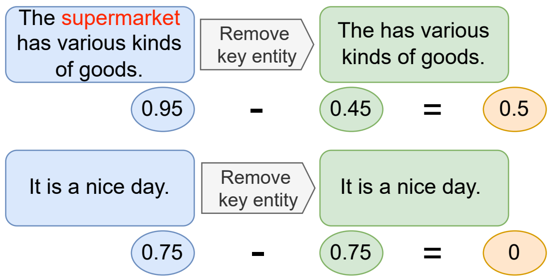
DELTASCORE: Story Evaluation with Perturbations

Zhuohan Xie **Miao Li** **Trevor Cohn** **Jey Han Lau**
School of Computing and Information Systems,
The University of Melbourne
{zhuohanx, miao4}@student.unimelb.edu.au, {t.cohn, laujh}@unimelb.edu.au

Abstract

Various evaluation metrics exist for natural language generation tasks, but they have limited utility for story generation since they generally do not correlate well with human judgments and are not designed to evaluate fine-grained story aspects, such as fluency and relatedness. In this paper, we propose DELTASCORE, an approach that utilizes perturbation to evaluate fine-grained story aspects. Our core idea is based on the hypothesis that the better the story performs in a specific aspect (e.g., fluency), the more it will be affected by a particular perturbation (e.g., introducing typos). To measure the impact, we calculate the *likelihood difference* between the pre- and post-perturbation stories using large pre-trained language models. We evaluate DELTASCORE against state-of-the-art model-based and traditional similarity-based metrics across two story domains, and investigate its correlation with human judgments on five fine-grained story aspects: fluency, coherence, relatedness, logicality, and interestingness. The findings of our study indicate that the DELTASCORE approach exhibits exceptional performance in evaluating intricate story aspects. An unexpected discovery was made in our experiment, where a single perturbation method was found to effectively capture a majority of these aspects.

(a) Perturbation "Add typos" affects the higher quality story (top) more.



(b) Stories conditioned on the prefix "I always go to the local supermarket". Perturbation "Remove key entity" affects the higher quality story (top) more.

Figure 1: Scenarios where higher quality stories (top) are affected more than lower quality ones (bottom) through aspect-specific perturbations (fluency: "Add typos"; relatedness: "Remove key entity"). Generative likelihood for original/perturbed story is in blue/green circle, and the DELTASCORE value is in orange circle.

1 Introduction

The advent of pre-trained language models (PLMs) and large language models (LLMs) (Radford et al., 2019; Lewis et al., 2020; Brown et al., 2020; Zhao et al., 2023) has enabled story generation models to produce plausible stories (Tan et al., 2021; Zhang et al., 2022b; Yang et al., 2022). In fact, the best models have been found to produce stories that are virtually indistinguishable from those written by humans (Karpinska et al., 2021; Dou et al., 2022; Xie et al., 2023). However, progress in the development of automatic evaluation metrics has not had the same momentum (Guan et al., 2021b). Although human evaluation remains the gold standard, it is often slow, expensive, and difficult to reproduce (Sai et al., 2023). Therefore, there is a pressing need for better automatic methods to evaluate story quality.

The predominant evaluation metrics for story evaluation have their origins in other natural language generation (NLG) tasks, including BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) for machine translation and ROUGE (Lin, 2004) for summarization. Nonetheless, recent efforts have led to the development of novel metrics for story evaluation, aiming to quantify story coherence (Guan and Huang, 2020; Ghazarian et al., 2021) or learn human preferences (Chen et al., 2022). Other works have directly utilized the likelihood of a story under a PLM (Vaswani et al., 2017; Han et al., 2022) or its conditional likelihood based on human references or other contextual factors, such as story title (Thompson and Post, 2020; Yuan et al., 2021). However, these approaches typically produce a single score that estimates the overall quality of the generated story. Chhun et al. (2022) argue that the quality of a story is composed of multiple fine-grained aspects, such as fluency and adherence to commonsense, and that an overall quality score has limited utility for story evaluation. In other words, a metric that produces a low overall score for a story does not reveal whether the story has fluency issues or certain elements of the story violate commonsense (Leiter et al., 2022).

In this paper, we propose DELTASCORE, an approach that assesses story quality by calculating the *likelihood difference* under a pre-trained language model between a story and its perturbed form, with the idea that higher quality stories will be affected more by the modification/perturbation compared to the lower quality ones. To provide fine-grained assessment of story quality, we experiment with perturbations that target a particular aspect (e.g., fluency). Figure 1 presents two examples to demonstrate the intuition of our approach: (1) when we modify the two stories in Figure 1a by randomly adding typos, the more fluent story (top) is impacted by the perturbation more than the less fluent story (bottom); (2) when we modify the two stories — conditioned on the story title "I always go to the local supermarket" — in Figure 1b by removing key entities, the story that is more related to the title (top) is again more affected. Through our empirical analysis, we have demonstrated that the DELTASCORE methodology outperforms existing model- and similarity-based metrics in evaluating intricate story quality aspects. Our investigation also revealed an intriguing finding, whereby the word jumbling perturbation technique displays exceptional accuracy in capturing diverse aspects. This discovery implies that there may be interdependence among these fine-grained aspects.

