Learning Personalized Alignment for Evaluating Open-ended Text Generation

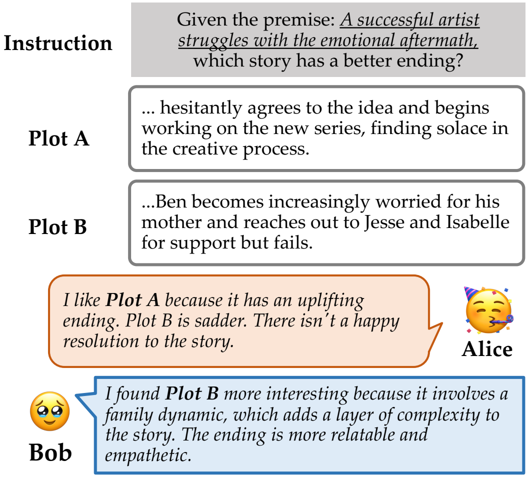
Anonymous ACL submission

Abstract

With rapid progress made in language qualities such as fluency and consistency via large language models (LLMs), there has been increasing interest in assessing alignment with diverse human preferences. Traditional metrics heavily rely on lexical similarity with human-written references and have been observed to suffer from a poor correlation with human evaluation. Furthermore, they ignore the diverse preferences of humans, a key aspect in evaluating open-ended tasks like story generation. Inspired by these challenges, we introduce an interpretable open-ended evaluation framework **PERSE** to assess the alignment with a specific human preference. It is tuned to deduce the specific preference from a given personal profile and evaluate the alignment between the generation and the personal preference. **PERSE** also explains its assessment by a detailed comment or several fine-grained scores. This enhances its interpretability, making it more suitable to tailor a personalized generation. Our 13B LLaMA-2-based **PERSE** shows a 15.8% increase in Kendall correlation and a 13.7% rise in accuracy on zero-shot reviewers compared to GPT-4. It also outperforms GPT-4 by 46.01% in the Kendall correlation on new domains, indicating its transferability.[1]

1 Introduction

Large language models (LLMs) have recently shown impressive generative capability in many generation tasks, gaining rapid improvement in language qualities such as fluency and consistency (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023). However, evaluating their performance in open-ended generation tasks is still challenging because of the diversity of the responses. Traditional automatic metrics suffer from the one-to-many problem in open-ended generation (Liu et al., 2016) and have shown poor correlation with human judgment (Krishna et al., 2021; Guan et al., 2021). Recently some studies have trained evaluation metrics on human ratings to better approximate human judgments (Sellam et al., 2020; Rei et al., 2020). However, these metrics mainly focus on objective qualities and ignore subjective assessment, such as surprise (Chhun et al., 2022) or interestingness (Bae et al., 2021).



Given the premise: <u>A successful artist struggles with the emotional aftermath,</u> which story has a better ending?

... hesitantly agrees to the idea and begins working on the new series, finding solace in the creative process.

...Ben becomes increasingly worried for his mother and reaches out to Jesse and Isabelle for support but fails.

*I like **Plot A** because it has an uplifting ending. Plot B is sadder. There isn't a happy resolution to the story.*

*I found **Plot B** more interesting because it involves a family dynamic, which adds a layer of complexity to the story. The ending is more relatable and empathetic.*

Figure 1: Two human reviewers have distinct preferences of LLM-generated stories from the same premise.

The subjective evaluation metrics are highly affected by diverse human preferences. For example, Figure 1 demonstrates two stories generated by Yang et al. (2023) from the same premise. Alice prefers Plot A for its uplifting ending while Bob favors Plot B because of the plot complexity and empathetic ending. This underscores the importance of an automatic personalized evaluation metric that can assess model generation based on different preferences. However, it is costly for each reviewer to provide a large number of personalized examples to demonstrate their preferences, making it infeasible to train a separate evaluation model for each reviewer and generalize the existing metric to unseen reviewers.

[1] Both datasets and code will be released.

Furthermore, the subjectiveness also makes the evaluation score more difficult to understand. AuPEL (Wang et al., 2023) introduces personalization as one of the evaluation aspects to compare two inputs without any explanation. The lack of transparency hinders the trustworthiness and reliability of evaluation and makes it difficult to assist in the development of generative models (Leiter et al., 2022). Thus, the key challenge of evaluation from a personalized aspect is how to model an unseen reviewer's preference from the limited annotated personalized context and give an interpretable explanation for its assessment.

In this paper, we introduce an LLM-based evaluation model (**PerSE**) to assess the alignment between the open-ended generation and a specific preference. **PerSE** is tuned to infer the preference from a limited-length profile and use this preference to evaluate the given generation. For the pointwise evaluation that takes a single input, **PerSE** provides an overall score along with an explanation. For pairwise comparison, it provides fine-grained scores on several aspects to interpret the alignment. We collect the different responses from various users for the same query to construct personalized instruction data. We fine-tune **PerSE** from LLaMA-2 (Touvron et al., 2023) on the personalized data to enhance their capability to infer preferences from the personal context. Compared with GPT-4, **PerSE** achieves a 15.8% higher Kendall correlation in pointwise evaluation of movie plot generation and a 13.7% higher accuracy in comparative evaluation of story generation on zero-shot reviewers. It also outperforms GPT-4 by 46.01% in the Kendall correlation on new domains when zero-shot transfer to other domains. Our contributions can be summarized as below:

- We develop an LLM-based evaluation model **PerSE** to assess the alignment between the open-ended generation and an in-context preference. By instruction-tuning on personalized data, **PerSE** significantly outperforms GPT-4 in evaluating the personal alignment.
- **PerSE** provides a detailed explanation for its assessment, which is a comment on the pointwise evaluation and fine-grained scores on the pairwise comparison. The interpretability of **PerSE** makes it more suitable as a guidance of personalized generation.
- We find that LLMs after reinforcement learning via human feedback tend to be less personalized and more cautious with negative comments. This makes them struggle with aligning strong personal preferences. However, when instruction-tuned with personalized data, even weak LLMs can show better performance in aligning with preference.

2 Related Work

Evaluation Metrics for Text Generation Many automatic metrics can be briefly divided into reference-based and reference-free metrics. Reference-based metrics evaluate the similarity between the reference and the model output based on lexical overlap (Papineni et al., 2002; Lin, 2004) or embedding distance (Zhang et al., 2019; Zhao et al., 2019). Meanwhile, reference-free metrics directly measure the quality of the model output without any reference. Usually, they are trained to evaluate generation from an overall perspective (Guan and Huang, 2020; Ghazarian et al., 2021) or along multiple axes (Chen et al., 2022; Xie et al., 2023). Recently, researchers have explored using large language models in evaluation metrics, such as GPTScore (Fu et al., 2023), GEMBA (Kocmi and Federmann, 2023), and InstructScore (Xu et al., 2023). However, these metrics mainly focus on objective qualities, where pre-trained language models have achieved great performance. In this paper, we explore LLM-based evaluators to evaluate how the generation aligns with personal preferences.

Human Evaluation for Generation Human evaluation is also used to evaluate different aspects of text quality, such as coherence (Xu et al., 2018; Peng et al., 2018), relevance (Yang et al., 2023, 2022; Jhamtani and Berg-Kirkpatrick, 2020), interestingness (Bae et al., 2021) and so on. To comprehensively cover all aspects, Chhun et al. (2022) suggested 6 human criteria for the story: relevance, coherence, empathy, surprise, engagement, and complexity. However, they showed that the inter-annotator agreement of human evaluation on these subjective aspects is low. Karpinska et al. (2021) also highlighted the perils of crowdsourced human judgments from Amazon Mechanical Turk due to under-qualified workers and lacking reproducibility details.

Personalization in Text Generation and Evaluation Personalization has been well studied in many recommendation systems (Das et al., 2007; Xu et al., 2022) and search applications (Croft et al.,

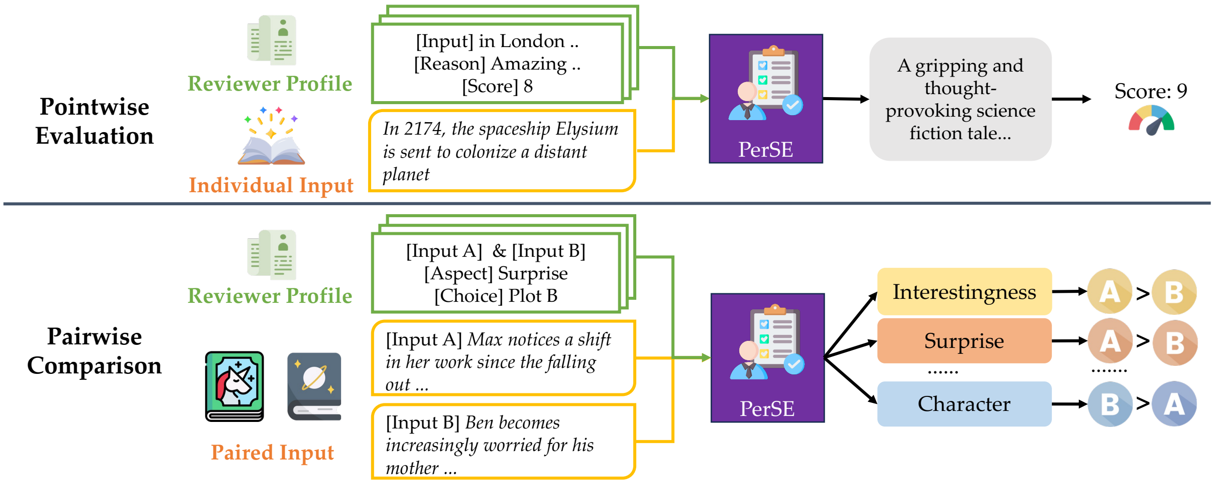


Figure 2: **PerSE** for the pointwise evaluation and pairwise comparison. The reviewer's preference is inferred from their prior reviews. The reviewer profiles $\boldsymbol{c}_u$ are in green and the input $\boldsymbol{x}$ is in orange.

2001; Shi et al., 2023). Recently, researchers have also highlighted its importance in natural language processing (Flek, 2020; Dudy et al., 2021). Several recent studies have investigated LLMs' capabilities in capturing personalization (Chen et al., 2023; Kang et al., 2023; Salemi et al., 2023) or prompting for personalized recommendations (Lyu et al., 2023; Chen, 2023; Li et al., 2023). Wang et al. (2023) introduces personalization score as one of the evaluation aspects and uses LLMs as evaluators. In this paper, we propose an interpretable evaluation model to align personal preference, which not only outputs an assessment but also a detailed explanation.

