Learning Personalized Alignment in Evaluating Open-ended Text Generation

Danqing Wang[1,2] **Kevin Yang**[2] **Hanlin Zhu**[2,3] **Xiaomeng Yang**[2]
Andrew Cohen[2] **Lei Li**[1] **Yuandong Tian**[2]
[1]Carnegie Mellon University [2]Meta AI [3] UC Berkeley
danqingw@andrew.cmu.edu
{yangkevin,andrewcohen,yuandong}@meta.com
bit.yangxm@gmail.com hanlinzhu@berkeley.edu leili@cs.cmu.edu

Abstract

Recent research has increasingly focused on evaluating large language models' (LLMs) alignment with diverse human values and preferences, particularly for open-ended tasks like story generation. Traditional evaluation metrics rely heavily on lexical similarity with human-written references, often showing poor correlation with human judgments and failing to account for alignment with the diversity of human preferences. To address these challenges, we introduce **PERSE**, an interpretable evaluation framework designed to assess alignment with specific human preferences. It is tuned to infer specific preferences from an in-context personal profile and evaluate the alignment between the generated content and personal preferences. **PERSE** enhances interpretability by providing detailed comments and fine-grained scoring, facilitating more personalized content generation. Our 13B LLaMA-2-based **PERSE** shows a 15.8% increase in Kendall correlation and a 13.7% rise in accuracy with zero-shot reviewers compared to GPT-4. It also outperforms GPT-4 by 46.01% in Kendall correlation on new domains, indicating its transferability [1].

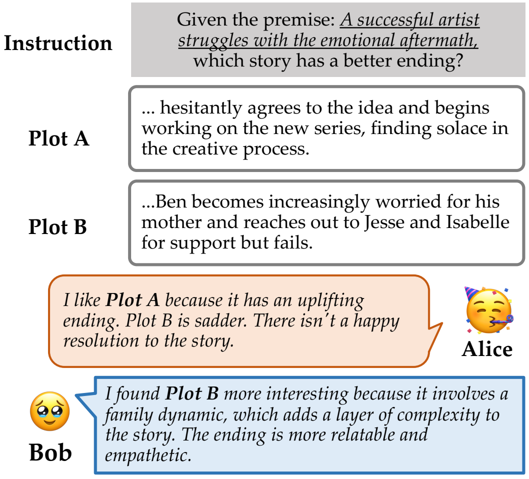

Instruction
Given the premise: A successful artist struggles with the emotional aftermath, which story has a better ending?
Plot A
... hesitantly agrees to the idea and begins working on the new series, finding solace in the creative process.
Plot B
...Ben becomes increasingly worried for his mother 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.
Alice
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.
Bob

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

1 Introduction

Large language models (LLMs) have recently demonstrated impressive generative capabilities across various tasks, with rapid improvements 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 remains challenging due to the diversity of responses. Traditional automatic metrics struggle with the one-to-many problem in open-ended generation (Liu et al., 2016) and often show poor correlation with human judgment (Krishna et al., 2021; Guan et al., 2021). Recent studies have trained evaluation metrics based on human ratings to better approximate human judgments (Sellam et al., 2020; Rei et al., 2020). However, these metrics primarily focus on objective qualities and tend to overlook subjective assessments, such as surprise (Chhun et al., 2022) or interestingness (Bae et al., 2021).

Subjective evaluation metrics are significantly influenced by diverse human preferences. For example, Figure 1 illustrates 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 due to its plot complexity and empathetic ending. This highlights the need for an automatic personalized evaluation metric that can assess model generations based on varying preferences. However, it is costly for each reviewer to provide a large number of personalized examples to demonstrate their preferences. This makes it impractical to train a separate evaluation model for each reviewer and to generalize the existing metric to unseen reviewers.

Moreover, the subjective nature of these evalu-

[1] Both datasets and code are released at https://github.com/facebookresearch/perse.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13274–13292
November 12-16, 2024 ©2024 Association for Computational Linguistics

ations makes the scores harder to interpret. AuPEL (Wang et al., 2023) incorporates personalization as one of the evaluation aspects to compare two inputs, but it does so without any explanation. This lack of transparency undermines the trustworthiness and reliability of evaluations and complicates the development of generative models (Leiter et al., 2022). Therefore, the key challenges in personalized evaluation are modeling an unseen reviewer's preference from a limited annotated personalized context and providing an interpretable explanation for the assessment.