2 Related Work

2.1 Automatic Story Evaluation

The conventional evaluation metrics used in natural language processing mainly focus on measuring lexical overlap between machine-generated text and its human reference (Papineni et al., 2002; Doddington, 2002; Lin, 2004). However, such methods are limited in their ability to capture semantic similarity, and they are vulnerable to small changes in morphology or even typos, as highlighted by Kaster et al. (2021). To overcome these limitations, recent methods have exploited the language understanding capability of pre-trained language models (PLMs) to capture semantic similarity by comparing embeddings (Zhao et al., 2019; Zhang et al., 2020). However, these evaluation metrics still rely on a single human reference, which may not be sufficient for story evaluation, given that there can be multiple good stories for a given condition.

Therefore, researchers have started to focus on reference-free evaluation metrics that are tailored to evaluate stories. For instance, Guan and Huang (2020) train a binary classification model to distinguish between original stories and their negative perturbations, while Ghazarian et al. (2021) extend this idea by generating negative samples from manipulated story plots instead of heuristic rules. These metrics place emphasis on evaluating story coherence. More recently, Chen et al. (2022) propose a new task that aims to evaluate human preference by training a model to differentiate highly-upvoted stories from lowly-upvoted ones, drawing labelled data from Reddit. This task is particularly useful for evaluating the subjective aspects of stories that cannot be easily captured by traditional metrics.

While conventional evaluation metrics typically provide a single score for overall quality, recent evaluation metrics have started to work assessing quality from different perspectives by offering different combinations of inputs, such as given prompts, generated text, and additional context (Zhong et al., 2022). For example, Yuan et al. (2021) consider text evaluation as a text generation problem, while Deng et al. (2021) exploit information alignment to evaluate the generated text.

Although these methods provide several assessments of text quality from a few perspectives, they do not correspond to more intuitive quality aspects such as fluency. To address this limitation, Zhong et al. (2022) consider text evaluation as a question answering problem and explore training a language model to answer specific questions, such as "*Is this a fluent sentence?*". Fu et al. (2023) propose utilizing the generalisation capability of large language models for evaluation, by prefixing instructions such as "*Generate a fluent story for the given title*" as part of the input to help the likelihood computation to align to a particular aspect.

2.2 Natural Text Perturbation

Perturbation has been widely used as a conventional technique to generate negative samples that can be utilized in both discriminative (Guan and Huang, 2020) and generative (Zhong et al., 2022) NLP tasks. To conduct behavioral tests of NLP models, Ribeiro et al. (2020) proposes a CheckList that provides a suite of negative examples constructed from a combination of perturbations. Similarly, Karpinska et al. (2022) designs a list of perturbation tests to evaluate the robustness of evaluation metrics for machine translation. Moreover, Sai et al. (2021) extends the idea of perturbations to evaluate the robustness of NLG evaluation metrics. To find blind spots of model-based evaluation metrics, He et al. (2022) designs perturbation tests. It is worth noting that all these perturbations are based on heuristic rules. In contrast, recent adversarial attacks, such as those proposed by Li et al. (2020); Morris et al. (2020), apply language models to generate adversarial examples, which can also be considered as another form of text perturbation. In this work, we experiment with perturbation for a different purpose: to evaluate story quality in a more fine-grained manner.

3 DeltaScore

We now describe the idea of our approach. Given a story condition (e.g. story title) $\boldsymbol{c} = c_1, ..., c_n$ containing n tokens, a model-generated story $\boldsymbol{s} = s_1, ..., s_m$ containing m tokens, and a perturbed/modified story $\boldsymbol{s}' = s'_1, ..., s'_{m'}$ containing m' tokens, DeltaScore calculates the likelihood difference under a language model:

$$\text{DELTASCORE}(\boldsymbol{s}) = \log p(\boldsymbol{s}|\boldsymbol{c}) - \log p(\boldsymbol{s}'|\boldsymbol{c}) \tag{1}$$

where $p(\boldsymbol{s}|\boldsymbol{c})$ represents the likelihood of s conditioned on c under a language model. In our experiments, we investigate several large language models with varying architectures (see § 3.1) and perturbation techniques that are designed to target specific aspects (see § 3.2).

3.1 Two Different Likelihood Calculations

We now explain how we compute $p(\boldsymbol{s}|\boldsymbol{c})$ with encoder-decoder PLMs (e.g. BART (Lewis et al., 2020) and T5 (Raffel et al., 2020)) and decoder-only PLMs (e.g. GPT3 (Brown et al., 2020)).

Denoting language model parameters as θ, we compute DeltaScore as follows for encoder-decoder PLMs:

$$\log p(\boldsymbol{s}|\boldsymbol{c}) = \frac{1}{m}\sum_{t=1}^{m} \log p(s_t|\boldsymbol{s}_{<t}, \boldsymbol{c}, \theta) \tag{2}$$

$$\log p(\boldsymbol{s}'|\boldsymbol{c}) = \frac{1}{m'}\sum_{t=1}^{m'} \log\ p(s'_t|\boldsymbol{s}'_{<t}, \boldsymbol{c}, \theta) \tag{3}$$

where, t denotes timestep in the sequence, and $\boldsymbol{s}_{<t}$ denotes all tokens before the current timestep. Intuitively, the story condition c is captured by the encoder, and the likelihood of the story s is produced by the decoder.