3 PerSE: Personalized Evaluation Model

We propose an LLM-based evaluation model for assessing the alignment between the generation and personal preference. **PerSE** delivers an interpretable evaluation from a particular reviewer's viewpoint without knowing the gold review.

Problem Formulation For the reference-free evaluation, the quality $\boldsymbol{y}$ of the input $\boldsymbol{x}$ is assessed from the perspective of a specific reviewer u. The evaluation model learns the mapping of $\mathrm{M}(\boldsymbol{x}, u) \rightarrow \boldsymbol{y}$. For the interpretable evaluation, an optional explanation $\boldsymbol{e}$ of the score is given. In this paper, we define the reviewer's profile $\boldsymbol{c}_u = \{(\boldsymbol{x}_u^{(1)}, \boldsymbol{e}_u^{(1)}, \boldsymbol{y}_u^{(1)}), (\boldsymbol{x}_u^{(2)}, \boldsymbol{e}_u^{(2)}, \boldsymbol{y}_u^{(2)}), \cdots\}$ as a series of annotated reviews. $\boldsymbol{x}_u^{(i)}$ and $\boldsymbol{y}_u^{(i)}$ are the i-th input and its score of the reviewer u, and $\boldsymbol{e}_u^{(i)}$ is the explanation from the reviewer[2].

As demonstrated in Figure 2, **PerSE** can provide a personalized review for the individual and pairwise evaluations. $\boldsymbol{x}$ is a single input in the point-wise evaluation while it is the concatenation of the paired input in the pairwise evaluation. Inspired by human evaluation in the open-ended generation, the explanation can be a detailed reason for an overall rating or several fine-grained scores on different aspects. Therefore, for the pointwise evaluation that outputs an overall score, we use the review comment as the explanation of the score. For the pairwise comparison, we add several fine-grained aspects to make the comparison more interpretable.

For pointwise evaluation, the output $\boldsymbol{y}$ is post-processed to a numerical score and then calibrated to 1 to 10. We use chain-of-thought (Wei et al., 2022) to make **PerSE** first generate the comment $\boldsymbol{r}$ and then output the score $\boldsymbol{y}$ based on it, which can be denoted by $\boldsymbol{y} = \mathrm{M}(\boldsymbol{x}, \boldsymbol{c}_u, \boldsymbol{e_r})$ and $\boldsymbol{e_r} = \mathrm{M}(\boldsymbol{x}, \boldsymbol{c}_u)$. For pairwise comparison, we pre-define a set of multiple aspects $\mathcal{A}$ based on the natural of open-ended generation. For each aspect $\boldsymbol{e_a} \in \mathcal{A}$, we add the aspect to the instruction and prompt the model with $\boldsymbol{y} = \mathrm{M}(\boldsymbol{x}, \boldsymbol{c}_u, \boldsymbol{e_a})$. By modeling the aspect in the prompt, we can use one unified model for all aspects and it can easily generalize to more aspects.

LLM-based Evaluation We reformulate the evaluation of personalized alignment as a generative task and prompt the model to generate the assessment $\boldsymbol{y}$ and the explanation $\boldsymbol{e}$ of the input $\boldsymbol{x}$ based on the review profile $\mathbf{c}_u$, where $\boldsymbol{e}$ is a review text $\boldsymbol{e_r}$ in the pointwise evaluation and

[2] To simplify, We assume the reviewer's preferences are consistent within the review time frame.

the fine-grained aspect $\boldsymbol{e_a}$ in pairwise comparison. Therefore, it can be represented as: $\mathrm{M}(\boldsymbol{x}, \boldsymbol{c}_u, \boldsymbol{e}) = \prod_{t=0}^{T} p(y_t|\rho(\boldsymbol{x}, \boldsymbol{c}_u, \boldsymbol{e}), \boldsymbol{y}_{<t})$. T is the length of generated evaluation $\boldsymbol{y}$. ρ is a prompt template that maps the reviewer's profile, input, and explanation $\boldsymbol{e}$ into a single instruction.

Personalized Instruction Tuning We construct instruction data to align the evaluation with personal preference. For reviewers in the training set $\mathcal{U}_{tr}$, we create preferences from two non-overlapped query spaces: $\mathcal{X}_{prior}$ and $\mathcal{X}_{tr}$. $\mathcal{X}_{prior}$ is used to build the user profile while $\mathcal{X}_{tr}$ is used as the new queries for finetuning. We create user profile $\boldsymbol{c}_u = \{(\boldsymbol{x}, \boldsymbol{a}_u, \boldsymbol{y}_u)|\boldsymbol{x} \in \mathcal{X}_{prior}\}$. To control the length of the user profile, we use a subset $\boldsymbol{c}_{u_k}$ as the limited-length profile by randomly sampling k reviews from $\boldsymbol{c}_u$. Therefore, the personalized dataset can be represented as $D_k = \{\{\boldsymbol{x}, \boldsymbol{c}_{u_k}, \boldsymbol{y}_u\}|x \in \mathcal{X}_{tr}, u \in \mathcal{U}_{tr}, \boldsymbol{c}_{u_k} \subseteq \boldsymbol{c}_u\}$. The training objective is $L = -\sum_{t=0}^{T} \log p(y_t|\rho(\boldsymbol{x}, \boldsymbol{c}_u), \boldsymbol{y}_{<t})$. Our dataset construction and the prompts we used are detailed in Appendix A.2.

4 Experiment Setup

To align the LLMs with specific personal preferences, we create several personal instruction data and fine-tune the LLaMA-2 chat version on it. We investigate the influence of contamination in LLM-based evaluation (Appendix A.1) and reproduce two personalized datasets from the existing dataset to alleviate the contamination (Appendix A.2). We list the brief introduction of the two datasets in Section 4.1. In Section 4.2 we describe the implementation of **PerSE** and several baselines. More details on the training can be found in Appendix B.

4.1 Datasets

Per-MPST We modify the movie review dataset MPST (Kar et al., 2018, 2020) for personalization. Each review includes a review text and a score from 1 (lowest) to 10 (highest). We anonymize the character and location names in the raw story and summarize it to alleviate the influence of the contamination issue. We then group reviews by reviewer ID and remove reviewers that have fewer than 6 reviews. For each reviewer, we sample different numbers of the prior reviewers ($k = 1$ to 5) as the profile. Due to the limited context length of LLaMA-2, we limit the maximum length of the prompt to 2500 words (about 4k tokens).

Per-DOC We use human evaluation results on system generated stories from Yang et al. (2023). There are 7000 unique examples from 403 annotators. Each example consists of two plots generated from the same premise. The annotators were asked to answer various questions and choose their preferred plot for each question. We derive five subjective aspects from the original questions: `Interestingness` (I), `Adaptability` (A), `Surprise` (S), `Character Development` (C), and `Ending` (E). `Interestingness` focuses on the appeal of the overall narrative; `Surprise` indicates unexpected elements or twists in the plot; `Character development` evaluates the emotional and personal connection between characters and events; `Ending` is about satisfaction or appreciation of the ending, and `Adaptability` measures the probability of further developing the story. We use the worker ID to cluster the annotations. Similarly, we removed annotators with fewer than 2 annotations. We keep $k = 1$ for the reviewer profile due to the length limitation.

We split the dataset into training and validation by 9:1 based on the identification of reviewers. The training set is used to create the personalized instruction data, while the validation set is used for inference. Reviewers in the validation set are unseen during the finetuning phase. The model is required to infer the preference of a zero-shot reviewer and evaluate based on this preference.

4.2 Experimental Setting

We implement **PerSE** based on LLaMA-7b-chat and LLaMA-13b-chat, tuning them on the personalized instruction data created from the training set of Per-MPST and Per-DOC. In our main experiments, we use $k = 3$ for Per-MPST and $k = 1$ for Per-DOC. For inference, we set the temperature to 0.8 and limit the maximum generation length to 600. We report Pearson, Spearman, and Kendall-Tau correlation coefficients to measure the agreement between human scores and the generated scores for each content-reviewer pair $(\boldsymbol{x}, u)$ in point-wise evaluation. For comparative evaluation, we view each aspect as a binary classification and report the accuracy for the (content, reviewer, aspect) tuple.

Baseline We set up a simple baseline that directly uses the average scores from prior reviews as the prediction. For Per-DOC, since we only have one comparison in the instruction ($k = 1$), we directly use this answer as the output. This baseline is named as **Reviewer Avg.**. On Per-MPST, we add baseline matrix factorization (MF) (Koren et al.,

Table 1: Statistics of Per-MPST and Per-DOC. Length is the number of words in the instruction, which includes the instruction template, reviewer preference, and plot query. **I**, **A**, **S**, **C**, and **E** stand for `Interestingness`, `Adaptability`, `Surprise`, `Character Development`, and `Ending`. k is the number of reviews; we fix $k = 1$ for Per-DOC due to the length.

		Per-MPST					**Per-DOC** ($k = 1$)				
		k=1	**k=2**	**k=3**	**k=4**	**k=5**	**I**	**A**	**S**	**C**	**E**
Train	# Reviewers	1412	1394	1385	1369	1336	172	171	156	160	155
	# Example	13254	13940	13794	13480	12041	1985	1856	1722	1785	1574
	Avg. Length	868.9	1235.2	1600.3	1964.0	2123.3	2410.9	2413.7	2411.7	2409.8	2409.6
Valid	# Reviewers	92	92	92	92	92	18	18	15	18	15
	# Example	915	920	920	906	833	234	224	161	162	173
	Avg. Length	857.9	1237.1	1597.2	1956.1	2108.4	2402.9	2399.2	2408.4	2421.4	2404.3

2009), which is commonly used in recommendation systems. The main idea is to recommend products based on the similarity of the user and the product. These two baselines do not have an interpretable explanation for their evaluation. On Per-DOC, both plot pairs and the annotators of the validation set have no overlapping with the training set, so the matrix factorization cannot apply to this setting. We also evaluate the zero-shot capability of LLMs, including the pre-trained LLaMA-2-chat from 7b to 70b and GPT-4, with the same prompts and generation configurations.

Table 3: The comparison of the generated review and the human-written review on Per-MPST. A higher score indicates a better alignment between the generation and the human reference. The reviews generated by **PERSE** are more similar to the human-written reviews.

