has shown that current automatic scoring techniques are not up-to-par with human scoring (Novikova et al., 2017; Guan et al., 2021; Colombo et al., 2023; Chhun et al., 2022, 2024; Chakrabarty et al., 2024).

In Chhun et al. (2022) the authors create the HANNA dataset consisting of human scores of different stories on 5 different criteria: relevance, coherence, empathy, surprise, and engagement. For each of these criteria, they ask humans to give a 5-point Likert score to stories generated by 10 different story generation systems using prompts from the WritingPrompts dataset (Fan et al., 2018). They then compared different automatic scoring techniques to the human scores using Kendall correlations. They found that the correlations between the automatic scores and human scores were weak and called for stronger automatic scoring methods. They found that larger pre-trained models like GPT-2 performed the best, however.

This prompted them to write a follow up paper (Chhun et al., 2024) where they further compared human scoring to automatic scoring using different LLMs for prompting. They found that while LLMs are consistent and have slightly higher ratings, they have correlations with human scores that are fairly similar to those of other automatic scoring methods. They therefore conclude that LLMs are currently the best proxy for human scoring of story generation. They also recommend future work on the use of fine-tuning models for this task.

Guan et al. (2021) address the overall low quality of automatic scoring methods by creating a benchmark called OpenMEVA for them. In contrast to Chhun et al. (2022), they use a single 5-point overall quality score rather than individual scores for several criteria. This overall score should be low for stories that have repetitive plots, unrelated events and conflicting logic, or globally chaotic scenes. They also found that state-of-the art methods corre-

[1] https://github.com/MeganDeer/auto-story-score

Prompt A

Story Passage:
King Aldric, determined...
Event:
King Aldric issues a decree.

Did this event happen in the story passage? Yes or No.

Prompt B

Story Passage:
King Aldric, determined...
Event:
King Aldric issues a decree.

Rate the story on a scale from 1 to 5 on Relevance (how closely the story passage follows the event).
1—The story has no relationship with the event at all.
2—The story only has a weak relationship with the event.
3—The story roughly matches the event.
4—The story matches the event, except for one or two small aspects.
5—The story matches the event exactly.
Do not include any numbers other than your rating in your answer.

Rating (1-5):

Prompt C

Prompt:
King Aldric issues a decree.
Target Story:
King Aldric, determined...

Rate the story on a scale from 1 to 5 on Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity.

Ratings:

Figure 1: The scoring prompts used for log-likelihood-based scoring (prompt A) which has been reproduced from Meta Research (2023), prompt-based scoring (prompt B), and the fine-tuned scorer (prompt C) from left to right on relevance. The fine-tuned model generates scores for all criteria at once.

late poorly with the human methods on this scale in their dataset.

Yang and Jin (2024) discusses different types of automatic scoring. In general, there are four different types that use LLMs. The first of these is embedding-based methods which use embeddings and matching algorithms to assign a score. These have many limitations. The next is probability-based methods. These methods use the generation probability from LLMs in computing their score. There are also generative-based methods, which simply prompt LLMs for a score. Finally, there are trained methods that fine-tune an LLM to assign a score.

Chen et al. (2023) found that generative-based methods are more effective than probability-based methods that use log-likelihood because their smooth distributions allow for better differentiation than the narrow range and peak structure of the probability-based methods.

3 Methodology

In this section, we describe our approach for evaluating automatic story-scoring methods for reranking in outline-conditioned story generation. We first outline the process we use for generating stories and how reranking is used within that. We then describe the criteria we use for the scoring within the reranking. Finally, we describe the three scoring methods to evaluate: log-likelihood-based scoring, prompt-based scoring and fine-tuning.

3.1 Generation

We use the generation component of the framework from Zhu et al. (2023) to generate a story. That is, given an outline, we prompt a model to generate each passage in the outline multiple times. Then, using a beam search, we generate the subsequent passages and select the path with the highest score as the final sequence of passages. We explore using different methods of scoring within this beam search.