For decoder-only PLMs, we have to concatenate $\boldsymbol{c}$ and $\boldsymbol{s}/\boldsymbol{s}'$ to form one sequence (e.g. $\boldsymbol{x} = x_1, ..., x_{n+m} = c_1, ..., c_n, s_1, ..., s_m$) to compute DeltaScore:

$$\log p(\boldsymbol{s}|\boldsymbol{c}) = \frac{1}{m}\sum_{t=n+1}^{n+m} \log\ p(x_t|\boldsymbol{x}_{<t}, \theta) \tag{4}$$

$$\log p(\boldsymbol{s}'|\boldsymbol{c}) = \frac{1}{m'}\sum_{t=n+1}^{n+m'} \log\ p(x'_t|\boldsymbol{x}'_{<t}, \theta) \tag{5}$$

Intuitively, we feed the full sequence including the story condition $\boldsymbol{c}$ and story $\boldsymbol{s}$ as input to the decoder-only PLM, although when computing the aggregate likelihood we only consider conditional likelihood for the $\boldsymbol{s}$ tokens.

3.2 Perturbations on Story Aspects

As noted by Xie et al. (2023), the fundamental aspects of story quality include fluency, coherence, relatedness, logicality, and interestingness. In order to perform a fine-grained evaluation of story quality, we propose aspect-aware perturbations of stories. We posit that perturbations designed for each aspect should have distinct objectives.

- **Fluency:** Perturbations on fluency usually focus on modifications on word or phrase level, which results in issues within one sentence. This usually includes intentionally using incorrect verb tenses, subject-verb agreement or word/phrases repetitions.

- **Coherence:** Perturbations on coherence usually focus on modifications on sentence level while keeping each individual sentence fluent, aiming to make the story have conflicts

Aspect	Perturbation	Original story	Perturbed story
Flu.	Typo	he went to see what the problem was	he went to see whta the problem was
	SubjVerbDis	he is the best student in the classroom .	he am the best student in the classroom .
Coh.	Jumble	We play badminton every evening .	badminton every We evening play .
	SentReorder	she did n't intend to buy anything . unfortunately she has poor impulse control ...	unfortunately she has poor impulse control . she did n't intend to buy anything ...
Rel.	RmkeyEntities	The supermarket has various kinds of goods	The has various kinds of goods
	StoryReplace	The supermarket has various kinds of goods	It is a nice day to hang out
Log.	Antonym	The boy got the gift he always wanted, he was so happy .	The boy got the gift he always wanted, he was so sad .
	Commonsense	they took me down to the lake . i threw my line out and caught several worms ...	they took me to the moon. i threw my line out and caught several stars ...
Int.	BlanderNarrative	i felt really angry, talked to my estranged father , and he gave me a gun! But I knew violence is not a solution here .	I felt upset and talked to my father about it . He advised me to handle the situation calmly , so I decided not to resort to violence .

Table 1: Selected perturbations that focus on each story quality aspect: Fluency (Flu.), Coherence (Coh.), Relatedness (Rel.), Logicality (Log.), and Interestingness (Int.). For stories on relatedness aspects, they are conditioned on the title "I always go to the local supermarket". Underlined perturbations are originally proposed by us.

between sentences. This usually includes intentionally repeating one sentence, replacing one sentence from one unrelated story, or reordering all sentences.

- **Relatedness:** Perturbations on relatedness need to alter the story so that it is less relevant to the story condition. It can be achieved by either removing key relevant information relevant to the prompt or replace the entire story by one story written for the different prompt.

- **Logicality:** Perturbations on logicality need to alter the story so that it disobey commonsense. This could include characters that defy the laws of physics or events that go against cultural norms such as "go trick or treating" on "Christmas".

- **Interestingness:** Perturbations on interestingness need to alter the story so that it is generally less interesting to read. This could include changing the tone of the story from dramatic to bland or altering descriptive words with flat ones.

Due to the interdependence of these quality aspects, perturbations designed for one aspect can potentially affect other aspects. In particular, aspects such as interestingness, are more subjective and influenced by individual perspectives and preferences. Hence, our taxonomy may not be universally applicable to all domains.

To provide a comprehensive evaluation of story quality, we first align the perturbations proposed in previous works (Sai et al., 2021; Guan et al., 2021b; He et al., 2022) to the fundamental aspects of story quality identified by Xie et al. (2023). In addition, we propose several new perturbation approaches, the majority of which are developed with the assistance of ChatGPT. The details of the prompts used for these perturbations can be found in Appendix A.

- **RmkeyEntities:** We leveraged ChatGPT to extract all entities related to the given title and subsequently remove them from the story.

- **StoryReplace:** We selected a story from different story conditions with the most similar story quality, as determined by the generative likelihood of the story without its condition.

- **Commonsense:** We utilized ChatGPT to modify stories with minimal adjustments to violate commonsense.

- **BlanderNarrative:** We requested ChatGPT to modify stories with minimal changes to render the narratives less interesting.

In some perturbations, it is possible to control the degree of perturbation. For instance, we can specify the percentage of words to shuffle in jumble. We set

Dataset	Condition	Story
ROC	[FEMALE] dad took me fishing .	we sat in a spot and waited for days ...
WP	tell me a story where the first line and last line ...	as i walked into the house , i was assailed by the smell of aging ...