In this paper, we introduce an LLM-based evaluation model, **PERSE**, designed to assess the alignment between open-ended generations and specific preferences. **PERSE** is tuned to infer preferences from a limited-length profile and uses this information to evaluate the generated content. **PERSE** provides an overall score along with an explanation for the scalar rating and offers fine-grained scores on several aspects to interpret the alignment for pairwise ratings. We curated two instruction-following datasets, Per-MPST and Per-DOC, to support personalized alignment in evaluation. **PERSE** is fine-tuned from LLaMA-2 (Touvron et al., 2023) to enhance its capability to infer preferences from reviewer profiles and apply these preferences in evaluations. Compared with GPT-4, **PERSE** achieves a 15.8% higher Kendall correlation in the scalar rating of movie plot generation and a 13.7% higher accuracy in the pairwise rating of story generation for zero-shot reviewers. It also outperforms GPT-4 by 46.01% in Kendall correlation when transferred to new domains. Our contributions can be summarized as follows:

- We develop an LLM-based evaluation model, **PERSE**, to assess alignment between open-ended generations and in-context preferences. By instruction-tuning on personalized data, it significantly outperforms GPT-4 in evaluating personal alignment.
- **PERSE** provides detailed explanations for its assessments. Its interpretability makes it particularly suitable for guiding personalized content generation.
- We curate two instruction-following datasets specifically for personalized alignment in the evaluation of open-ended generations.
- We find that LLMs, after reinforcement learning via human feedback, tend to be less personalized and more cautious with negative comments, which hinders their ability to align with strong personal preferences. However, when instruction-tuned with personalized data, even less powerful LLMs can perform better in aligning with preferences.

2 Related Work

Evaluation Metrics for Text Generation Automatic metrics can be broadly categorized 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). In contrast, reference-free metrics directly assess the quality of the model output without any reference. These metrics are usually 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 (Fu et al., 2023; Kocmi and Federmann, 2023; Xu et al., 2023), or as judges (Bai et al., 2023; Zheng et al., 2023). In this paper, we investigate LLMs' capabilities in learning personalized alignment on subjective aspects, which is crucial for evaluating open-ended generation.

Human Evaluation for Generation Human evaluation is employed to assess various 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), and interestingness (Bae et al., 2021). To comprehensively cover all aspects, Chhun et al. (2022) suggested six human criteria for storytelling: relevance, coherence, empathy, surprise, engagement, and complexity. However, they found that the inter-annotator agreement for human evaluation on these subjective aspects is low. Karpinska et al. (2021) also pointed out the risks of crowdsourced human judgments from Amazon Mechanical Turk due to underqualified workers and a lack of reproducibility details.

Personalization in Text Generation and Evaluation Personalization has been extensively studied in many recommendation systems (Das et al., 2007; Xu et al., 2022) and search applications (Croft et al., 2001; Shi et al., 2023). Recently, researchers have emphasized its importance in natural language processing (Flek, 2020; Dudy et al., 2021). Several

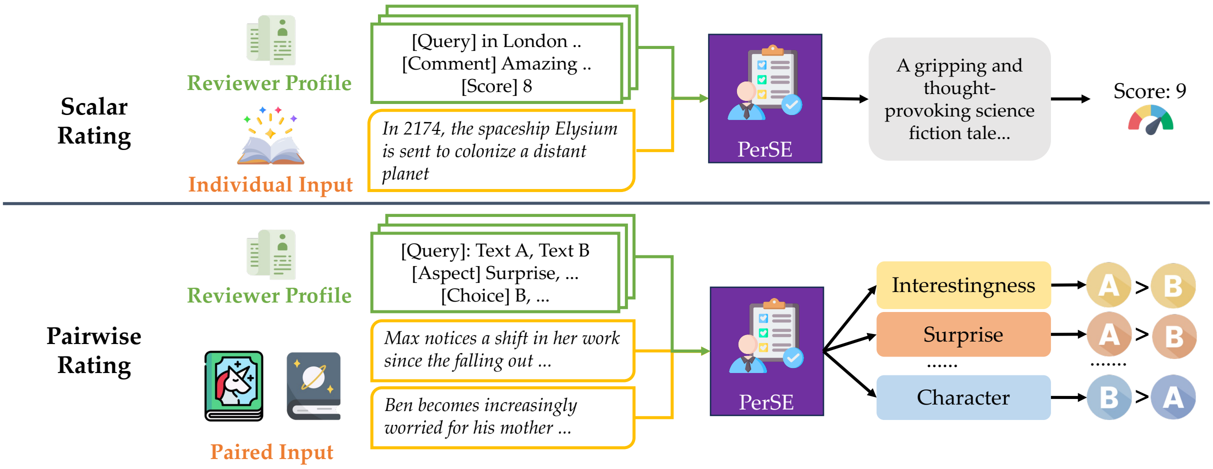


Figure 2: **PerSE** provides the scalar rating and pairwise rating for the personalized alignment in evaluation. **PerSE** infers the reviewer's preference from the profile with historical reviews and employs the preference to provide an interpretable evaluation.

recent studies have explored LLMs' capabilities in capturing personalization (Chen et al., 2023; Kang et al., 2023; Salemi et al., 2023) or in prompting for personalized recommendations (Lyu et al., 2023; Chen, 2023; Li et al., 2023). Wang et al. (2023) introduced a personalization score as one of the evaluation aspects, using LLMs as evaluators. In this paper, we propose an interpretable evaluation model to align with personal preferences, which not only provides an assessment but also a detailed explanation.