	BLEU	**ROUGE**	**BERTScore**	**BARTScore**
LLaMA-7b	2.213	0.253	0.829	-9.049
LLaMA-13b	2.847	0.262	0.833	-9.228
LLaMA-70b	3.014	0.256	0.832	-8.538
GPT-4	3.040	0.252	0.831	-6.853
PERSE-7b	3.988	0.292	**0.834**	-6.741
PERSE-13b	**4.108**	**0.294**	**0.834**	**-6.577**

5 Results and Analysis

We demonstrate the pointwise evaluation in the individual setting on Per-MPST and the pairwise comparison in the pairwise personalized setting on Per-DOC.

5.1 Main Results

Table 2: Pearson, Spearman, and Kendall correlations with human ratings for each $(\boldsymbol{x}, u)$ pair on Per-MPST. We use three reviews ($k = 3$) to represent reviewers' preferences. All results have a p-value less than 0.05. **PERSE**-7b is comparable to GPT-4 and **PERSE**-13b significantly outperforms GPT-4.

	Pearson	**Spearman**	**Kendall**
Reviewer Avg.	0.301	0.302	0.230
Matrix Factorization	0.308	0.313	0.269
LLaMA-2-7b	0.146	0.117	0.094
LLaMA-2-13b	0.172	0.182	0.147
LLaMA-2-70b	0.214	0.232	0.181
GPT-4	0.315	0.312	0.253
PERSE-7b	0.307	0.329	0.263
PERSE-13b	**0.345**	**0.368**	**0.293**

Pointwise Evalution As shown in Table 2, **PERSE**-13b significantly outperforms all baselines on correlations with unseen reviewers, and **PERSE**-7b is comparable to GPT-4. In particular, **PERSE**-13b achieves a typical high 0.345 Pearson correlation between its predictions and human scores, indicating that our model effectively captures the reviewer's preference from the given reviews. On the other hand, the results show that it is difficult for LLMs to directly infer the reviewer's preference without instruction-tuning. All LLaMA-2 baselines underperform the traditional baselines such as average or MF. This observation is consistent with Kang et al. (2023) who show that pre-trained LLMs struggle to understand reviewers' preferences and use them for a personalized score. However, the traditional baselines lack an interpretable review to explain their decision, hindering their applications in analyzing detailed model performance. We believe one possible reason is that both the pre-training phase and RLHF are aligning the model towards more objective and common human values, hindering personalization. This is also observed by Kirk et al. (2023) who claims that the aggregate fine-tuning process may not well represent all human preferences and values. However, we observe that with targeted instruction-tuning on only a few training data, LLMs can effectively infer personalized preferences and align with them.

One of the disadvantages of the score-based evaluation is that they do not provide an interpretable explanation for their scores, such as the

Table 4: Fine-grained prediction accuracy for each $(\boldsymbol{x}, u, \boldsymbol{a})$ on Per-DOC with $k = 1$. **PerSE**-7b and **PerSE**-13b were trained on all aspects. **PerSE** outperforms all baselines in all aspects. The p-value for t-test are smaller than 0.05.

	Interestingness	**Adaptability**	**Surprise**	**Character**	**Ending**	**Average**
Reviewer Avg.	0.466	0.478	0.460	0.469	0.515	0.477
LLaMA-2-7b	0.466	0.491	0.453	0.481	0.503	0.479
LLaMA-2-13b	0.422	0.451	0.477	0.481	0.517	0.470
LLaMA-2-70b	0.517	0.507	0.431	0.505	0.545	0.501
GPT-4	0.502	0.496	0.596	0.506	0.543	0.529
PerSE-7b	0.572	0.565	**0.619**	0.565	0.560	0.576
PerSE-13b	**0.621**	**0.570**	0.616	**0.607**	**0.597**	**0.602**

simple baseline and MF method. The lack of transparency makes the assessment less reliable. Thus, we further investigate how the interpretable reviews generated by **PerSE** align with the human-written reviews. We present the results on Per-MPST in Table 3. We use two common lexical-similarity-based metrics, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), and two model-based metrics, BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021) to measure the generation quality [3]. We can see that **PerSE**-7b and **PerSE**-13b outperform other baselines on all metrics. It indicates that **PerSE** can better model the preference of a specific reviewer and generate a personalized review from this perspective.

Pairwise Comparison We present the accuracy in Table 4. Our **PerSE** achieves the best performance in all aspects. Compared to **PerSE**-13b, **PerSE**-7b achieves comparable performance on `Surprise` but lags behind on other aspects. For baselines, the pre-trained LLaMA only achieved comparable performance with the simple baseline, with around 50% accuracy on most aspects. One possible reason for the poor performance is that we only have $k = 1$ review due to the context length limitation, making it more difficult to capture the preference. Meanwhile, GPT-4 does better in capturing `Surprise` than other LLM baselines but does not show advantages in other aspects.

5.2 Analysis

Here, we show some additional experiments to investigate personalization modeling in LLMs. More experiments are in Appendix C.

PerSE achieves a higher correlation with more reviews. We explore how many reviews are required to establish the reviewer's preference in Figure 3. For **PerSE**-7b and **PerSE**-13b, we train the models on different subsets of Per-MPST as shown in Table 1. $k = 0$ indicates that there are no personalized examples in the instruction, which is a baseline for evaluation without personalization. We randomly selected a score between 1 to 10 for the simple baseline for $k = 0$. The poor performance on $k = 0$ for all baselines suggests that an overall score does not work for evaluation. When we increase the number of reviews, it is easier for **PerSE**-13b to capture the reviewer's preference. However, for weaker baselines such as pre-trained LLaMA-2, they fail to benefit from more reviews. Furthermore, the simple average baselines also drop after 4 reviews. This indicates that although more reviews provide more information about the reviewer, it also increases the complexity of the context and may introduce noise. Therefore, if not limited by the context length, we suspect that the performance of **PerSE**-13b will also drop after achieving its maximum capability of inferring from complicated context with potential noise.

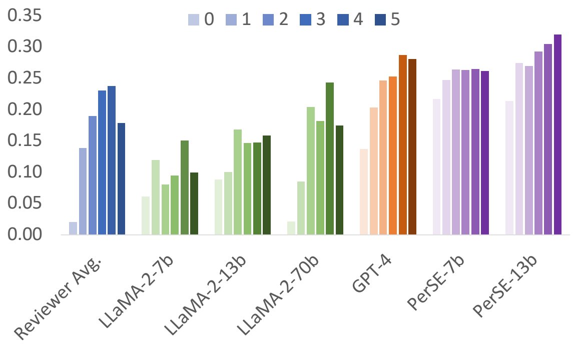


Figure 3: Kendall correlation on Per-MPST with different numbers of reviews (k) in reviewer history. Having more reviews benefits **PerSE**-13b, but the increased complexity may harm the performance of LLaMA.

More reviews improve the robustness of PerSE. Previous studies have shown that large language models are sensitive to the example order (Lu et al., 2022). Moreover, the assumption that the prefer-

[3] We use ROUGE-1 here. BARTScore is negative because it uses the average log-likelihood of the fine-tuned BART as the score.

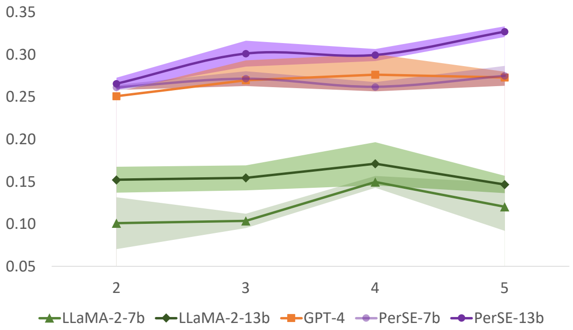


Figure 4: Kendall correlation on Per-MPST with different orders of reviews. The shadow indicates the variance while the line is the average performance among three trials. **PerSE** is more stable than baselines.

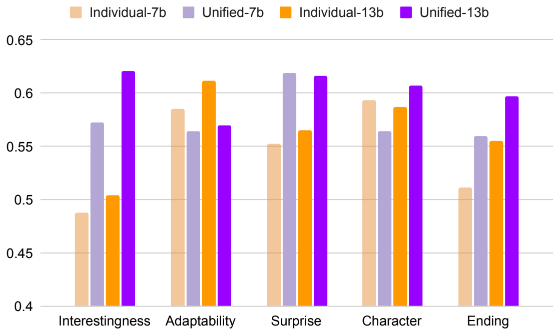


Figure 5: Accuracy of the unified and individual models on Per-DOC. Unified training improves performance.

ence is constant during these reviews may not hold in a real scenario. Therefore we randomly shuffle the reviews and test for three times to mitigate the potential influence of example order. We demonstrate the average performance with lines and the standard deviation by shadow regions in Figure 4. We can see that **PerSE**-13b stably outperforms other baselines on average. Furthermore, more reviews increase the robustness of **PerSE** to the order change, indicated by a smaller shadow region. It shows **PerSE** successfully captured the implicit reviewer's preference from these reviews. In contrast, the pre-trained LLaMA-2 are sensitive to the order, with a larger variance shadow.

Joint training benefits the individual aspects. We investigate the influence of joint training of different aspects on Per-DOC by training an individual model on each aspect and comparing the performance. As illustrated in Figure 5, the performance in most aspects is enhanced by the joint training, where the models are exposed to more data, i.e., different aspects can benefit each other. For example, the performance of capturing `Interestingness` and `Surprise`, and evaluation of the quality of `Ending` are weaker under the individual setting, but are enhanced by other aspects during the joint training, resulting in significant improvement. For separate models, they are better at capturing the preference for `Adaptability` and `Character Development`. We hypothesize that these two aspects are related to the setting of the plot, which is more structured. This may lead to a clearer preference that is easier to capture with single-aspect data.