Table 2: Sampled examples of given story condition and its generated story for each dataset.

Objective	Metric	FT	B/F	ST	MS
Similarity	BLEU	✗	B	✗	✗
	BERTScore	✗	B	✗	✗
	MoverScore	✗	B	✗	✗
Discriminative	UNION	✓	F	✓	✗
	MANPLTS	✓	F	✓	✗
	StoryER	✓	F	✓	✗
	CTC	✓	B&F	✗	✓
	UNIEVAL	✓	F	✗	✓
Generative	BARTScore	✗	B&F	✗	✓
	GPTScore	✗	F	✗	✓

Table 3: Statistics of Compared evaluation metrics. "FT" indicates whether the metric requires additional synthetic data to fine-tune on. "R-B/F" indicates whether the metric is reference-based (B) or reference-free (F). "ST" indicates whether the metric is originally designed for story evaluation. "MS" indicates whether the metric produces scores that consider multiple aspects.

the degrees to 0.4, 0.9, and 0.8 for typo, jumble, and antonym perturbations, respectively. In § 5.3, we investigate the effects of varying the perturbation degrees. However, due to space constraints, we only present a selection of perturbations in Table 1.

4 Meta-evaluation Experiments

In this section, we provide an overview of the meta-evaluation datasets used in our experiments as well as the evaluation metrics we compare DELTASCORE to.

4.1 Benchmarks and Compared Approaches

Benchmarks In our study, we leverage human evaluation scores from Xie et al. (2023), who conducted human evaluations on two widely-used story datasets: ROCStories (ROC) (Mostafazadeh et al., 2016), which contains concise stories that incorporate commonsense, and WritingPrompts (WP) (Fan et al., 2018), which focuses on fictional stories. The evaluations were conducted on five aspects of the stories: fluency, coherence, relatedness, logicality, and interestingness, rated on an ordinal scale from 1 (worst) to 5 (best). We provide examples of the stories in Table 2, while more detailed information about the datasets can be found in Appendix B.

Compared Evaluation Metrics We describe the evaluation metrics we compare in our experiments and classify them into three categories: similarity, discriminative, and generative, as shown in Table 3. While some of these metrics were not originally proposed for story evaluation, we adapted them to fit our settings. For discriminative metrics, we used the models provided by the authors without additional fine-tuning. We used the reference-free version of CTC and BARTScore to ensure consistency in comparisons, namely CTC (Consistency) and BARTScore ($\boldsymbol{c} \rightarrow \boldsymbol{s}$). Our preliminary experiments showed that reference-free versions perform better than reference-based ones. For UNIEVAL and GPTScore, we adapted the questions and prompts accordingly. More detailed information about the evaluation metrics can be found in Appendix C.

4.2 Meta-evaluation Results

In this study, we prioritize the story-level Kendall correlation as the evaluation metric due to the non-linear relationship between human and automatic evaluations and the ordinal scale used in human annotations (Kendall, 1938). We compare DELTASCORE, which is based on OPT (Zhang et al., 2022a) with three perturbations, to the evaluation metrics presented in Table 3. We select the best-performing metric from each category for comparison, and the results are shown in Figure 2. Further analysis of perturbation effects is presented in § 5.1 and model effects in § 5.2. Detailed information can be found in Appendix D.

Our findings suggest that similarity evaluation metrics such as BLEU, BERTScore, and MoverScore exhibit poor performance in open-ended story evaluation, which is consistent with previous studies (Guan and Huang, 2020; Xie et al., 2023). This implies that reference is not necessary in assessing open-ended stories. Interestingly, discriminative evaluation metrics do not demonstrate impressive results despite being designed specifically for story evaluation. This may be due to their extensive training on synthetic data and not on our data, which has different features that could affect their performance in our evaluation scenarios. Fi-

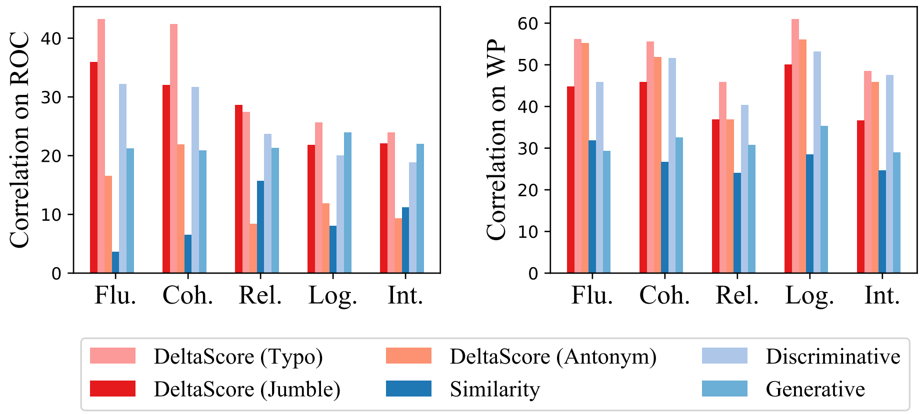


Figure 2: Story-level Kendall ($|\tau|$) correlations of different metrics on five aspects of ROC and WP. Similarity, Discriminative and Generative indicate the best performing evaluation metrics from such category. Higher bar indicates better performance.