3 Learning Personalized Alignment in Evaluating Open-ended Generation

We propose an LLM-based evaluation model for assessing the alignment between generated content and personal preferences. **PerSE** provides a reference-free, interpretable evaluation from a specific reviewer's perspective.

Given an open-ended generation as the query $\boldsymbol{x}$ and the personal preference p_u of reviewer u, the goal is to provide a scalar or pairwise rating y_u along with an explanation $\boldsymbol{e}_u$ to indicate how well the query aligns with the reviewer's personal preference. The reviewer's preference is defined by their profile, which includes their historical comments $p_u = \{(\boldsymbol{x}^{(i)}, \boldsymbol{e}_u^{(i)}, y_u^{(i)})\}_{i=0}^{K}$, where K is the number of historical comments. The tuple $(\boldsymbol{x}^{(i)}, \boldsymbol{e}_u^{(i)}, y_u^{(i)})$ is sampled from the historical set $\mathcal{M}_p$. Typically, the number of a reviewer's historical comments is limited, making it challenging to train a separate rating model for each reviewer[2].

As demonstrated in Figure 2, **PerSE** provides a scalar rating for individual input text and a pairwise rating between two inputs. For the scalar rating, the reviewer's comment serves as the explanation $\boldsymbol{e}_u$, and their rating (ranging from 1 to 10) represents the alignment score y_u. For a more interpretable comparison between two inputs, we conduct a fine-grained assessment across a set of aspects $\mathcal{A}$. For each aspect, **PerSE** determines the winner between the two inputs and uses this as the rating. The aspects and their ratings collectively form the explanation $\boldsymbol{e}_u$. The win rate across all aspects is considered the alignment score y_u.

3.1 Evaluate with Personal Preference

We use LLMs to evaluate open-ended generations based on the reviewer's profile p_u. The LLM is instructed to analyze the implicit personal preferences from the historical comments in the profile p_u and predict the score y_u and explanation $\boldsymbol{e}_u$ for the new query $\boldsymbol{x}$ based on these preferences. In scalar ratings, the query $\boldsymbol{x}$ is a single generation, while in pairwise ratings, it consists of two generations. A high score indicates that the generation aligns well with the preference and suggests that the reviewer is likely to rank it highly. Conversely, a low score suggests that the reviewer may not favor it. For the same query $\boldsymbol{x}$, there could be distinct scores $y_{u_i} \neq y_{u_j}$ for different reviewers $u_i \neq u_j$. The prompt templates are listed in Figure 9 in the Appendix.

3.2 Curate Personalized Alignment Datasets

To learn the alignment between personal preferences and generated content, we curate person-

[2] For simplicity, we assume the reviewer's preferences are consistent within the review time frame.

alized alignment data for evaluating open-ended generation. We utilize two data sources: existing evaluation datasets that include reviewer identity and crowd-sourced annotations for open-ended generation systems (such as MTurk).

However, we find that if a dataset has been exposed to most LLMs during their pre-training, it can suffer from severe contamination issues during evaluation. The LLM-based evaluation model may memorize the ground-truth scores and perform well on the exposed test set, but its performance drops dramatically on non-memorized data, making it difficult to generalize to new cases. We discuss the influence of contamination issues on the evaluation of open-ended generation in Appendix A.

To address this, we recreate existing evaluation datasets to alleviate contamination issues through anonymization and summarization. We extract identifiable entities, such as characters and locations, and replace them randomly. We then summarize the raw content to omit details while retaining the main content. These strategies help significantly resolve the contamination issue. The details of the data processing and an analysis of its effect are provided in Appendix B.

3.3 Learning Personalized Alignment in Evaluation

The goal is to enhance LLMs' capabilities in inferring new preferences from an in-context reviewer's profile and applying these preferences in evaluation. We split our dataset into two non-overlapping sets based on the query: the historical reviews $\mathcal{M}_p$ is used for the reviewer profile, and $\mathcal{D}$ is for personalized alignment. We also divide the reviewers in the dataset into $\mathcal{U}_{\text{ift}}$ and $\mathcal{U}_{\text{test}}$, which are used for instruction-tuning and testing, respectively. Importantly, there is no overlap between the reviewers in $\mathcal{U}_{\text{ift}}$ and $\mathcal{U}_{\text{test}}$, ensuring that the model cannot directly apply any memorized preferences from fine-tuning during inference time.

The instruction-tuning dataset is defined as $D_{\text{ift}} = \{(\boldsymbol{x}, p_u, \boldsymbol{e}_u, y_u) | (\boldsymbol{x}, \boldsymbol{e}_u, y_u) \in \mathcal{D}, u \in \mathcal{U}_{\text{ift}}\}$, where $p_u \subseteq \{(\boldsymbol{x}, \boldsymbol{e}_r, y_r) \in \mathcal{M}_p \,| r = u\}$. This dataset includes all reviewers from $\mathcal{U}_{\text{ift}}$, with profiles built on the historical set $\mathcal{M}_p$. **PERSE** learns personalized alignment for evaluation based on instruction fine-tuning on D_{ift}.