3.2 Criteria

We follow Chhun et al. (2022) in forming and defining criteria for scoring passages and evaluating the final stories. We look at the following criteria, which were determined in Chhun et al. (2022) to be good measures of story quality according to the social-sciences literature:

1. **Relevance**: How well the story matches its prompt.
2. **Coherence**: How much the story makes sense.
3. **Empathy**: How well a reader will understand the character's emotions.
4. **Engagement**: How much a reader will engage with the story.
5. **Complexity**: How elaborate the story is.

We leave out the criterion of surprise, which measures how surprising the end of the story is, because we are implementing criteria to score individual passages in the story rather than the entire story.

3.3 Log-likelihood-based Scoring

Zhu et al. (2023) uses log-likelihood-based scoring for reranking in their system. That is, they ask the model the prompts outlined in prompt A of Figure 1 and then calculate the score as the log-likelihood of the answer "yes" being in the response. They also only score each passage on coherence, relevance and commentary. In this case, commentary is used to determine whether or not the passage is actually a part of a story, or just commentary about a story. We also add the commentary criterion to the prompt-based and fine-tuned scoring as we found that, without it, the generated stories were often formulated as brainstorms rather than stories. This acts as our probability-based method.

3.4 Prompt-based Scoring

The next scoring method that we compare is using simple prompt-based scoring. Here, we prompt the model to assign the story a score from 1 to 5 on the criteria in Section 3.2. This acts as our generative-based method.

Additionally, we add guidelines on the definitions of each criterion to better guide the model. We also expand the criteria to include all of the criteria listed in Section 3.2. An example prompt can be found in prompt B of Figure 1.

3.5 Fine-tuned Scorer

The third scoring method in the comparison is a fine-tuned model. This is our trained method.

We fine-tuned a model to score the criteria presented in Chhun et al. (2022). We used LORA Quantization with `llama2-7b` and trained on the HANNA dataset from Chhun et al. (2022) with the prompts presented in Chhun et al. (2024), which can be found in prompt C of Figure 1. This dataset includes story prompts as well as the story generated from these prompts and the scores given by human raters to the story. We trained to match these human scores and used mean-squared error as our loss.

4 Experimental setup

4.1 Outlines

In order to generate the stories using our three methods, we first need to have outlines with which to generate them. We use outlines in the format of those included in Yang et al. (2023). An example outline can be found in Figure A.2. To our knowledge, there is no dataset of outlines that follows this format. Their framework does include a component for generating outlines, but these outlines tend to be very long and detailed — almost story-like — leading to even longer stories that are more difficult for a human to evaluate all at once. Therefore, we asked OpenAI's `gpt-4o` to generate outlines from story prompts contained in the WritingPrompts dataset (Fan et al., 2018). The prompt for generating these outlines is found in Figure A.1. We sampled 16 outlines using this method and one of the present authors ensured that the story plots were interesting and coherent.

4.2 Fine-tuning

To evaluate the fine-tuning, we hold out a test set of data from the HANNA dataset (Chhun et al., 2022). We then calculate the Kendall correlations between the predictions of our fine-tuned model on this test set and the human scores in HANNA. We compare these correlations to the results of the evaluations done in Chhun et al. (2024).

Additionally, we use our fine-tuned model to generate scores on the dataset from OpenMEVA (Guan et al., 2021) to test it on unseen data. Since OpenMEVA has one unified score, while the HANNA dataset has six different scores, we train a basic linear model to learn the weighting of each individual score to contribute to the overall OpenMEVA

score. We use Pearson correlation to evaluate the relationship between our results and human scores, as this statistic is also used in Guan et al. (2021), thus allowing for a direct comparison.

4.3 Human Evaluation

Automatic scoring is used during the story generation process for reranking; therefore, we rely solely on human evaluation to compare the final stories. These are not the same human-sourced numerical scores that are found in several of the available datasets, which we used to calculate the correlations in Section 5.1. This was a separate human-subject evaluation that we conducted, the results of which are described in Section 5.2.

To conduct the human evaluation, we first generate 16 stories for the three experimental conditions: the original pipeline from Zhu et al. (2023) with log-likelihood-based scoring, the modified pipeline with simple prompt-based scoring, and the modified pipeline with the fine-tuned scoring model. For each of these experimental conditions, `llama2-7b-chat` was used for generation. It was also used for the log-likelihood-based and prompt-based scoring. As mentioned in Section 3.5, `llama2-7b` was fine-tuned for the fine-tuned scorer, using a beam-width of 3 for all of the stories.