Perturbation	**ROC**					**WP**				
	Flu.	**Coh.**	**Rel.**	**Log.**	**Int.**	**Flu.**	**Coh.**	**Rel.**	**Log.**	**Int.**
w/o perturbation	32.0	27.6	25.1	21.1	19.7	31.6	36.0	34.7	38.5	30.6
Typo	22.9	11.4	13.3	8.2	11.4	42.8	35.6	39.8	41.9	33.0
SubjVerbDis	11.9	10.5	8.8	4.3	0.9	5.8	6.2	17.9	13.2	17.3
Jumble	35.9	40.0	**33.8**	**29.6**	23.2	52.7	49.2	36.6	53.9	42.8
SentReorder	20.0	16.1	10.1	7.4	4.7	4.1	17.6	5.6	12.4	3.4
RmkeyEntities	15.9	16.9	13.3	9.6	7.6	36.5	27.7	31.7	33.5	21.9
StoryReplace	6.2	2.2	9.1	9.6	7.8	38.5	36.6	37.6	38.5	33.8
Antonym	14.3	22.1	10.9	21.3	16.8	41.5	39.5	33.9	42.7	36.6
Commonsense	16.9	15.8	17.5	2.2	3.8	29.6	39.0	26.3	37.6	26.5
BlanderNarrative	6.7	10.4	11.5	1.5	0.7	13.9	17.7	13.0	19.3	7.2

Table 4: Story-level Kendall correlation ($|\tau|$) between DELTASCORE with GPT-3.5 and human evaluations on ROC and WP. Highlight indicates the scores where DELTASCORE achieves better performance over applying generative likelihood from GPT-3.5 directly as the evaluation metric.

nally, we observe that DELTASCORE significantly outperforms other metrics, particularly with the jumble perturbation, achieving the best results in all aspects of both datasets.

5 More Analysis

5.1 Perturbations for Different Aspects

In this section, we investigate the performance of DELTASCORE with various perturbations that target different aspects of story quality. We employ GPT-3.5[1] here as it is the largest generative language model, and has demonstrated predominant performance in other tasks. We present the complete set of results in Table 4, and visualize the performance gains of DELTASCORE over GPT-3.5 in Figure 3.

Our analysis reveals that some perturbations do not improve the performance of DELTASCORE, despite their targeting of specific aspects of story quality. For example, sentenceReorder does not provide any improvement on DELTASCORE, possibly due to the focus of LLMs on local logic rather than the global narrative arc of a story. As such, reordering sentences, which does not affect the local logic, is not helpful in enhancing DELTASCORE. Furthermore, we observe that interestingness, the most subjective aspect, is the most challenging for DELTASCORE to improve, as designing a perturbation that targets it is difficult.

5.2 Extending to Other Generative LLMs

In order to test the generalization of our proposed approach, we extend it to other generative LLMs. In addition to the OPT and GPT-3.5 models that we demonstrated previously, we also include two representatives of encoder-decoder models, namely BART and FLAN-T5 (Chung et al., 2022), as well

[1]It is also called text-davinci-003, and the most capable model from GPT family that provides generative likelihood at the time of our experiments

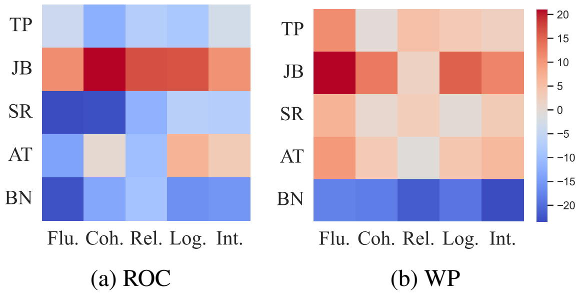


Figure 3: We present improvements of DELTASCORE on GPT-3.5 over applying GPT-3.5 generative likelihood directly as evaluation metric. Warmer color indicates greater improvement. 'TP' indicates typo, 'JB' indicates jumble, 'SR' indicates sentencereorder, 'AT' indicates antonym, and 'BN' indicates blandernarrative.

Arch.	Model	Size	#Data	Objectives
En-De	BART	406M	160GB	Denoising
	FLAN-T5	11B	-	Denoising
Cs-De	BLOOM	7B	366BT	LM
	OPT	66B	180BT	LM
	GPT-3.5	175B	300BT	LM

Table 5: Statistics of selected generative models classified by architecture (Arch.) to encoder-decoders (En-De) and Casual decoders (Cs-De). #Data indicates the pre-trained data scale (either in the number of tokens or storage size, 'BT' indicates "billion tokens"). Size indicates model parameters. LM indicates language modeling.

as one more casual decoder model, BLOOM (Scao et al., 2022).

In our preliminary experiment, we observed that larger models tend to demonstrate better performance within the same model framework. As such, we applied the largest model size that we could run for each model. The model details are presented in Table 5, while the results are shown in Table 6.

It has been observed that DELTASCORE, when applied with the three perturbations, demonstrates a strong performance across all generative models, irrespective of their varying model architectures. In general, it offers a better evaluation metric compared to directly applying likelihood. Among the three perturbations, jumble appears to be the most effective, providing substantial improvement in nearly all aspects across the two datasets. As a result, jumble yields the best-performing DELTASCORE for each aspect.