4 Experiment

We introduce our two datasets in Section 4.1. In Section 4.2, we describe the implementation of **PERSE** and several baselines. Further details on the fine-tuning process are listed in Appendix D.

4.1 Datasets

We recreate our dataset, Per-MPST, from the existing movie review dataset MPST (Kar et al., 2018, 2020), as detailed in Section 3.2. We reorganize the released human annotations from Zhu et al. (2023) for personalized alignment, resulting in Per-DOC. As described in Section 3.3, we split the reviewers by a 9:1 ratio into $\mathcal{U}_{\text{ift}}$ and $\mathcal{U}_{\text{test}}$, building the instruction-tuning and test sets based on these groups. The instruction-tuning set is used for learning personalized preference alignment, while the test set evaluates model performance on unseen reviewers. The statistics are listed in Table 1.

Per-MPST MPST is a movie review dataset collected from IMDb[3]. It includes a synopsis and multiple comments for each movie. Each comment contains a review text and a score ranging from 1 (lowest) to 10 (highest). We group comments by reviewer ID and remove reviewers with fewer than 6 comments to ensure there are at least 5 historical comments. We create different versions by sampling various numbers of historical reviews ($K = 1$ to 5) as the reviewer profile. Additionally, we remove queries with more than 2500 words (about 4k tokens) to fit within the context window of LLMs.

Per-DOC This dataset contains 7,000 unique examples from 403 annotators, based on the released annotations from Zhu et al. (2023). Each example consists of two plots generated from the same premise, and annotators were asked to answer various questions and choose their preferred plot for each question. We define five subjective aspects: `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` pertains to the satisfaction or appreciation of the ending, and `Adaptability` measures the potential for further developing the story. We removed annotators with fewer than 2 annotations and use $K = 1$ for the reviewer profile.

[3] https://www.imdb.com/

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

4.2 Experimental Setting

We implement **PerSE** based on LLaMA-7b-chat and LLaMA-13b-chat, tuning them on the D_{ift} of Per-MPST and Per-DOC for scalar and pairwise ratings, respectively. In our main experiments, we use $k = 3$ for Per-MPST and $k = 1$ for Per-DOC. During inference, we set the temperature to 0.8 and limit the maximum generation length to 600 tokens. We report Pearson, Spearman, and Kendall-Tau correlation coefficients to measure the agreement between ground-truth scores and the predicted scores y_u in scalar ratings. For pairwise ratings, we report the accuracy for each aspect.

Baseline We establish a basic baseline that directly uses the average scores from historical reviews as the predicted score. For $k = 1$, we use the historical score as the output. This baseline is named **Reviewer Avg.**, reflecting the average score this reviewer gives based on historical comments. On Per-MPST, we add the baseline **Matrix Factorization (MF)** (Koren et al., 2009), commonly used in recommendation systems. It decomposes the user-item interaction matrix into the product of the user matrix and the product matrix, with the main idea being to recommend products based on the similarity between the user and the product. These two baselines do not provide an interpretable explanation for their evaluation. On Per-DOC, both the plot pairs and the annotators in the test set have no overlap with the instruction-tuning set, making the matrix factorization baseline unsuitable in this case. We also evaluate the capabilities of vanilla LLMs, including LLaMA-2-chat models from 7b to 70b and GPT-4 [4], using the same prompts and generation configurations.

5 Results and Analysis

We report the performance of the scalar rating on the test set of Per-MPST and the pairwise rating on Per-DOC. The reviewers in the test set $\mathcal{U}_{\text{test}}$ have no overlap with those in the instruction-tuning set $\mathcal{U}_{\text{ift}}$.

5.1 Key Findings

PerSE's Scalar Rating Achieves the Highest Correlation with Human Ratings. As shown in Table 2, **PerSE**-13b significantly outperforms all baselines in terms of correlations with human ratings. Specifically, **PerSE**-13b achieves a Pearson correlation of 0.345 between its predictions and human scores, indicating that **PerSE** effectively captures the reviewer's preferences from the given profile. The comparison between **PerSE** and the vanilla LLMs demonstrates that it is challenging for vanilla LLMs to align evaluations with personal preferences without personalized alignment instruction tuning. Moreover, we observe that both the average score of the reviewer's historical reviews and the simple baseline MF are strong baselines. This observation aligns with Kang et al. (2023), which shows that vanilla LLMs struggle to understand user preferences. One possible reason is that both the pre-training phase and instruction-tuning via reinforcement learning with human feedback (RLHF) focus on aligning the model towards objective and common human values, which hinders their ability to provide more personalized responses. This is also noted by Kirk et al. (2023), who claim that the aggregate fine-tuning process may not adequately represent all human preferences and values. However, **PerSE** demonstrates that this capability can be easily regained through instruction-tuning on a small amount of high-quality personalized alignment data.