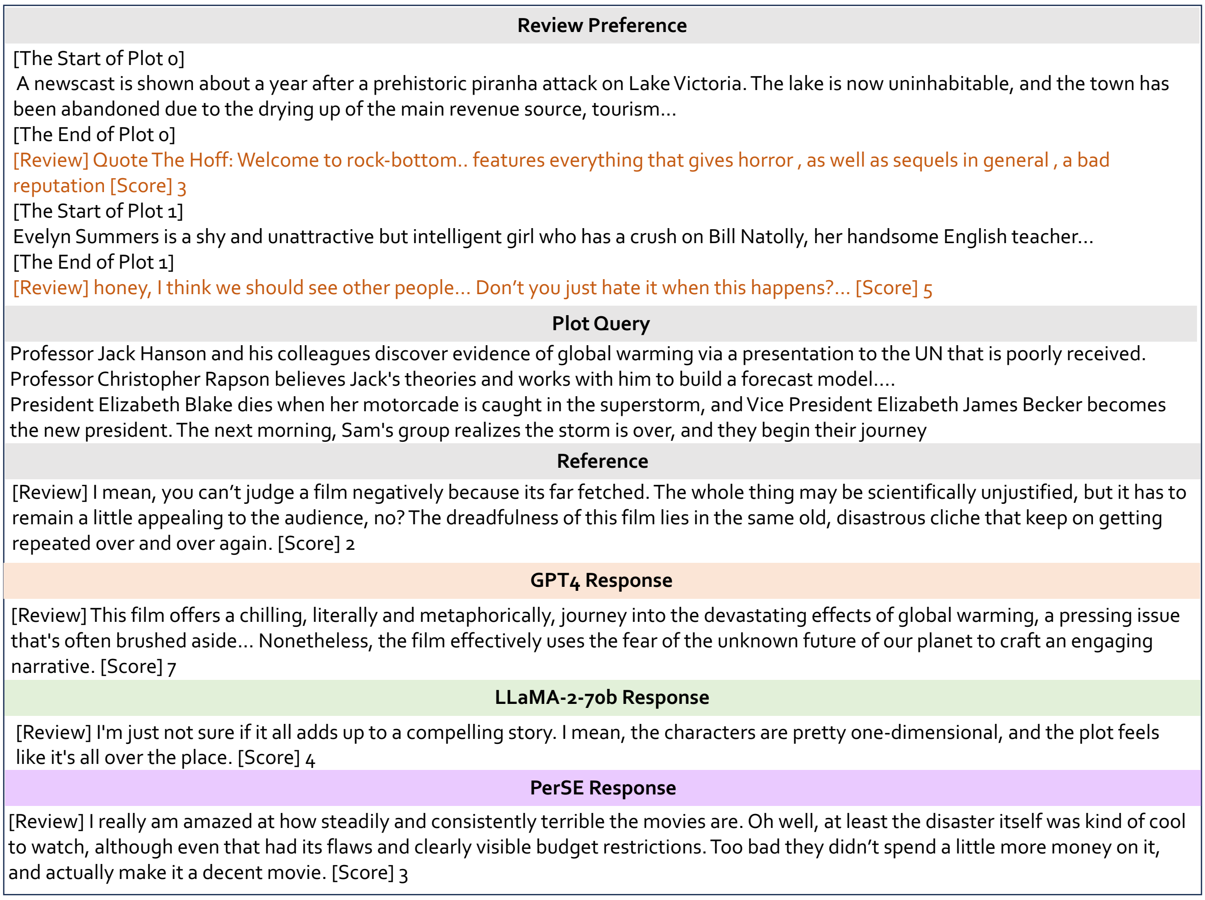
PerSE shows great generalization on other domains and language models We evaluate the generalization and transferability of **PerSE** by applying it to the new domain (Amazon book review[4]) in a zero-shot manner. We use the review-enhanced setting of **PerSE**, which was fine-tuned in Per-MPST to predict a personalized review and score for each book based on the user's preference. The detailed date process can be found in Appendix A.2. In Table 5 we can see that **PerSE** model outshines other baselines even if it never fine-tunes on the new domain. On the other hand, we fine-tune another **PerSE** based on Mistral 7B (Jiang et al., 2023) to investigate whether the proposed method can generalize to other LLMs. Results in Table 6 show that our method can enhance the capability of different LLMs in both in-domain and out-of-domain settings.

Table 5: Zero-shot performance on Amazon book review. The experimental setting is the same as Table 2.

	Pearson	**Spearman**	**Kendall**
Reviewer Avg.	0.146	0.180	0.177
LLaMA-7b	0.066	0.127	0.124
LLaMA-13b	0.070	0.122	0.112
LLaMA-70b	0.116	0.150	0.146
GPT-4	0.152	0.165	0.162
PerSE-7b	0.170	0.238	0.219
PerSE-13b	**0.217**	**0.247**	**0.237**

GPT-4 tends to be nice and general. In Figure 6 we show an example from Per-MPST. From the annotated reviews, we can see that this reviewer is critical about the plots, and especially cares about the novelty. However, even given this reviewer's preference, GPT-4 predicts a positive review, which may be caused by alignment towards safety or harmlessness. LLaMA-2-70b is stricter but gives a score of 4. **PerSE** cares more about the steady terribleness and only gives 3, which is more consistent with this reviewer's true score. Moreover, from the review

[4] https://nijianmo.github.io/amazon/index.html



Review Preference

[The Start of Plot 0]
A newscast is shown about a year after a prehistoric piranha attack on Lake Victoria. The lake is now uninhabitable, and the town has been abandoned due to the drying up of the main revenue source, tourism...
[The End of Plot 0]
[Review] Quote The Hoff: Welcome to rock-bottom.. features everything that gives horror , as well as sequels in general , a bad reputation [Score] 3
[The Start of Plot 1]
Evelyn Summers is a shy and unattractive but intelligent girl who has a crush on Bill Natolly, her handsome English teacher...
[The End of Plot 1]
[Review] honey, I think we should see other people... Don't you just hate it when this happens?... [Score] 5

Plot Query

Professor Jack Hanson and his colleagues discover evidence of global warming via a presentation to the UN that is poorly received. Professor Christopher Rapson believes Jack's theories and works with him to build a forecast model....
President Elizabeth Blake dies when her motorcade is caught in the superstorm, and Vice President Elizabeth James Becker becomes the new president. The next morning, Sam's group realizes the storm is over, and they begin their journey

Reference

[Review] I mean, you can't judge a film negatively because its far fetched. The whole thing may be scientifically unjustified, but it has to remain a little appealing to the audience, no? The dreadfulness of this film lies in the same old, disastrous cliche that keep on getting repeated over and over again. [Score] 2

GPT4 Response

[Review] This film offers a chilling, literally and metaphorically, journey into the devastating effects of global warming, a pressing issue that's often brushed aside... Nonetheless, the film effectively uses the fear of the unknown future of our planet to craft an engaging narrative. [Score] 7

LLaMA-2-70b Response

[Review] I'm just not sure if it all adds up to a compelling story. I mean, the characters are pretty one-dimensional, and the plot feels like it's all over the place. [Score] 4

PerSE Response

[Review] I really am amazed at how steadily and consistently terrible the movies are. Oh well, at least the disaster itself was kind of cool to watch, although even that had its flaws and clearly visible budget restrictions. Too bad they didn't spend a little more money on it, and actually make it a decent movie. [Score] 3

Figure 6: An example for evaluating the individual story from the given reviewer's preference. The reference is the ground-truth reviews given by this reviewer. More cases are shown in the Appendix C.

Table 6: Mistral-7b-based **PERSE** outperforms the original pre-trained model and achieves comparable performance with **PERSE**-7b. The setting is the same as Table 2.

	Pearson	**Spearman**	**Kendall**
In-domain: Per-MPST			
Mistral 7B	0.166	0.128	0.106
PERSE-Mistral	0.302	0.320	0.250
Out-of-domain: Amazon Book Review			
Mistral 7B	0.088	0.102	0.098
PERSE-Mistral	0.170	0.218	0.204

preference, we find that, unlike most people, this reviewer does not pay much attention to complicated themes. However, GPT-4's "one-size-fits-all" evaluation offers a high score for this theme. **PERSE** cares more about the visual preference of this reviewer, giving a more reviewer-specific rating. This indicates that **PERSE** can better evaluate stories based on personalized preferences rather than a general and nice evaluation principle without any personalized preference.

6 Conclusion and Discussion

In this paper, we focus on the evaluation of the alignment between open-ended generation and personal preference. We introduce **PERSE**, an LLM-based evaluation model that can provide an interpretable evaluation from the perspective of an unseen reviewer. It infers an unseen reviewer's preference based on a few annotated reviews and aligns its evaluation toward this preference. Besides the score, it also provides a detailed explanation (such as a review or multi-aspect comparison) for its evaluation, making it more interpretable. By instruction-tuned on personalization data, the LLaMA-2-based **PERSE** outperforms GPT-4 in both individual and pairwise settings. The comprehensive analysis of the personalized alignment highlights the importance of personalized finetuning to avoid the over-alignment with common human values by RLHF. The interpretability of **PERSE** makes it a better suit for personalized generation systems and recommendation systems.

Limitation

While this research makes notable strides in addressing the challenge of personalized evaluation, it is not without its limitations. For example, we assume that the preference is consistent within the prior reviews, which may not reflect the preference change in real-world scenarios. It would be interesting to model the preference shift over time and evaluate the context based on potential future preferences. Additionally, the current context length of large language models limits the number of reviews, which might affect the comprehensive understanding of a reviewer's preference. However, with the development of large language models with longer context windows, we believe that more reviews can be utilized for better modeling of the reviewer's preference. Furthermore, we show that with instruction tuning on personalization data, the small LLaMA-2 can outperform the larger GPT-4. More exploration can be done in large-scale LLMs to see the scalability of our method.

Ethics Statement

As we conduct extensive research to enhance and personalize the capabilities of Large Language Models (LLMs) such as the **PerSE** presented in this paper, we are ever-conscious of the ethical implications of our work.

One ethical concern is to ensure fairness and avoid potential bias in the personalization of LLMs. While **PerSE** aims to evaluate content based on individual preferences, we carefully construct the instruction data to alleviate the potential undesirable behaviors during the finetuning. We also enhance the transparency of the personalized evaluation by introducing interpretable metrics, as suggested in Kirk et al. (2023).

The other ethical consideration relates to privacy and consent. The two datasets Per-MPST and Per-DOC are reproduced from the existing publicly released datasets MPST (Kar et al., 2018, 2020) and DOC (Yang et al., 2023)under their licenses. They are sourced ethically and the privacy of individuals is always respected. All data used is aggregated and anonymized to safeguard personal information.

In conclusion, while the **PerSE** holds tremendous potential for fostering personalized human-AI interaction, careful consideration must be given to the ethical implications of its development and usage. We remain committed to conducting our research responsibly, adhering to ethical guidelines, to ensure that our contributions to AI advancements promote transparency, fairness, and respect for privacy.

References

Byung-Chull Bae, Suji Jang, Youngjune Kim, and Seyoung Park. 2021. A preliminary survey on story interestingness: Focusing on cognitive and emotional interest. In *Interactive Storytelling: 14th International Conference on Interactive Digital Storytelling, ICIDS 2021, Tallinn, Estonia, December 7–10, 2021, Proceedings 14*, pages 447–453. Springer.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Hong Chen, Duc 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*, pages 1739–1753.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307.16376*.

Zheng Chen. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

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 *29th International Conference on Computational Linguistics (COLING 2022)*.