It should be noted that DELTASCORE appears to exhibit better performance on WP as compared to ROC. Our speculation is that this can be attributed to the fact that the stories in WP are significantly more intricate in nature, providing more opportunities for the perturbations to target a greater number of aspects. Consequently, this complexity can potentially provide a benefit to DELTASCORE.

We have made an interesting observation that larger models generally exhibit better performance, with GPT-3.5 and OPT being the highest performing models and BART being the lowest. BLOOM and FLAN-T5 are positioned in the middle of the spectrum. However, this trend is not always consistent, as evidenced by our findings. Specifically, in this study, OPT emerged as the top-performing model, being the best in 6 out of 10 aspects, while GPT-3.5 won only 2 out of 10 aspects, despite being 2.6× larger and trained on 1.7× more tokens of data. This observation aligns with previous research indicating that larger models do not necessarily guarantee better performance.

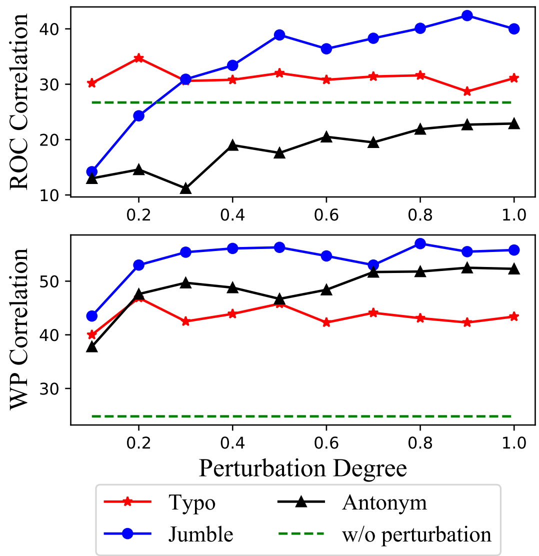


Figure 4: Effects of perturbation degrees on DELTASCORE with OPT.

5.3 Influence of Perturbation Degrees

In our study, we investigated the impact of perturbation degrees on DELTASCORE using OPT, as it demonstrated dominant performance in Table 6. We present the outcomes in Figure 4, and trends for specific perturbations can be observed. It appears that jumble and antonym do not offer significant improvement when the degree of perturbation is minor, but they exhibit a substantial increase in performance when more words are jumbled, ultimately

Metric	ROC					WP				
	Flu.	Coh.	Rel.	Log.	Int.	Flu.	Coh.	Rel.	Log.	Int.
DELTASCORE (with BLOOM 7B)										
w/o perturbation	26.4	21.3	21.1	15.8	20.3	18.1	15.4	17.5	17.6	9.3
Jumble	35.4	36.4	21.6	23.4	**26.8**	52.7	52.5	43.6	56.6	44.7
Typo	31.7	28.2	22.0	15.3	19.0	48.4	46.3	39.0	52.6	40.2
Antonym	11.8	16.1	6.5	9.6	7.7	49.9	46.3	32.5	50.0	42.7
DELTASCORE (with OPT 66B)										
w/o perturbation	32.0	26.7	25.3	22.0	23.5	20.8	24.8	22.2	25.7	17.1
Jumble	**43.2**	**42.4**	27.4	25.6	23.9	56.1	**55.5**	**45.9**	**60.9**	**48.5**
Typo	35.9	32.0	28.6	21.8	22.1	44.8	45.8	36.8	50.1	36.6
Antonym	16.5	21.9	8.4	11.9	9.3	55.2	51.8	36.9	56.0	45.9
DELTASCORE (with FLAN-T5 XXL 11B)										
w/o perturbation	21.0	22.1	26.8	23.2	22.5	13.6	10.6	12.1	10.8	4.2
Jumble	35.7	28.2	19.9	19.8	22.2	**59.2**	53.7	41.8	57.4	46.9
Typo	24.2	26.6	27.7	24.6	23.0	34.3	30.5	28.1	34.8	23.1
Antonym	9.4	16.1	3.2	7.5	5.4	52.6	48.0	37.1	53.8	46.3
DELTASCORE (with BART-large-cnn 406M)										
w/o perturbation	24.9	11.6	10.9	11.3	14.6	32.2	25.6	31.0	30.2	23.0
Jumble	29.2	17.7	9.9	12.2	12.6	49.1	40.1	44.5	46.5	41.5
Typo	22.9	11.4	13.3	8.2	11.4	42.8	35.6	39.8	41.9	33.0
Antonym	14.3	22.1	10.9	21.3	16.8	41.5	39.5	33.9	42.7	36.6
DELTASCORE (with GPT-3.5 175B)										
w/o perturbation	32.0	27.6	25.1	21.1	19.7	31.6	36.0	34.7	38.5	30.6
Jumble	35.9	40.0	**33.8**	**29.6**	23.2	52.7	49.2	36.6	53.9	42.8
Typo	29.5	28.2	22.7	15.4	11.7	40.7	43.1	32.6	47.2	35.4
Antonym	21.4	26.7	15.1	21.9	18.5	48.3	47.3	37.6	51.6	42.1

Table 6: Absolute value of Story-level Kendall correlation ($|\tau|$) between different metrics and human evaluations on ROC and WP. We **bold** the best scores in each aspect and we highlight the scores where DELTASCORE achieves better performance over the likelihood based metrics from the same model.

reaching a stable value. In contrast, typo appears to perform more reliably, regardless of the degree of perturbation. We hypothesize that the language model may not be highly sensitive to small changes in word order, given its masked word infilling objective during training. However, it consistently responds to typos in words, potentially due to the resulting difference in embeddings caused by tokenization.