PerSE's Explanation of Scale Rating Aligns

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

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**

Table 3: The comparison of the generated explanation and the human-written review on Per-MPST. A higher score indicates a better alignment between the generated explanation 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**

with Reviewer's Comments. We further investigate whether the explanations of the scalar ratings provided by **PERSE** align with the reviewer's comments. We use four widely used evaluation metrics in text generation to compare the explanation $\boldsymbol{e}_u$ with the ground-truth review text. These metrics include two 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)[5]. A higher score indicates that the generation is more similar to the reference. Table 3 shows that **PERSE**-7b and **PERSE**-13b outperform other baselines across all metrics. This suggests that **PERSE** can better model the preferences of a specific reviewer and generate a personalized review from this perspective.

PERSE's Pairwise Rating Outperforms Other Baselines on All Aspects of Per-DOC. As shown in Table 4, **PERSE** achieves the best performance across all aspects. Compared to **PERSE**-13b, **PERSE**-7b achieves comparable performance on `Surprise` but falls behind on other aspects. Similarly, the vanilla LLaMA models struggle with modeling personal preferences and rarely surpass the simple baselines, achieving around 50% accuracy on most aspects. This is partly due to having only $K = 1$ historical review for each reviewer, which requires the LLMs to have a strong capability in inferring personal preferences from a limited profile. Meanwhile, although GPT-4 demonstrates relatively high accuracy in capturing `Surprise`, its performance in other aspects is not satisfactory.

5.2 In-Depth Exploration

We conduct additional experiments to investigate personalization alignment in **PERSE**. Some of these experiments are detailed in Appendix E.

Personalized Evaluation of Open-ended Generation is Necessary. In Table 3, $K = 0$ indicates that there are no personalized examples in the instruction, meaning that p_u is empty and only the query $\boldsymbol{x}$ is provided in the prompt. This represents a 'one-score-fits-all' evaluation for open-ended generation, where the same score is predicted for all users given the same query. The poor performance of all baselines under this setting indicates that the variance in reviewers' preferences has a significant influence on the evaluation, highlighting that a 'one-score-fits-all' approach is ineffective for evaluating open-ended generation. We further calculate the review score variance of Per-MPST in Table 10 in the Appendix to emphasize the necessity of personalized alignment.

With a Richer Reviewer Profile, PERSE Aligns Better with the Personal Preference. We further explore how many reviews are required to establish a 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. As observed, with an increase in the number of historical reviews K in the profile, **PERSE**-13b's performance consistently improves. This indicates that after personalized alignment tuning, **PERSE** can better capture the personal preferences of historical reviewers. However, we also find that after 4 reviews, the performance of most baselines, including the average reviewer score baseline, declines. While a longer context provides more information about personal preferences, it also introduces challenges due to increased context complexity and noise. Therefore, we assume that increasing the number of historical reviews beyond a certain point may not further enhance **PERSE**'s performance.

More Historical Reviews Make PERSE More

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

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**

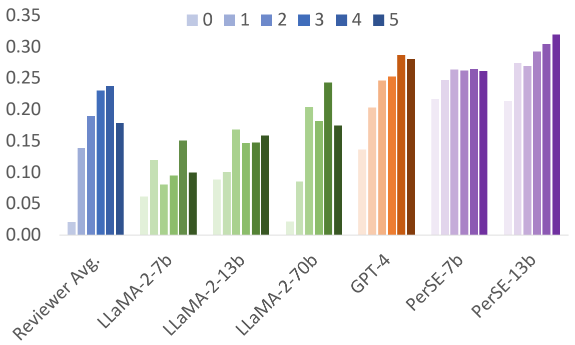


Figure 3: Kendall correlation on Per-MPST with different numbers of historical reviews (K) in reviewer profile. Having more reviews benefits **PerSE**-13b, but the increased complexity may harm the performance of the vanilla LLaMA models.

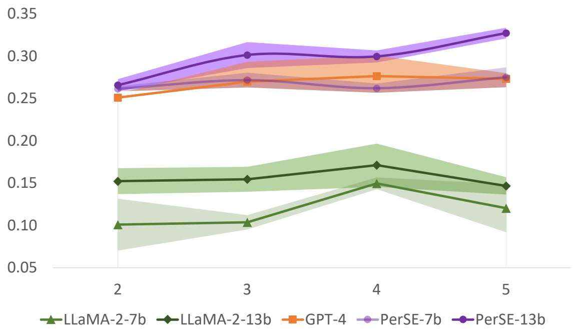


Figure 4: Kendall correlation on Per-MPST with different orders of reviews. The shadow indicates the variance while the line is the average performance among three trials. **PerSE** is more stable than baselines.