64 participants were recruited from Prolific, subject to the constraints that they are fluent in English and have a 99-100% approval rating. Each participant was presented with a Word-document survey containing three stories, one from each experimental condition, all generated from the same outline. The order of the experimental conditions was permuted in each survey. The participants were asked to read the three stories and rate the coherence, empathy, and relevance of each using a 1-5 Likert scale. They were also asked, for each outline point, to highlight the passage in each story that best corresponds, in order to further assess relevance. Finally, they were asked to rank the stories in order of the likelihood that they would purchase them, and explain the reasoning behind their decision. We call this the preference ranking. Further details on this evaluation can be found in Table A.1.

We chose to change some of the criteria from Chhun et al. (2022). Again, we left out surprise as we generated passage-by-passage and surprise is, by the definition in Chhun et al. (2022), only applicable to the end of the story. We also replaced the criteria of engagement (Chhun et al., 2022) and "interestingness" (Yang et al., 2023) with preference ranking because the latter is less abstract and better grounded in a ecologically valid task. Yamshchikov and Tikhonov (2023) also claim that human raters may be misinterpreting "interestingness." We also chose to leave out complexity as the annotators were comparing stories that were generated with the same outline, and thus should all be roughly equally elaborate.

To evaluate the results of the human study, we used an ANOVA test to determine the significance of relevance, coherence, and empathy. Before computing ANOVA, we checked for a normal distribution using a histogram and Levene's test. We further evaluated relevance using the highlighted passages that were chosen to correspond to each outline point, calculating the specificity, precision, and recall of the words in the passages compared to the actual alignment of generated words to outline points in the model. We used the Bradley-Terry-Luce (BTL) method to linearise the relative preference rankings into a global ranking of the three scoring methods with respect to one another.

5 Results

5.1 Fine-tuning

Kendall correlations		
Criteria	Fine-tuned Model	Beluga-13B 1[a]
Relevance	0.18	0.21
Coherence	0.30	0.26
Empathy	0.29	0.27
Surprise	0.30	0.17
Engagement	0.33	0.11
Complexity	0.39	0.26

[a] Correlations from Chhun et al. (2024).

Table 1: Kendall correlations of the scoring model fine-tuned for three epochs to human-sourced scores in the HANNA dataset, rounded to two decimal places, along with human correlations to the Beluga-13b 1 model as reported in Chhun et al. (2024).

In order to evaluate the fine-tuning alternative, we use Kendall correlations to compare to those reported in Chhun et al. (2024). We found that for all criteria, we were able to achieve better correlations than those from Chhun et al. (2024) to the human-sourced scores found in the HANNA dataset. This indicates that the fine-tuning was able to improve conformity on this dataset. The results for a model trained with 3 epochs can be seen in

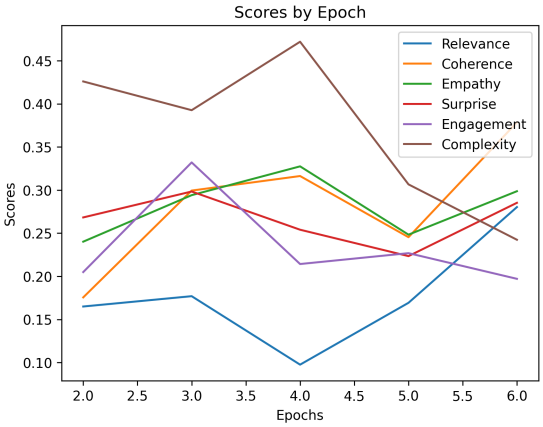

Figure 2: The Kendall correlations of the fine-tuned model trained for different numbers of epochs.

Table 1. Although relevance was unable to beat the score from Chhun et al. (2024) within 3 epochs, beyond 6 epochs, it did ($\tau = 0.28$).