6 Conclusion

This paper introduces DELTASCORE, a novel approach for evaluating fine-grained story quality by comparing the difference in likelihood between pre- and post-perturbations. Our experiments demonstrate that DELTASCORE produces stronger correlations with human evaluation results across all story quality aspects compared to current state-of-the-art evaluation metrics. We also show that DELTASCORE can be applied to various pre-trained language models, including encoder-decoder and causal-decoder models, outperforming the direct application of generative likelihood as an evaluation metric.

Limitations

Our work explores a limited set of perturbations for story evaluation, but it is likely that there are many more perturbations that could be obtained through different approaches. Although we only apply this perturbation method to story generation in this paper, it has the potential to be adapted for evaluating various aspects of text generation using specifically designed perturbations. This opens up a fruitful area for future research.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Informa-*

tion Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. Storyer: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1739–1753. Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé 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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5794–5836. International Committee on Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

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, Online. Association for Computational Linguistics.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021a. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393, Online. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021b. 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 (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1450–1470, Seattle, United States. Association for Computational Linguistics.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James R. Glass, and Yulia Tsvetkov. 2022. On the blind spots of model-based evaluation metrics for text generation. *CoRR*, abs/2212.10020.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9540–9561. Association for Computational Linguistics.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation. *CoRR*, abs/2203.11131.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. 2016. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):26:1–26:39.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, and et al Matthias Gallé. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. Can very large pretrained language models learn storytelling with A few examples? *CoRR*, abs/2301.09790.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4393–4479. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022b. Persona-guided planning for controlling the protagonist's persona in story generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3346–3361, Seattle, United States. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2023–2038. Association for Computational Linguistics.

A Perturbation Prompts

We use following prompts for perturbations.

- Relatedentities: Find all entities in the given story that is relevant to the given title. Please only print entities in the given story, and separate them by ','. "title": {title}, "story": {story}
- Commonsense: Revise the following story such that certain elements does not make sense. The revision should be minimal, e.g. by changing a few words. "story": {story}
- BlanderNarrative: Revise the following story to make it less interesting (e.g. expected ending, no plot twist). The revision should be minimal. "story": {story}

B Meta-evalution Datasets Details

Xie et al. (2023) evaluate a range of recently proposed story generation models: 1) **KGGPT2** (Guan et al., 2020), 2) **MTCL** (Xu et al., 2020),

3) **HINT** (Guan et al., 2021a), 4) **PROGEN** (Tan et al., 2021), 5) **BART** (Lewis et al., 2020) and 6) **GPT-3** (Brown et al., 2020), and also human reference stories. They conduct a meta-evaluation, which consists of two parts: crowdsourcing through Amazon Mechanical Turk and in-house evaluation by colleagues. We primarily rely on the crowdsourced evaluations since the authors suggest that in-house annotators may have a natural bias towards assigning higher scores to machine-generated stories.

C Compared Evaluation Metrics Details

BARTScore (Yuan et al., 2021) evaluates generated text as a text generation task by calculating the conditional likelihood of the generated text under BART. This metric can evaluate various aspects of text quality by using different combinations of input conditions and output. We use the reference-free version of BARTScore in the direction of $\boldsymbol{c} \rightarrow \boldsymbol{s}$.

CTC (Deng et al., 2021) evaluates generated text as an information alignment task, and they propose various approaches for alignment calculations. In our evaluation, we use the reference-free alignments, which is the "consistency" version of CTC.

UNIEVAL (Zhong et al., 2022) evaluates generated text as a question answering task, where different questions are asked for each aspect. The evaluation models are trained on text summarization and dialogue generation tasks and are shown to have zero-shot transfer ability to data-to-text generation task by asking corresponding new questions.

UNION (Guan and Huang, 2020) frames story evaluation as a classification task by constructing negative samples of original stories using heuristic rules, and training a neural discriminator to differentiate them. The trained model is then used for evaluating new stories. While their work focuses on the ROC and WP datasets, we do not fine-tune the UNION model on our own dataset.

MANPLTS (Ghazarian et al., 2021) is an extension of UNION that focuses on constructing negative samples that are more similar to machine-generated stories. They do this by manipulating storylines and generating stories based on these manipulated storylines using a story generation model. Like UNION, we use the models provided by authors.

StoryER (Chen et al., 2022) combines the approach of separating original stories from negative samples, as in UNION and MANPLTS, with considering human preference by training the model to differentiate highly-upvoted stories from lowly-upvoted ones. In addition to the preference score, they train their model to produce ratings and comments on predefined story aspects to provide explanations for the evaluation.

We also incorporate conventional similarity-based metrics, including: **BLEU** (Papineni et al., 2002), which measures n-gram overlaps between stories and human references, **BERTScore** (Zhang et al., 2020), and **MoverScore** (Zhao et al., 2019), which measures semantic similarity from embeddings obtained from BERT (Devlin et al., 2019).