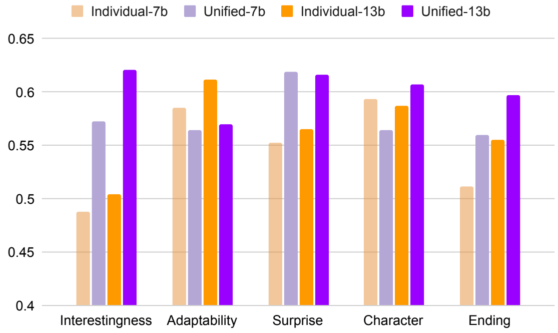


Figure 5: Accuracy of the unified and separate models on Per-DOC. Unified training improves performance.

Robust. Previous studies have shown that LLMs are sensitive to the perturbation of in-context examples (Lu et al., 2022). To investigate how robust **PerSE** is to the order of historical reviews, we randomly shuffle the historical reviews in the profile and rerun the experiments three times. We present the average performance with lines and the standard deviation as shaded regions in Figure 4. We can see that **PerSE**-13b consistently outperforms other baselines on average and has a smaller shaded region, indicating that **PerSE** is more robust to changes in the order of the profile than the other baselines. Furthermore, as the number of reviews increases, **PerSE**'s performance converges, suggesting that its performance is more stable with a larger profile. This implies that **PerSE** better captures the reviewer's preference and can coherently provide personalized scores for new queries without being affected by the order of reviews. In contrast, the vanilla LLaMA-2 models are more sensitive to order, showing a larger variance in the shaded regions.

Training with Ratings from Different Aspects Helps PerSE Better Understand Preferences. We investigate the influence of joint training on different aspects using Per-DOC. We compare the performance of a unified model, trained on all aspects together, with that of models trained separately on each aspect. As illustrated in Figure 5, joint training improves performance in most aspects, as exposure to different aspects allows the models to benefit from one another. For example, the capabilities to capture `Interestingness` and `Surprise`, as well as to evaluate the quality of `Ending`, are weaker in the individual setting. However, these aspects are enhanced during joint training, resulting in significant improvement. For separate models, they are better at capturing preferences for Adaptability and Character Development. We believe these two aspects are related to the plot's setting, which is more structured. This structure may lead to a clearer preference that is easier to capture with single-aspect data.

PERSE Generalizes Well to Other Domains without Fine-tuning. We evaluate the generalization of **PERSE** by applying it to a new domain (Amazon book review[6]) in a zero-shot manner. We recreate a personalized dataset from the Amazon book reviews for scalar rating using the pipeline described in Section 3.2. Detailed data statistics can be found in Appendix B. To fit the scoring range of the Amazon dataset, which is 1 to 5, we calibrate **PERSE**'s predictions (originally 1 to 10). We directly use **PERSE**, tuned on Per-MPST, to predict personalized reviews and scores for each book based on the user's preference. As shown in Table 5, **PERSE** outperforms other baselines even without fine-tuning on the new domain, indicating that **PERSE** can be effectively applied to new domains with limited or no fine-tuning data.

Personalized Tuning in PERSE Works Well with Other LLMs. In the main experiment, we use LLaMA-2 as the backbone LLM. Here, we also investigate whether our training process is applicable to other LLMs. We use the same data and training method to fine-tune Mistral 7B (Jiang et al., 2023). Results in Table 6 show that our method enhances the capability of the Mistral 7B model 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**

Compared with PERSE's Personalized Alignment, GPT-4's Generic Assessment Fails to Align with Specific Reviewer Preferences. In Figure 6, we present an example from Per-MPST. The annotated reviews reveal that this reviewer is critical of plots and particularly values novelty. However, even with this reviewer's preferences provided, GPT-4 predicts a positive review. We believe this is because GPT-4 is overly aligned towards safety and harmlessness, making it cautious in giving negative responses. While LLaMA-2-70b is stricter and assigns a score of 4, **PERSE** focuses more on the consistent terribleness and assigns a score of 3, which aligns more closely with the reviewer's true score. Moreover, we find that, unlike most people, this reviewer does not prioritize complicated themes. Despite this, GPT-4's "one-size-fits-all" evaluation offers a high score for such themes. **PERSE**, on the other hand, pays more attention to this reviewer's visual preferences, providing a more reviewer-specific rating. This indicates that **PERSE** can better evaluate stories based on personalized preferences rather than relying on a general and universally applicable evaluation principle without any personalized preference.