W Bruce Croft, Stephen Cronen-Townsend, and Victor Lavrenko. 2001. Relevance feedback and personalization: A language modeling perspective. In *DELOS*. Citeseer.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7828–7838, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv: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 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, 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 (Volume 1: Long Papers)*, pages 6394–6407.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2020. Narrative text generation with a latent discrete plan. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3637–3650, Online. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*.

Sudipta Kar, Gustavo Aguilar, Mirella Lapata, and Thamar Solorio. 2020. Multi-view story characterization from movie plot synopses and reviews. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5629–5646, Online. Association for Computational Linguistics.

Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

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

Pan Li, Yuyan Wang, Ed H Chi, and Minmin Chen. 2023. Prompt tuning large language models on personalized aspect extraction for recommendations. *arXiv preprint arXiv:2306.01475*.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hui Shi, Yupeng Gu, Yitong Zhou, Bo Zhao, Sicun Gao, and Jishen Zhao. 2023. Every preference changes differently: Neural multi-interest preference model with temporal dynamics for recommendation. *ICML*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Han Lau. 2023. Deltascore: Evaluating story generation with differentiating perturbations. *arXiv preprint arXiv:2303.08991*.

Jiajing Xu, Andrew Zhai, and Charles Rosenberg. 2022. Rethinking personalized ranking at pinterest: An end-to-end approach. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 502–505.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. 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*, pages 4393–4479.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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 (EMNLP)*.

A More details about Personalization Dataset

There are two main challenges in constructing personalized evaluation datasets. First, it is difficult to collect preference labels for a long context. It requires the reviewer to identify their preferences first and read the whole context to provide a review. This process is time-consuming and costly. The second challenge is these online movies and reviews have been implicitly exposed to the training phase of many LLMs. This contamination problem may lead to evaluation bias for LLM-based evaluation models.

A.1 Contamination in evaluating existing stories

Many online stores have been exposed to LLMs during pretraining, which may lead to bias during evaluation. We first investigate how such contamination affects LLMs when evaluating general stories. We use average movie ratings from the non-commercial IMDB dataset,[5] containing plots evaluated by thousands of reviewers with scores ranging from 1 to 10. We let GPT-4 evaluate the movie plot and ask it to identify the movie title to check the memorization.

We consider a movie to be "known" by GPT-4 if the title is correct, and split the results into three groups based on memorization status: GPT-4 knows both plots, knows one, or knows neither. We calculated prediction accuracy (Accu.), consistency (Cons.), and bias first within each group. Consistency measures how many judgments are consistent after changing the order in which the two plots are presented. Bias first is defined as an inappropriate preference for the first one. It is calculated by subtracting the percentage where GPT-4 favors the first plot by the true percentage of the first.

We investigate the memorization problem in two settings: the pointwise evaluation is to predict a score (1 to 10) for a single story, and the pairwise evaluation is to compare two plots.

Pairwise evaluation We create 200 movie pairs, where each pair consists of two movie plots whose ratings differ by 1 point. We ask GPT-4 to identify the titles and then conduct a pairwise comparison [6]. Results on the original IMDB movie plots are reported in the 'Raw' rows of Table 7. We can see that GPT-4 knows at least one of the movies in the pair. Moreover, if GPT-4 knows exactly one of the two plots, it is more consistent in its judgment and has a lower position bias. We find it is because GPT-4 tends to choose the known plot. To alleviate the effect of memorization, we ask GPT-4 to identify the characters and local names in the plot and randomly replace them with similar names, ('Anonymized' in Table 7); doing so reduces the percentage of both known pairs by 18%. However, 96% of pairs still have at least one known plot. Therefore, we further summarize the anonymized plot ('Summarized'), reducing both known to 42.5% and increasing neither known to 23.5%. In all three groups, the summarized plots have the highest consistency and lowest position bias. Moreover, compared to the other two groups, neither known group exhibits much lower accuracy despite keeping the main plot points, indicating that memorization can result in misleadingly high performance in evaluation.

We further calculate the 'Bias Known' on the 'One known' group by subtracting the percentage that GPT-4 favors the known plot by the true percentage where this plot is better. In Table 8, we can see that for all raw, anonymized, and summarized plots, GPT-4 has an obvious tendency for the known plot when it can identify one of the plot pairs. This tendency is more obvious in the summarized plots. We suppose it is because, with the data processing, the uncertainty of the prediction increases. It makes the model more conservative, believing in what it has known. However, GPT-4 also shows high consistency and low position bias on the 'neither known' group (see Table 7), indicating that when facing two novel stories, it can get rid of the effect of memorization and evaluate based on the plots.

Pointwise evaluation We also investigate the influence of memorization on pointwise evaluation. Similarly, we ask the GPT-4 to identify the movie title and give a score (1 to 10) for this plot. We divide the results into 'Known' and 'Unknown' according to the success of the title identification. We calculate the correlation between the prediction scores and the average scores in IMDB. The results are shown in Table 9. The percentage of known significantly decreases after anonymization and summarization, indicating the effectiveness of alleviating memorization issues. Although the correlation on known plots is very high, it drops after

[5] https://developer.imdb.com/non-commercial-datasets/

[6] We used the gpt-4-0613 version from https://openai.com/gpt-4 with default settings.

Table 7: GPT-4 in comparing two movies. The plot with the correct predicted title is viewed as a known plot by GPT-4. Cons. is the percentage of consistent results when swapping the order. Bias First is the percentage where GPT-4 favors the first answer more than the ground truth. Percent is the percentage of each story type (raw/anonymized/summarized) recognized as 'both known', 'one known', or 'neither known'. Overall, memorization leads to greater position bias and lower consistency.

		Accu. ↑	Cons. ↑	Bias First ↓	Percent
Both Known	Raw	0.714	63.0%	16.5%	91.0%
	Anonymized	0.712	60.7%	17.8%	73.0%
	Summarized	0.753	73.4%	12.9%	42.5%
One Known	Raw	0.778	78.9%	-11.1%	9.0%
	Anonymized	0.804	71.7%	-6.5%	23.0%
	Summarized	0.632	82.4%	1.5%	34.0%
Neither Known	Raw	/	/	/	0.0%
	Anonymized	0.500	62.5%	25.0%	4.0%
	Summarized	0.660	85.1%	4.3%	23.5%

GPT-4 fails to identify the plots. It shows that the memorization issue makes the evaluation of GPT-4 unreliable.

Table 8: Prediction on 'One known' Group in pairwise comparison of GPT-4. The 'Raw', 'Anonymized', and 'Summarized' have the same meaning with Table 7. 'Bias known' is defined as the case that GPT-4 more favors the known plot than the ground-truth.

	Bias Known
Raw	0.222
Anonymized	0.283
Summarized	0.397

Personalized evaluation We also explored the influence of memorization in personalized evaluation for different LLMs. We provided one review from the same reviewer as the few-shot example and asked LLMs to predict a personalized score for the new plot. We experimented on randomly chosen 400 reviewers of Per-MPST with $k = 1$ and calculated the Kendall correlation between the human ratings and the predicted score in Figure 8 for LLaMA-2 and GPT-4. Similarly, LLMs achieved a high correlation with human ratings in original plots, but the performance degraded after anonymization and summarization. Although the main plots remain the same, with only slight differences in recognizable details, it greatly affected the results. Both experiments highlight that the memorization results in great bias in LLM-based evaluation models, making them unreliable for both general evaluation and personalized evaluation.

Overall, for LLM-based evaluation, contamination leads to an unfairly high rating on exposed plots, compared to unexposed ones.

A.2 Data processing

To address these problems, we create a less biased personalized evaluation dataset by anonymization of famous characters and summarization from existing plots. We demonstrate our pipeline in Figure 7.

For each reviewer, we first randomly pick several examples from this reviewer's prior reviews [7] For each plot, if it is already published online, we rewrite it to avoid contamination. Specifically, we use oasst-30b (Köpf et al., 2023) to anonymize and summarize the plots. It is a 30B LLaMA-based model finetuned on OpenAssistant Conversations for alignment.

Anonymization and Summarization The anonymization makes the character and location names less identifiable and the summarization avoids the text-level memorization while keeping the main idea of the plot. The anonymization is two-step: it first creates the name mapping and then replaces the name. It ensures that the model will not hallucinate new content during the name replacement.

In Figure 8, we investigate how the anonymization and summarization affect the evaluation performance. LLMs achieved a high correlation with human ratings in original plots, but the performance degraded after anonymization and summarization. Although the main plots remain the same, with only slight differences in recognizable details, it greatly affected the results. It indicates that these techniques can effectively alleviate the memorization problem.

Preference labels. Note that we do not have access to personal profiles that directly describe

[7] We assume that the reviewer's preferences are consistent within the review time frame.

Table 9: Performance of GPT-4 in predicting average movie scores. Percent is the percentage of each type of stories (raw/anonymized/summarized) being recognized as ‘known’, ‘Unknown’. Memorization heavily affects performance, but its impact decreases with anonymization and summarization.

		Pearson	**Spearman**	**Kendall**	**Percent**
Known	Raw	0.680	0.718	0.590	84.5%
	Anonymized	0.682	0.680	0.548	57.5%
	Summarized	0.621	0.648	0.552	27.0%
Unknown	Raw	0.460	0.470	0.364	15.5%
	Anonymized	0.216	0.289	0.222	42.5%
	Summarized	0.232	0.271	0.217	72.5%

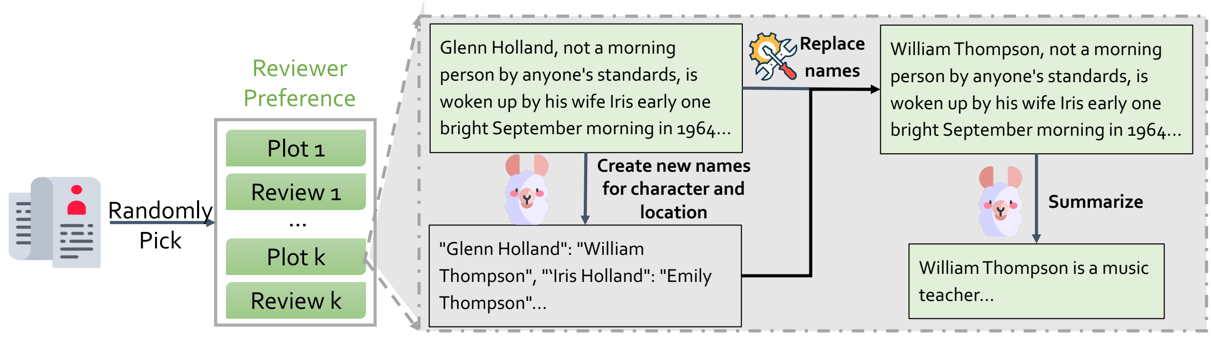

Reviewer Preference
Plot 1
Review 1
...
Plot k
Review k
Randomly Pick
Glenn Holland, not a morning person by anyone's standards, is woken up by his wife Iris early one bright September morning in 1964...
Replace names
William Thompson, not a morning person by anyone's standards, is woken up by his wife Iris early one bright September morning in 1964...
Create new names for character and location
"Glenn Holland": "William Thompson", "'Iris Holland": "Emily Thompson"...
Summarize
William Thompson is a music teacher...