C.1 UNIEVAL and GPTScore

UNIEVAL We ask the following questions for each aspect. Note that we try to use the narrative/vocabulary as close to the original questions Zhong et al. (2022) use in their efforts as possible.

- Fluency: Is this a fluent utterance?
- Coherence: Is this a coherent utterance?
- Relatedness: Is this claim consistent with the document?
- Logicality: Is this utterance consistent with the commonsense?
- Interestingness: Is this an interesting utterance?

GPTScore We ask the following questions for each aspect. Note that we try to use the narrative/vocabulary as close to the original questions (Fu et al., 2023) use in their efforts as possible.

- Fluency: Generate a fluent story for the given title: {title}, and story: {story}
- Coherence: Generate a coherent story for the given title: {title}, and story: {story}
- Relatedness: Generate a story related to the given title: {title}, and story: {story}
- Logicality: Generate a story that adhere to commonsense for the given title: {title}, and story: {story}
- Interestingness: Generate an interesting story for the given title: {title}, and story: {story}

D Correlations

Metric	ROC					WP				
	Flu.	**Coh.**	**Rel.**	**Log.**	**Int.**	**Flu.**	**Coh.**	**Rel.**	**Log.**	**Int.**
Similarity Based Evaluation Metrics										
BLEU	3.4	4.4	0.8	4.6	0.4	13.4	11.4	9.6	11.2	7.2
BERTScore	3.5	5.0	14.0	5.7	7.3	31.8	26.7	24.0	28.5	24.6
MoverScore	3.6	6.5	15.7	8.0	11.2	16.4	17.2	20.1	20.7	23.7
Story Evaluation Metrics										
UNION	0.7	12.7	12.8	0.8	3.9	17.4	21.9	18.1	22.9	21.8
MANPLTS	21.3	32.8	23.2	14.7	12.4	1.1	5.5	1.7	4.6	6.7
StoryER	6.5	4.5	4.0	3.7	9.7	15.9	13.1	14.1	17.9	26.1
Unified Evaluation Metrics										
CTC	22.9	27.3	14.3	11.1	8.3	45.9	51.6	40.3	53.1	47.5
UNIEVAL	32.2	31.7	23.7	20.0	18.8	39.3	41.3	38.6	50.7	39.1
GPTScore	21.2	20.9	21.3	23.9	22.0	29.3	32.6	30.8	35.3	28.9
DELTASCORE (with BLOOM 7B)										
w/o perturbation	26.4	21.3	21.1	15.8	20.3	18.1	15.4	17.5	17.6	9.3
Jumble	35.4	36.4	21.6	23.4	**26.8**	52.7	52.5	43.6	56.6	44.7
Typo	31.7	28.2	22.0	15.3	19.0	48.4	46.3	39.0	52.6	40.2
Antonym	11.8	16.1	6.5	9.6	7.7	49.9	46.3	32.5	50.0	42.7
DELTASCORE (with OPT 66B)										
–	32.0	26.7	25.3	22.0	23.5	20.8	24.8	22.2	25.7	17.1
Jumble	**43.2**	**42.4**	27.4	25.6	23.9	56.1	**55.5**	**45.9**	**60.9**	**48.5**
Typo	35.9	32.0	28.6	21.8	22.1	44.8	45.8	36.8	50.1	36.6
Antonym	16.5	21.9	8.4	11.9	9.3	55.2	51.8	36.9	56.0	45.9
DELTASCORE (with FLAN-T5 XXL)										
w/o perturbation	21.0	22.1	26.8	23.2	22.5	13.6	10.6	12.1	10.8	4.2
Jumble	35.7	28.2	19.9	19.8	22.2	**59.2**	53.7	41.8	57.4	46.9
Typo	24.2	26.6	27.7	24.6	23.0	34.3	30.5	28.1	34.8	23.1
Antonym	9.4	16.1	3.2	7.5	5.4	52.6	48.0	37.1	53.8	46.3
DELTASCORE (with BART-large-cnn)										
w/o perturbation	24.9	11.6	10.9	11.3	14.6	32.2	25.6	31.0	30.2	23.0
Jumble	29.2	17.7	9.9	12.2	12.6	49.1	40.1	44.5	46.5	41.5
Typo	22.9	11.4	13.3	8.2	11.4	42.8	35.6	39.8	41.9	33.0
Antonym	14.3	22.1	10.9	21.3	16.8	41.5	39.5	33.9	42.7	36.6
DELTASCORE (with GPT-3.5)										
w/o perturbation	32.0	27.6	25.1	21.1	19.7	31.6	36.0	34.7	38.5	30.6
Jumble	35.9	40.0	**33.8**	**29.6**	23.2	52.7	49.2	36.6	53.9	42.8
Typo	29.5	28.2	22.7	15.4	11.7	40.7	43.1	32.6	47.2	35.4
Antonym	21.4	26.7	15.1	21.9	18.5	48.3	47.3	37.6	51.6	42.1

Table 7: Absolute value of Story-level Kendall correlation ($|\tau|$) between different metrics and human evaluations on three CoudSourcing datasets. We **bold** the best scores in each aspect and we highlight the scores where DELTASCORE achieves better performance over the likelihood based metrics from the same model.