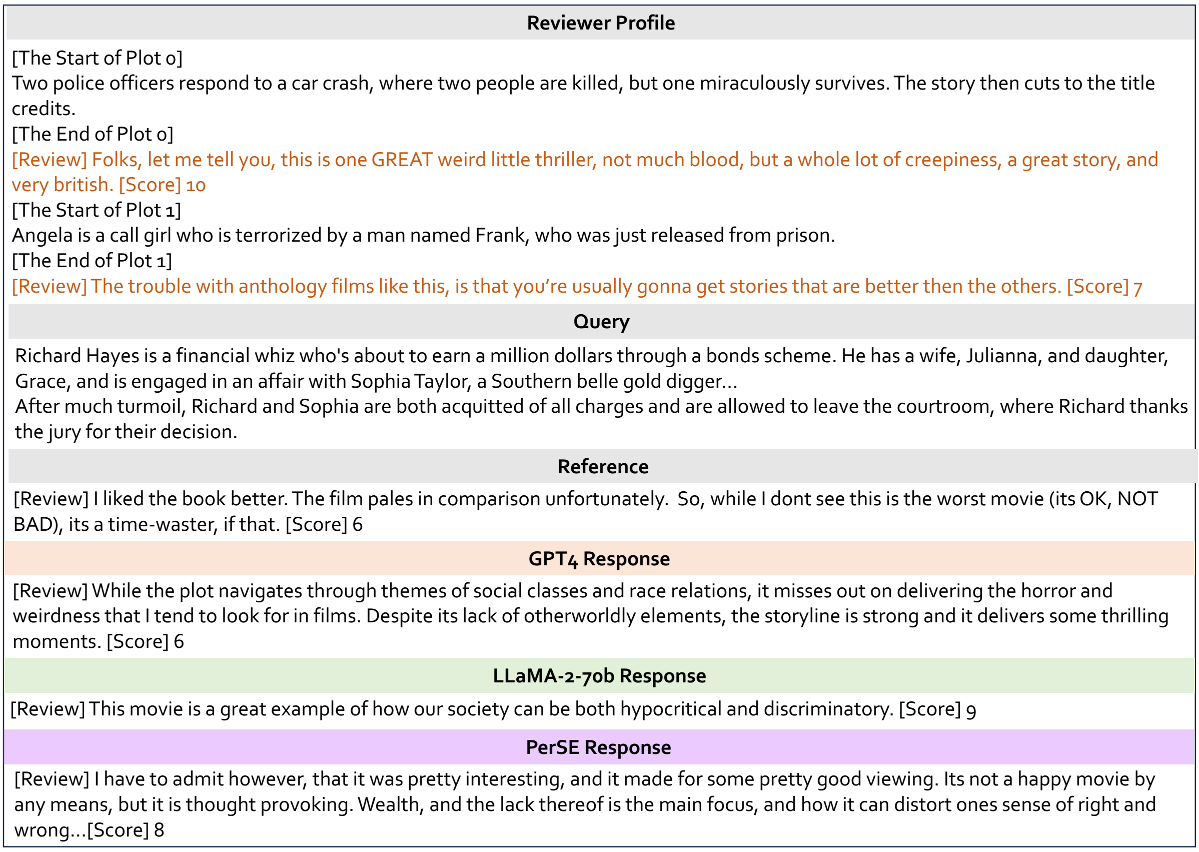
Table 6: Mistral-7b-based **PERSE** outperforms the original pre-trained model and achieves comparable performance with **PERSE**-7b. The setting is the same as Table 2.

	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

6 Conclusion and Discussion

In this paper, we focus on learning personalized alignment in evaluating open-ended generation. We introduce **PERSE**, an LLM-based evaluation model that provides an interpretable evaluation from the perspective of an unseen reviewer. It infers the reviewer's preferences based on their profile and applies these preferences to the evaluation of open-ended generation. In addition to providing a score, **PERSE** also offers a detailed explanation. By being instruction-tuned on personalized alignment data, the LLaMA-2-based **PERSE** outperforms GPT-4 in both scalar and pairwise ratings. Our comprehensive analysis of personalized alignment underscores the importance of personalized fine-tuning to avoid over-alignment with common human values, often a result of RLHF. The interpretability of **PERSE** makes it particularly well-suited for personalized generation and recommendation systems.

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



Reviewer Profile

[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

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

[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

[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

[Review] This movie is a great example of how our society can be both hypocritical and discriminatory. [Score] 9

PerSE Response

[Review] I have to admit however, that it was pretty interesting, and it made for some pretty good viewing. 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 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 E.

Limitation

While this research makes notable strides in addressing the challenge of personalized evaluation, it is not without limitations. For instance, we assume that preferences remain consistent across prior reviews, which may not account for changes in preferences in real-world scenarios. It would be interesting to model how preferences shift over time and evaluate content based on potential future preferences. Additionally, we demonstrate that with instruction tuning on personalization data, the smaller LLaMA-2 can outperform the larger GPT-4. More exploration could be conducted with large-scale LLMs to assess 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 **PerSE**, we remain ever-conscious of the ethical implications of our work.

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

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

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*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. *ArXiv*, abs/2306.04181.

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*.

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)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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

A Contamination Issues in Evaluation

Many existing evaluation datasets have been exposed to LLMs during their pretraining, leading to contamination issues when assessing model performance on these datasets. LLMs might achieve excellent performance on contaminated cases by memorizing the ground truth but perform poorly on others, rendering the assessment unreliable.

To address this, we investigate how such contamination affects LLMs when evaluating open-ended generation under three evaluation settings: scalar rating, pairwise rating, and personalized alignment. We evaluate GPT-4's performance on the IMDb dataset[7]. This movie dataset includes the synopsis and a score (1 to 10) for each movie. GPT-4 is tasked with identifying the movie based on the synopsis and evaluating the quality of the synopsis. We consider a movie to be memorized by GPT-4 if it can correctly predict the movie title from the synopsis. Additionally, we evaluate GPT-4's performance on a preprocessed version of the IMDb dataset that incorporates anonymization and summarization as described in Section 3.2.