Thus, we found that different training times were required to achieve the best results for different criteria. Figure 2 shows the graph of Kendall correlations trained for different numbers of epochs. Relevance, coherence, and surprise do better with more training. However, complexity has the best performance with 4 epochs and engagement does the best with 3 epochs of training. After that, their performance noticeably decreases due to overfitting. Empathy does the best on 4 epochs, but performance continues to improve with more training. Overall, 3 epochs achieve the best overall balance between Kendall correlation and training time.

These results imply that it is much more difficult to train for relevance and coherence. In fact, the Kendall correlations for coherence are extremely low before 6 epochs. In the case of relevance, this may be because it requires the language model to pay attention to more information. For instance, the model needs to attend more closely to the early portion of the prompt, whereas other criteria are only concerned with the story itself.

Another possible contributor to this difficulty could be the lack of clarity on what the criteria mean. Chhun et al. (2022) also found less than favourable Kendall correlations for relevance, but the instructions that they gave to their annotators to score relevance were very underspecific, asking them merely to "measure how well the story matches the prompt" (Chhun et al., 2022). Coherence, furthermore, enjoys no consensus on how it is used in natural language generation (Yamshchikov and Tikhonov, 2023). This may have affected how the annotators scored coherence in the HANNA dataset.

Because of these results, we decided to use the model trained for 3 epochs for the criteria of complexity, engagement, and empathy and the model trained for 6 epochs for relevance and coherence in our reranking. We do not use surprise for our reranking.

We also tested our fine-tuned model on the OpenMEVA dataset (Guan et al., 2021) to evaluate its performance on unseen data. The ROCStories and WritingPrompts datasets (Mostafazadeh et al., 2016; Fan et al., 2018) are evaluated separately in Guan et al. (2021), but their stories are interspersed in the available data. Therefore, we conducted our evaluation on a mixture of both datasets and compared our results using Pearson correlations to the human-sourced scores found in OpenMEVA for the ROCStories and WritingPrompts datasets, as this was the statistic presented in Guan et al. (2021).

Our model achieved a Pearson correlation of 0.2281, outperforming BERTScore-F1 (Zhang* et al., 2020), which scored 0.1271 on ROCStories and 0.0329 on WritingPrompts (Guan et al., 2021), as well as RUBER-BERT (Ghazarian et al., 2019), which scored 0.1434 and 0.2116, respectively (Guan et al., 2021). However, it falls short of the best-performing method, UNION (0.4119/0.3256), from Guan and Huang (2020). While our fine-tuned model does not achieve the highest performance, it remains competitive with other methods evaluated in Guan et al. (2021). This is notable given that our method was not explicitly trained to assess the specific aspects targeted in the OpenMEVA evaluation, such as repetition and conflicting logic.

5.2 Human Study

In total, we collected surveys from 64 participants on 16 different stories for each experimental condition. Each story was evaluated by 4 participants on relevance, coherence, empathy and preference. Additionally, we had participants annotate which passages they believed correlated to each outline point.

5.2.1 Ratings

We evaluated the significance of relevance, coherence and empathy using ANOVA as outlined in Section 4.3.

For coherence, as seen in Figure 3, the generated

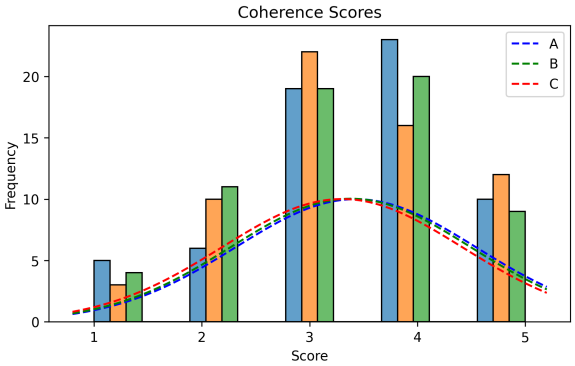

Figure 3: Histogram of the coherence ratings given by human raters to each of the three methods. A: log-likelihood-based scoring, B: simple prompt-based scoring and C: fine-tuned scoring.