Figure 7: The flowchart to construct our dataset. We use oasst-30b (Köpf et al., 2023), an instruction-tuned LLaMA-based model for anonymization and summarization. The prompts are listed in Figure 11.

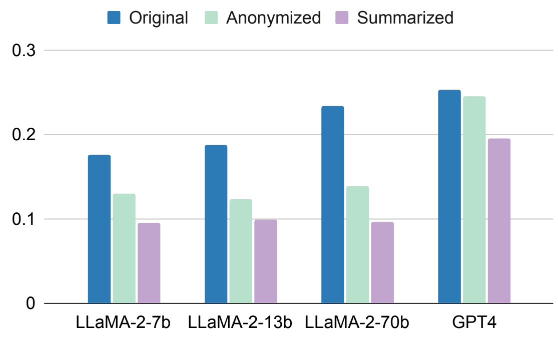


Figure 8: Kendall correlation between the LLM’s personalized prediction on movie scores with human ratings. Personalized predictions of all LLMs are also affected by memorization.

the story genres reviewers would like. Instead, we use existing reviews from the reviewer as the preference labels that typically reflect the evaluation principles and practices. For example, given the reviews in Figure 1, we can infer that Reviewer 1 favors happy endings while Reviewer 2 cares more about the plot complexity.

Finally, we repurpose two personalized story datasets: **Per-MPST** and **Per-DOC**.

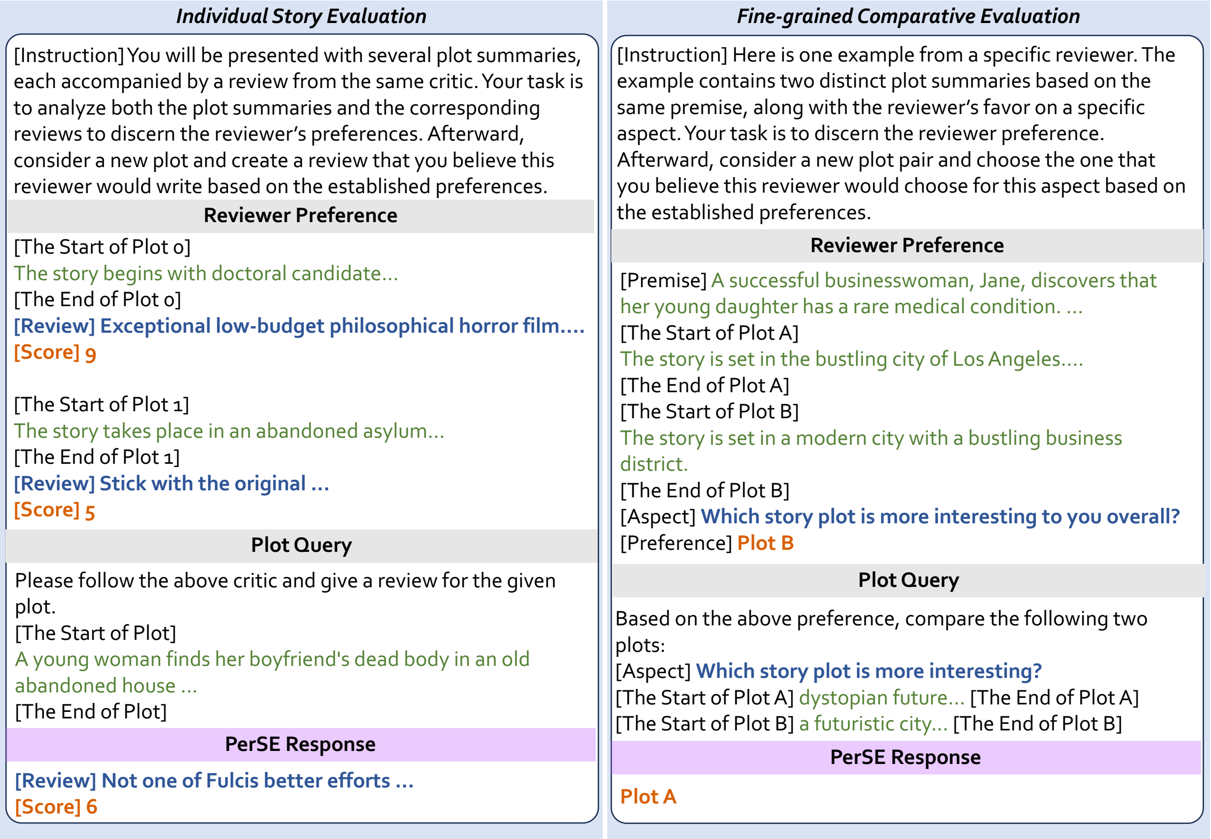
For Per-DOC, we define five aspects based on the questions in Yang et al. (2023):

1. `Interestingness`: Which story plot is more interesting to you?
2. `Adaptability`: In your opinion, which one of the plots above could generate a more interesting book or movie (when a full story is written based on it)?
3. `Surprise`: Which story plot created more suspense and surprise?
4. `Character Development`: Which story’s characters or events do you identify with or care for more?
5. `Ending`: Which story has a better ending?

These aspects evaluate the three key elements in the story: Interestingness and Surprise for the plot, Character development for the character, and Ending and Adaptability for the setting. For each question, there are four options: plot A, and plot B, both are good, and neither is good. We remove the examples with the answer of ‘Both’ and ‘Neither’ because they do not show preference.

We illustrate the length distribution of the movie plot in Per-MPST and the story in Per-DOC in Figure 10b and 10c. For Per-MPST, we also provide the length distribution of the raw plots in Figure 10a.

Variation of Score Preference across Reviewers We display the variation of score preferences in Per-MPST as per the table. We computed the count of reviewers for each query along with the average (mean) and standard deviation (std) of the re-



Individual Story Evaluation

[Instruction] You will be presented with several plot summaries, each accompanied by a review from the same critic. Your task is to analyze both the plot summaries and the corresponding reviews to discern the reviewer's preferences. Afterward, consider a new plot and create a review that you believe this reviewer would write based on the established preferences.

Reviewer Preference

[The Start of Plot 0]
The story begins with doctoral candidate...
[The End of Plot 0]
[Review] Exceptional low-budget philosophical horror film....
[Score] 9

[The Start of Plot 1]
The story takes place in an abandoned asylum...
[The End of Plot 1]
[Review] Stick with the original ...
[Score] 5

Plot Query

Please follow the above critic and give a review for the given plot.
[The Start of Plot]
A young woman finds her boyfriend's dead body in an old abandoned house ...
[The End of Plot]

PerSE Response

[Review] Not one of Fulcis better efforts ...
[Score] 6

Fine-grained Comparative Evaluation

[Instruction] Here is one example from a specific reviewer. The example contains two distinct plot summaries based on the same premise, along with the reviewer's favor on a specific aspect. Your task is to discern the reviewer preference. Afterward, consider a new plot pair and choose the one that you believe this reviewer would choose for this aspect based on the established preferences.

Reviewer Preference

[Premise] A successful businesswoman, Jane, discovers that her young daughter has a rare medical condition. ...
[The Start of Plot A]
The story is set in the bustling city of Los Angeles....
[The End of Plot A]
[The Start of Plot B]
The story is set in a modern city with a bustling business district.
[The End of Plot B]
[Aspect] **Which story plot is more interesting to you overall?**
[Preference] **Plot B**

Plot Query

Based on the above preference, compare the following two plots:
[Aspect] **Which story plot is more interesting?**
[The Start of Plot A] dystopian future... [The End of Plot A]
[The Start of Plot B] a futuristic city... [The End of Plot B]

PerSE Response

Plot A

Figure 9: The demonstrate of **PERSE**. The input is in green, the detailed review and fine-grained aspects are in blue, and the review scores are in orange.

viewers' scores. The table demonstrates that while the average scores for queries are almost identical, there is a sizable std difference. This underlines the necessity for an evaluation method that takes into consideration varying preferences.

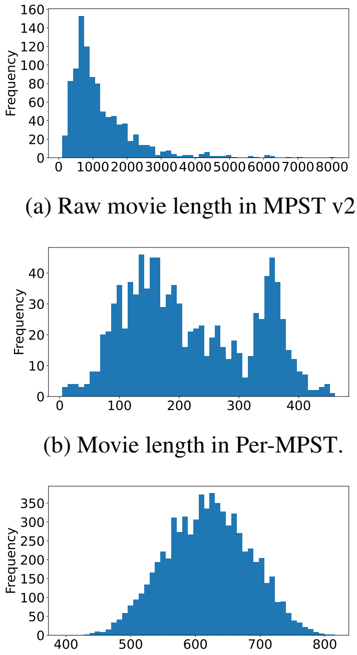
Table 10: Score variance in Per-MPST

	# Review	Score Mean	Score STD
Train	28.37	6.69	1.97
Test	4.64	6.84	1.39

Zero-shot Amazon dataset We preprocess the dataset to create a personalized version, which consists of 120 evaluation examples. Each example features one annotated review serving as the user profile (k=1): Opted for the 5-core subset of the book domain, where every user and item has a minimum of 5 reviews. Expanded the included brief book description using LLM-based retrieval. Anonymized the character and location names present in the raw book description. We then condensed this data to lessen potential contamination issues, an approach aligned with Per-MPST standards. We zero-shot test our PerSE model, fine-tuned on Per-MPST, directly onto this dataset without any further fine-tuning. To correspond with the scoring range within the Amazon dataset, we calibrate our PerSE score (1 to 10) to span from 1 to 5 after getting the prediction.

A.3 Prompts

We demonstrate the framework of **PERSE** in Figure 9 and list the detailed prompts used in Appendix A.1 and **PERSE** in Table 11. We anonymize the raw plot by asking LLMs to identify characters and local names and create new names for them. Based on the JSON mapping it generates, we replace those names with new names. We do not directly ask LLMs to replace names because they sometimes hallucinate new plots during the replacement. For the characters with the same family names, LLMs can create new character names that still have the same last names (but not the same as the original last names). For example, 'Glenn Holland' and 'Iris Holland' are mapped to 'William Thompson' and 'Emily Thompson'.