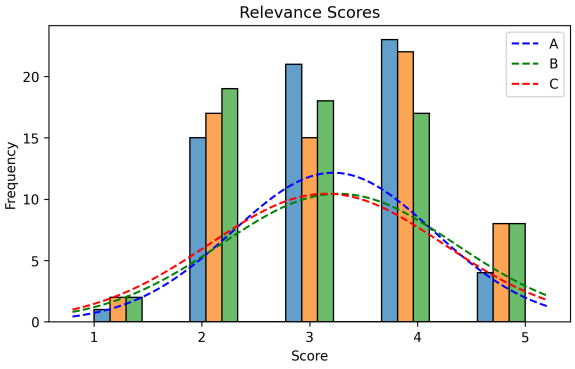

Figure 4: Histogram of the relevance ratings given by human raters to each of the three methods. A: log-likelihood-based scoring, B: simple prompt-based scoring and C: fine-tuned scoring.

histogram and the Levene score of 0.9973 suggest that it is normally distributed. The p-value from the ANOVA test was 0.58. Therefore, the differences of the human ratings of coherence may not be significant.

For relevance, as seen in Figure 4, the histogram generated appears to be normal and we got a Levene score of 0.23, suggesting that it may be normal. The p-value achieved from our ANOVA test was 0.757, and so no significance was demonstrated.

For empathy, similarly to relevance, the data appear to be somewhat normal with a Levene score of 0.19, and are shown in Figure 5. However, the p-value again failed to demonstrate significance at 0.397.

Overall, human raters rated coherence an average of 3.37, relevance 3.21, and empathy 3.29 across all of the stories.

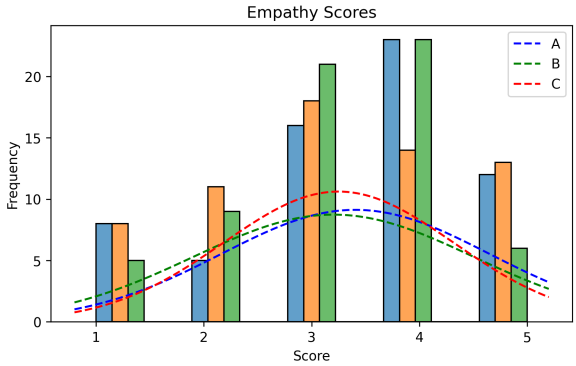

Figure 5: Histogram of the empathy ratings given by human raters to each of the three methods. A: log-likelihood-based scoring, B: simple prompt-based scoring and C: fine-tuned scoring.

Outline Annotation Statistics			
Method	Specificity	Precision	Recall
Log-likelihood	0.3901	0.4220	0.5345
Simple prompt	0.3527	0.3606	0.4733
Fine-tuning	0.4126	0.4449	0.5780

Table 2: Specificity, precision, and recall of human annotations for identifying corresponding passages in generated stories. Annotators selected passages they believed best matched each outline point. Specificity indicates how many words (as a ratio of passage length) they selected which overlapped with the actual, corresponding, generated passages. Precision and recall are both ratios in which the numerator is the sum of the number of annotators that selected each correctly annotated word. Precision divides this by the product of the total number of correctly annotated words and the number of annotators. Recall divides it by the number of words in the generated passage multiplied by the number of annotators. Each of these three scores is then aggregated over all of the passages using a macro average.

5.2.2 Outline Annotation

As described in Section 4.3, we further instructed the human subjects to annotate which passages from the generated story they believed to correspond best to each outline point. We then calculated the specificity, precision, and recall of these selections relative to the actual corresponding passages from the alignment used by the LLM generator. The results can be found in Table 2.

From these results, we can see that generating with fine-tuned scoring produces the highest specificity, precision and recall, suggesting that it adheres most closely to the input outlines and/or that it compels human raters to select longer passages to

hedge their uncertainty. However, considering that all scores are fairly close and the relevance ratings in Section 5.2.1 are not statistically significantly different, it is likely that these methods are similar in their effectiveness in scoring relevance.

5.2.3 Preference Ranking

To evaluate the preference ranking we used the Bradley-Terry-Luce method, from which we determined the following preference ranking, where high scale values indicate higher preference:

1. Prompt-based scoring (scale 0.884)
2. Log-likelihood-based scoring (scale 0.464)
3. Fine-tuned scoring (scale -0.1349)

Therefore, in spite of their performances in respect of relevance, the simple prompt-based scoring method was the most preferred, while the fine-tuned scoring was the least preferred.