(a) Raw movie length in MPST v2.

(b) Movie length in Per-MPST.

(c) Story length in Per-DOC.

Figure 10: Length Distribution of Per-MPST and Per-DOC. The x-axis is the length and the y-axis is the frequency.

B Training Details

Each model in our experiments was trained on 8 x 80G A100 GPU with a learning rate of 1e-5. We set the batch size to 4 for **PerSE**-7b and 2 for **PerSE**-13b. **PerSE**$_{\text{ind}}$-7b and **PerSE**$_{\text{ind}}$-13b converged after 2k/6k steps on Per-MPST respectively. We trained two unified models on Per-DOC for all aspects by finetuning 7b and 13b LLaMA-2-chat. **PerSE**$_{\text{comp}}$-7b converged after 1k steps and **PerSE**$_{\text{comp}}$-13b converged after 2k steps. It took about 10 hours for these two models. For the ablation study, we also trained one model for each aspect on Per-DOC and each model converged after 500 steps for 7b and 2k steps for 13b. The total training time was around 5 x 5 hours. We plot the curve of the training loss in Figure 12.

C More Analysis

PerSE infers the preference instead of copying scores from context. In Figure 13, we show another example on Per-MPST. From the reviews, we can find the reviewer loves horror elements. However, the new plot and its level of terror are not satisfactory, which makes the reviewer give it a low score. Both GPT-4 and LLaMA-2-70b emphasize the horror theme and predict a high score for this plot. We suppose that they are affected by the high review scores in the reviewer's preference, ignoring the analysis of the new plot. In contrast, **PerSE** focuses on the boring profiling of the plot, which is more similar to what the reviewer cares about. It gives a score of 5, which is different from the existing review scores but close to the real score this review has for this plot.

PerSE can provide diverse reviews for the same plot based on different preferences. In Figure 14, we demonstrate the reviews of the same plot from two reviewers A and B with different preferences. We can see that both the reviewer A and B have read the book. Reviewer A is a critical reviewer and has a high standard for good movies, leading to low scores in the annotated reviews. He then gives a score of 2 because of his disappointment with the movie adaptation. In contrast, reviewer B is relatively tolerant and likes to score high. Although the movie is much worse than the book, the reviewer still gives a score of 6. However, GPT-4 and LLaMA-2-70b give similar high scores in both cases, ignoring the reviewer's preference. Instead, **PerSE** is able to give personalized scores for different reviewers, predicting 1 for reviewer A and 8 for reviewer B. Although the predicted score of reviewer B is not as close as GPT-4, it illustrates the positive attitude it captures.

PerSE achieves better performance on fine-grained comparative evaluation. We illustrate one example from Per-DOC in Figure 15. **PerSE** successfully predicts the preference on 4 out of 5 aspects, while GPT-4 correctly predicts 3 aspects and LLaMA-2-70b only has 2 success. GPT-4 predicts Plot A for all aspects, ignoring the difference between aspects and outputs an overall evaluation. Instead, **PerSE** cares more about the distinctive attribute of each aspect and gives judgment according to the aspect.

The review text in profile helps PerSE infer preferences. We also investigate the role of review text in the reviewer profile in **PerSE**. We removed the review text $\boldsymbol{a}_{u_i}$ in the reviewer profile $\boldsymbol{c}_u$ and retrained the model. The results are shown in Table 11. We can observe that for all models the performance degraded after removing the detailed review content. This highlights the importance of incorporating the explanation of the review score when evaluating the story, especially for unseen reviewers. This performance decrease is more significant for pre-trained LLaMA models because they are more sensitive to the prompts.

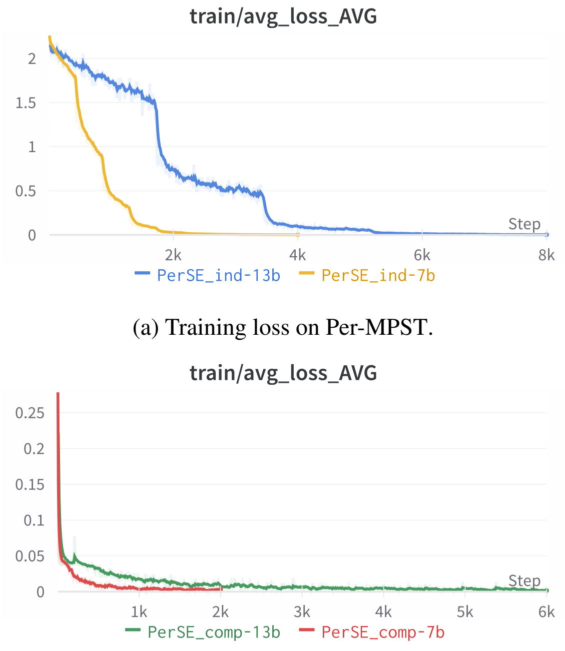
Anonymization	Here is one plot: {plot} Please create a JSON mapping of current character and location names to new, distinctive names. In this mapping, the current names will act as keys and the new names as values. For instance, if you were to change the name 'Diego' to 'Sherry Evans', the corresponding JSON entry would be: {{'Diego': 'Sherry Evans'}}. The task requires you to replace all character and location names in the text with alternative names, and then provide the mapping relationship as a JSON object.
Summarization	Provided below is a narrative: {plot} Kindly analyze this story and provide a clear and succinct summary of the key events.
Individual Story Evaluation	Here we have one plot. Please give a score for 1 to 10 for the following plot, where 1 is the lowest and 10 is the highest. If you already know the plot, give the name. But remember do not depend on any public review score you already remember. [Plot] {plot} Please only reply a JSON-format with the following keys: "Score", "Title". If you cannot identify the title, respond with "N/A" for that field.

Pairwise Story Evaluation	Here we have two plots: plot1 and plot2. Please based on the description to choose which one is better and give your reasons. If you know the movie title of this plot, please tell me the titles as well.
	[Plot1] {plot1}
	[Plot2] {plot2}
	Please only reply a JSON-format with the following keys: "Choice", "Reason", "Plot1 Title", "Plot2 Title". If you cannot identify the title, respond with "N/A" for that field.

Figure 11: Prompts used in Section A.1. The blue text is the placeholder for plots.

Table 11: The ablation study on the review content on Per-MPST. We utilized three reviews ($k = 3$) to represent the reviewer's preferences. The results are the average of three replicate experiments with p-values less than 0.05. Removing review content leads to performance degradation.

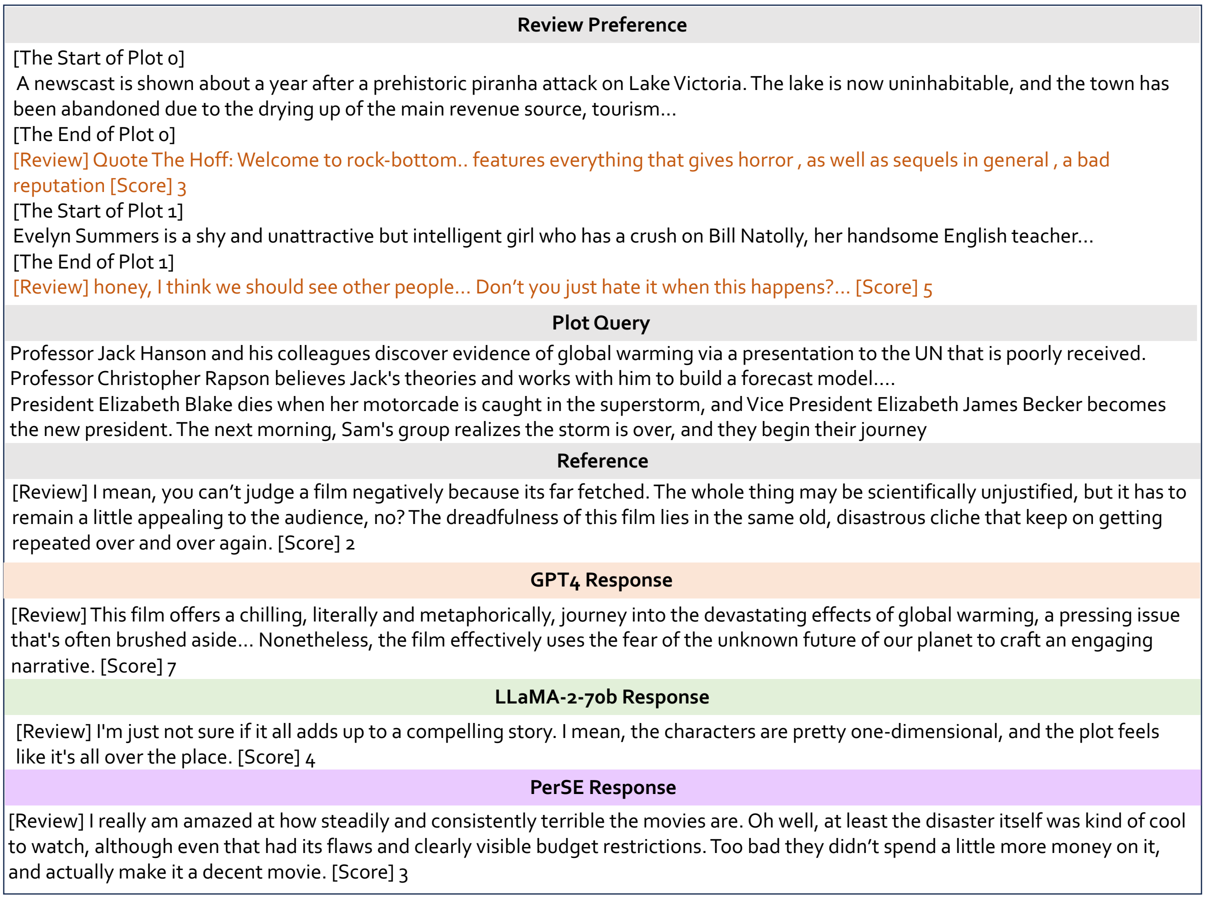
	Pearson	**Spearman**	**Kendall**
LLaMA-2-7b	0.123	0.130	0.104
w/o content	-0.007	-0.002	-0.005
LLaMA-2-13b	0.163	0.191	0.154
w/o content	-0.039	-0.035	-0.028
PERSE-7b	0.322	0.340	0.272
w/o content	-0.088	-0.064	-0.057
PERSE-13b	0.381	0.378	0.301
w/o content	-0.088	-0.069	-0.057



(a) Training loss on Per-MPST.