In their explanations of why they chose their rankings, many participants cited coherence as a reason. This may be in part due to the fact that they had been asked earlier in the survey to rate the coherence of the stories, although they were primed to the same extent for empathy and relevance, and empathy was cited far less frequently, whereas relevance was hardly ever mentioned. This suggests that coherence has a greater impact on a reader's preference for a generated story than the other criteria that participants were asked to score.

On the other hand, when comparing the preference ranking to the coherence histogram in Figure 3 and the p-value for coherence, there are clearly other factors that have influenced their decision.

Many participants additionally mention how engaging or interesting the story is as a factor in their choice. As mentioned in Section 4.3, we chose this question to replace the criteria of engagement and "interestingness" from Chhun et al. (2022) and Yang et al. (2023). This was an effective question. Many participants also mentioned that they enjoyed the stories due to the overall structure or flow of the story in how it was presented. One participant, for example, wrote about why they preferred a prompt-based scoring story:

> I preferred Story 3 because it presents an exciting, high-stakes conflict surrounding an advanced weapon and the dangerous implications of its existence. The tension between the characters and the mystery about their father's involvement with a shadowy organization adds depth to the narrative, and I find the mix of technology, moral dilemmas, and intrigue particularly engaging.

5.3 Discussion

These results suggest that there may be no significant difference between prompting, log-likelihood and fine-tuning as reranking methods in this framework. It also calls into question how thorough previous work has been in this area. Tasks have been ill-defined, instructions to annotators have been lacking in specificity or ecologically questionable, and some papers (Yang et al., 2023, e.g.,) have advocated for switching from one method to another without experimentally determining which was better.

Log-likelihoods derived from LLMs are a competitive approach, although there is evidence that prompting is preferable overall. The bad news is that even the effort of fine tuning does not seem to provide a significant improvement to automated reranking. This could suggest that there is still a need for better human-rated datasets of generated story output.

6 Conclusion

In this paper, we compared three different automatic scoring methods when used in a reranking framework for story generation. Our experiments were unable to prove a significant difference among these methods when their outputs are assessed by human raters. Future work could explore alternative reranking techniques, such as reinforcement-learning-based methods. Additionally, there is a need for more annotated datasets of stories. We hope that this research contributes to more effective and controllable story generation systems in NLP.

Limitations

A key limitation of our approach lies in the datasets used for fine-tuning. Beyond the previously mentioned issues regarding the wording of instructions for annotators, this dataset was designed to score entire stories, whereas our task focused on scoring individual passages within a larger narrative. Using a model fine-tuned for full-story scoring on smaller passages may not be an effective solution, highlighting the need for more specialized datasets.

Additionally, stories generated without reranking could have been compared as a baseline. We did not pursue this approach since Yang et al. (2022) found through ablation studies that reranking was essential for the generation process. Now that LLMs have improved, however, there may be a different outcome.

Acknowledgments

This research was supported by a grant from NAVER corporation.

References

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 2024 CHI Conference on Human Factors in Computing Systems*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of IJCNLP-AACL 2023*, pages 361–374.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chlo'e Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836.

Cyril Chhun, Fabian M. Suchanek, and Chlo'e Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2023. The glass ceiling of automatic evaluation in natural language generation. In *Findings of IJCNLP-AACL 2023*, pages 178–183.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6394–6407.

Levon Haroutunian, Zhuang Li, Lucian Galescu, Philip Cohen, Raj Tumuluri, and Gholamreza Haffari. 2023. Reranking for natural language generation from logical forms: A study based on large language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1067–1082.

Meta Research. 2023. doc-storygen-v2.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Jekaterina Novikova, Ondvrej Duvsek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. Chae: Fine-grained controllable story generation with characters, actions and emotions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6426–6435.

Ivan P. Yamshchikov and Alexey Tikhonov. 2023. What is wrong with language models that can not tell a story? In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 58–64.

Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation. *Preprint*, arXiv:2408.14622.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3378–3465.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: towards better automatic storytelling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*.