(b) Training loss on Per-DOC.

Figure 12: Average training loss



Review Preference

[The Start of Plot 0]
A newscast is shown about a year after a prehistoric piranha attack on Lake Victoria. The lake is now uninhabitable, and the town has been abandoned due to the drying up of the main revenue source, tourism...
[The End of Plot 0]
[Review] Quote The Hoff: Welcome to rock-bottom.. features everything that gives horror , as well as sequels in general , a bad reputation [Score] 3
[The Start of Plot 1]
Evelyn Summers is a shy and unattractive but intelligent girl who has a crush on Bill Natolly, her handsome English teacher...
[The End of Plot 1]
[Review] honey, I think we should see other people... Don't you just hate it when this happens?... [Score] 5

Plot Query

Professor Jack Hanson and his colleagues discover evidence of global warming via a presentation to the UN that is poorly received. Professor Christopher Rapson believes Jack's theories and works with him to build a forecast model....
President Elizabeth Blake dies when her motorcade is caught in the superstorm, and Vice President Elizabeth James Becker becomes the new president. The next morning, Sam's group realizes the storm is over, and they begin their journey

Reference

[Review] I mean, you can't judge a film negatively because its far fetched. The whole thing may be scientifically unjustified, but it has to remain a little appealing to the audience, no? The dreadfulness of this film lies in the same old, disastrous cliche that keep on getting repeated over and over again. [Score] 2

GPT4 Response

[Review] This film offers a chilling, literally and metaphorically, journey into the devastating effects of global warming, a pressing issue that's often brushed aside... Nonetheless, the film effectively uses the fear of the unknown future of our planet to craft an engaging narrative. [Score] 7

LLaMA-2-70b Response

[Review] I'm just not sure if it all adds up to a compelling story. I mean, the characters are pretty one-dimensional, and the plot feels like it's all over the place. [Score] 4

PerSE Response

[Review] I really am amazed at how steadily and consistently terrible the movies are. Oh well, at least the disaster itself was kind of cool to watch, although even that had its flaws and clearly visible budget restrictions. Too bad they didn't spend a little more money on it, and actually make it a decent movie. [Score] 3

Figure 13: The score given by the reviewers on the new plot is very different from the comments with annotations. While LLaMA-2-70b and GPT-4 give a more similar score, **PERSE** is able to infer the preference and provide a score that is closer to the true score but far away from the annotated scores.

Review A Preference

[The Start of Plot 0]
Zara encounter a man beating a transsexual prostitute. She tries to intervene but is raped and beaten unconscious..... Zara wakes up and is sad that the man who raped her is dead.
[The End of Plot 0]
[Review] Irredeemable! The viewer is taken on a ride through the tunnel of sado-masochistic grotesquerie, beginning at the end and traveling with jolts and stops back to the start. [Score] 1
[The Start of Plot 1]
Henri Ford is a professor at UCLA who is still grieving the loss of his partner Ryan Miller three years earlier....
[The End of Plot 1]
[Review] Miscast! Actually the novel is rather good. And the film does not depart all that much from it, just enough to create questions.! [Score] 5

Review B Preference

[The Start of Plot 0]
Two police officers respond to a car crash, where two people are killed, but one miraculously survives. The story then cuts to the title credits.
[The End of Plot 0]
[Review] Folks, let me tell you, this is one GREAT weird little thriller, not much blood, but a whole lot of creepiness, a great story, and very british. [Score] 10
[The Start of Plot 1]
Angela is a call girl who is terrorized by a man named Frank, who was just released from prison.
[The End of Plot 1]
[Review] The trouble with anthology films like this, is that you're usually gonna get stories that are better then the others. [Score] 7

Plot Query

Richard Hayes is a financial whiz who's about to earn a million dollars through a bonds scheme. He has a wife, Julianna, and daughter, Grace, and is engaged in an affair with Sophia Taylor, a Southern belle gold digger...
After much turmoil, Richard and Sophia are both acquitted of all charges and are allowed to leave the courtroom, where Richard thanks the jury for their decision.

Reference A	Reference B
[Review] As someone who has both read the novel and seen the film, I have a different take on why the film was such a flop. First, any comparisons between novel and film are purely superficial. [Score] 2	[Review] I liked the book better. The film pales in comparison unfortunately. So, while I dont see this is the worst movie (its OK, NOT BAD), its a time-waster, if that. [Score] 6
GPT4 Response	**GPT4 Response**
[Review] A story that confronts the uncomfortable realities of racial disparities, privilege, and the media's role in shaping public opinion; this plot, albeit intense, presents a new way of examining societal issues. [Score] 7	[Review] While the plot navigates through themes of social classes and race relations, it misses out on delivering the horror and weirdness that I tend to look for in films. Despite its lack of otherworldly elements, the storyline is strong and it delivers some thrilling moments. [Score] 6
LLaMA-2-70b Response	**LLaMA-2-70b Response**
[Review] Overall, this is a well-crafted and engaging novel that will keep you hooked until the very end. [Score] 8	[Review] This movie is a great example of how our society can be both hypocritical and discriminatory. [Score] 9
PerSE Response	**PerSE Response**
[Review] In short, it is a perfect example of how elites feel about the great unwashed.I could perhaps forgive the film its sins if it were simply incompetent. But it is not incompetent. [Score] 1	[Review] Its not a happy movie by any means, but it is thought provoking. Wealth, and the lack thereof is the main focus, and how it can distort ones sense of right and wrong...[Score] 8

Figure 14: Reviews from two reviewers on the same plot. **PerSE** is able to give personalized scores based on preference.

Reviewer Preference	[Premise] A lonely man in his seventies befriends a group of street kids who teach him to find meaning in the simplest of joys, and he, in turn, teaches them about a different perspective on life. They rediscover life and its small joys together. [The Start of Plot A] Settings The story is set in an urban city with a mix of old and new architecture. Characters Sophie Wong Sophie Wong is 16 years old, a street kid who has been living on the streets since the age of 12, when she ran away from an abusive home.Mark Chen Mark Chen is 25 years old, a caring and compassionate social worker who befriends Edward and the street kids.Edward James Edward James is 75 years old, a retired math teacher, living alone in a small apartment since his wife died three years ago. Outline 1. Edward becomes lost in his grief after his wifes death and becomes detached from the world around him. 2. Sophie and the other street kids discover him sleeping on a park bench one night and, sensing his loneliness, initiate a friendship with him. 3. Mark, the social worker, recognizes Edwards situation and offers his help, which brings him closer to the street kids and helps him find a new purpose in life. [The End of Plot A] [The Start of Plot B] Settings The story is set in a small town in the United States. Characters Tito Robles Tito Robles is 15, a street kid who is the leader of the group he befriends John with, and together, they find meaning in life.Jane Davis Jane Davis is 40, Drews wife, and a friendly and welcoming presence in the town.Ben Smith Ben Smith is 45, a retired military man who lives in the same town and provides help and advice to John and the street kids when they need it.John Doe John Doe is 75, a retired man with a small house and a lonely life.Drew Davis Drew Davis is 50, the local bartender and a friend of John, who helps him connect with the street kids and their way of life. Outline 1. John becomes friends with Tito and the street kids, and together they rediscover the simple joys of life despite their different ages and backgrounds. 2. Drew, Jane, Ben, and other townspeople play important roles in helping the group of friends and teaching them about life and caring for one another. 3. The man decides to help the street kids and provides them with a house filled with toys and games. [The End of Plot B] **[Interestingness] Plot A [Adaptability] Plot B [Surprise] Plot A [Character Development] Plot A [Ending] Plot A**
Plot Query	[Premise] A struggling artist, living in a small town, stumbles upon an antique store that holds a mysterious painting with the power to change the course of her life, but at what cost? [The Start of Plot A] Settings The story is set in a small, rural town in the American South. Characters Maddie James Maddie James is 30 years old, Emmas best friend and roommate, with a quirky personality and a passion for art.Charles Carson Charles Carson is 45 years old, Emmas high school art teacher, who saw her potential and pushed her to pursue her artistic ambitions.Emma Watson Emma Watson is 24 years old, with wild, curly hair and big, expressive eyes. Outline 1. Emma discovers the mysterious painting at the antique store and starts to experience strange occurences around her town, leading her to suspect the true power of the art work. 2. Motivated by her desire to understand the paintings power, Emma begins to research and is guided by her art teacher and mentor towards her potential as an artist. 3. Emma starts to experience success as an artist and is approached by a powerful art dealer who reveals the true nature and power of the mysterious painting and offers her a tempting deal that threatens her family and friends. [The End of Plot A] [The Start of Plot B] Settings The story is set in a small town surrounded by vast, open fields and rolling hills. Characters Jackson Wrightson Jackson Wrightson is 29 years old, an art appraiser and Elaras ex-boyfriend, who is both supportive and a source of tension in her life.Elara Kassin Elara Kassin is 32 years old, with a kind heart and a struggling artist living in a small town.Lila Williams Lila Williams is 26 years old, Elaras best friend and a supportive companion who helps Elara on her journey to uncover the truth.Iris Beller Iris Beller is 61 years old, a kind and wise antique store owner, who serves as a confidante and mentor to Elara.Adrian Roth Adrian Roth is 33 years old, charming with disheveled hair and a mysterious demeanor, runs an antique store with a secret to hide. Outline 1. Elara discovers the mysterious painting at Adrians antique store, but quickly realizes the painting is more than just a simple work of art. 2. Elara starts to experience strange dreams and visions, causing her to explore the paintings true purpose and the consequences of her involvement in its magic. 3. Elara, with the help of Lila, Jackson, and Iris, uncovers Adrians true intentions and the dark ritual required to harness the paintings power. [The End of Plot B]
Reference	**[Interestingness]** Plot B **[Adaptability]** Plot A **[Surprise]** Plot A **[Character Development]** Plot B **[Ending]** Plot A
GPT-4	**[Interestingness]** Plot A **[Adaptability]** Plot A **[Surprise]** Plot A **[Character Development]** Plot A **[Ending]** Plot A
LLaMA-2-70b	**[Interestingness]** Plot A **[Adaptability]** Plot B **[Surprise]** Plot A **[Character Development]** Plot A **[Ending]**Plot A
PerSE	**[Interestingness]** Plot B **[Adaptability]** Plot A **[Surprise]** Plot B **[Character Development]** Plot B **[Ending]** Plot A

Figure 15: One case of comparative evaluation on Per-DOC. **PERSE** is more similar to this reviewer. However, it fails to capture the preference of `Surprise` in this case.