Hanlin Zhu, Andrew Cohen, Danqing Wang, Kevin Yang, Xiaomeng Yang, Jiantao Jiao, and Yuandong Tian. 2023. End-to-end story plot generator. *Preprint*, arXiv:2310.08796.

A Appendix

Could you, using the following format: {outline}

please generate an outline for a short story with the following writing prompt: '{prompt}'

The outline should have at most 8 nodes.

Outline:

Figure A.1: Prompt for generating outline. Here, {outline} is replaced with an example outline in the required format and {prompt} is replaced with the story prompt that we want to use to generate the outline.

Premise: A humble cheesemaker, renowned for making the best cheese in town, embarks on a daring adventure to slay a fearsome dragon that threatens the kingdom
Setting: A quaint village surrounded by rolling hills and dense forests, within a kingdom plagued by a dragon.
Entities and Characters:

- **Milo:** Milo is the best cheesemaker in the village, known for his delicious and unique cheeses

  ⋮

Outline:

1. **Event:** The village is terrorized by a dragon.
   Characters: Milo, King Aldric
   Setting: Milo's cheese shop.

   (a) **Event:** King Aldric issues a decree.
   Characters: King Aldric
   Setting: The King's castle.
   ⋮

2. **Event:** Milo prepares for the journey.
   Characters: Milo, Fiona
   Setting: Milo's home.
   ⋮

Figure A.2: Example outline for input to the generation framework.

Example Survey

Instructions

- You will be presented with three stories in this document.
- After each story, you will find a set of questions related to that story.
- At the end, there will be a final question asking you to compare all three stories.

Please carefully read all stories and questions. Write your answers in the designated highlighted spaces. Also be sure to highlight the story when instructed to. Important: Answer all questions.

Story 1	Jake Hunter was known for his lightning-fast reflexes and uncanny ability to anticipate his opponents' moves ...
How coherent was the story (how clear and sensible is it and how well does it flow logically together)? (1-5)	(1) Not at all coherent (The entire story is unclear, doesn't make sense at all and is inconsistent throughout). (2) It is logically inconsistent or doesn't make sense for most of the story. (3) The story is logically consistent and makes sense overall, but has some inconsistencies or parts that don't make sense. (4) The story is coherent but it has one or two inconsistencies or incoherences. (5) The story is entirely coherent - there are no inconsistencies and it makes sense.
2. How well did you understand the character's emotions? (1-5)	(1) The characters seemed apathetic to you. (2) At least one character slightly related to you on an emotional level. (3) You recognized specific, but not necessarily strong, emotions (e.g. sadness, joy, fear. . .) in at least one character. (4) At least one character emotionally involved you, but minor details prevented you from completely relating to them. (5) At least one character completely involved you on an emotional level.
The following is a bullet point story outline used to generate this story. First, **highlight the text in the story that corresponds to each outline point using the specified color for each point** (for example the first outline point is yellow, the second is green, etc.). Then answer the question below.	• Introduction to Jake's unique ability and career – EVENT:Jake's latest victory showcasing... SETTING:The boxing ring CHARACTERS:Jake Hunter, Coach Reynolds – EVENT:Coach Reynolds discusses Jake's ... • ...

(Continued)	
How well do you feel like this text captures each outline point? How well are the characters, scene and event represented in each passage? (1-5)	(1) The text has no relationship with the outline points at all. (2) The text only has a weak relationship with the outline points. (3) The text roughly follows the outline points. (4) The text follows the outline points, except for one or two small aspects. (5) The text follows the outline point exactly.
Story 2	...
Story 3	...
For the following question, place your ranking for each story inside the highlighted brackets.	
Rank the three stories based on how likely you would be to purchase them, assuming you were given the money to do so. Assign a rank of 1 to the story you are most likely to buy, 2 to the next, and 3 to the least likely. Each rank must be assigned to only one story.	() Story 1 () Story 2 () Story 3
Why did you prefer the story you did? Please explain your preferences in 1-3 sentences below.	

Table A.1: An example survey given to participants in the human study. Participants are given the same questions for each story and then asked to rank their preference between stories. The order of the stories is changed for each survey. Each story in the survey is generated from the same outline.