A.1 Memorization in Scalar Rating

We ask GPT-4 to predict the score for a given synopsis. We then group the results based on the memorization status into two sets: 'Memorized' and 'Un-memorized'. We calculate the correlation between the predicted scores and the ground-truth scores. The results are presented in Table 7.

GPT-4's predictions show a very high correlation with the ground-truth scores for memorized cases. However, the performance drops dramatically for un-memorized cases. This indicates that the memorization issue makes the evaluation of GPT-4's performance unreliable: rather than analyzing the quality of the movie based on the given synopsis, GPT-4 relies on its memory of the score. We also find that the percentage of memorized cases significantly decreases after applying anonymization and summarization, demonstrating their effectiveness in alleviating memorization issues.

A.2 Memorization in Pairwise Rating

We create 200 movie pairs, each consisting of two movie synopses with scores differing by at least 1 point. We ask GPT-4 to identify the titles and

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

Table 7: Performance of GPT-4 in predicting a scalar rating for a single synopsis. Percent is the percentage of each type of synopsis (raw/anonymized/summarized) being recognized as ‘memorized’, or ‘unmemorized’. Memorization heavily affects performance, but its impact decreases with anonymization and summarization.

		Pearson	**Spearman**	**Kendall**	**Percent**
Memorized	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%
Un-memorized	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%

Table 8: GPT-4 in comparing two synopsis. 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. Overall, memorization leads to greater position bias and lower consistency.

		Accu. ↑	**Cons.** ↑	**Bias First** ↓	**Percent**
Both Memorized	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 Memorized	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 Memorized	Raw	/	/	/	0.0%
	Anonymized	0.500	62.5%	25.0%	4.0%
	Summarized	0.660	85.1%	4.3%	23.5%

predict which synopsis is better. We calculate prediction accuracy (Accu.), consistency (Cons.), and bias towards the first synopsis. Consistency measures how many judgments remain consistent after changing the order of the two plots. Bias towards the first is defined as an inappropriate preference for the first synopsis. It is calculated by subtracting the percentage where GPT-4 favors the first plot from the true percentage of the first being better.

Results are reported in Table 8. Similarly to scalar ratings, the neither-memorized group exhibits much lower accuracy compared to the other two groups, despite maintaining the main plot points. This indicates that memorization can lead to misleadingly high performance in evaluation. When GPT-4 memorizes one of the two plots, it is more consistent in its judgment and shows a lower position bias. This occurs because GPT-4 favors the memorized plot regardless of its order in the pair. The use of anonymization and summarization reduces the both-memorized cases to 42.5% and increases the neither-memorized cases to 23.5%.

We further calculate the ’Bias memorized’ by subtracting the percentage that GPT-4 favors the memorized plot from the true percentage where this plot is actually better. In Table 9, we observe that for all raw, anonymized, and summarized plots, GPT-4 shows a clear tendency to choose the memorized plot. This tendency is more pronounced in the summarized plots. We believe this is because data processing increases the uncertainty of the prediction, causing the model to be more conservative and rely on what it has memorized. However, GPT-4 also demonstrates high consistency and low position bias in the ’neither memorized’ group (see Table 8), indicating that when evaluating two novel stories, it can overcome the effects of memorization and assess based on the actual plots.

Table 9: Prediction on ‘One Memorized’ Group in pairwise comparison of GPT-4. The ‘Raw’, ‘Anonymized’, and ‘Summarized’ have the same meaning in Table 8.

	Bias Memorized
Raw	0.222
Anonymized	0.283
Summarized	0.397

A.3 Memorization in Personalization

We also explored the influence of memorization in personalized alignment during evaluation. We evaluate the performance of GPT-4 and vanilla LLaMA-2, as described in Section 3.1, on Per-MPST. We use $K = 1$ and calculate the Kendall correlation between human ratings and the predicted scores, as shown in Figure 7.

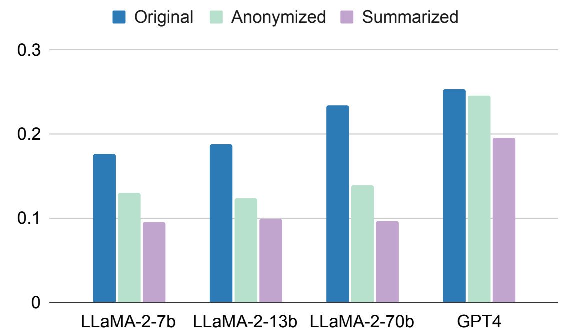


Figure 7: Kendall correlation between the LLM's personalized prediction with human ratings. Personalized predictions of all LLMs are also affected by memorization.

Similarly, LLMs achieved a high correlation with human ratings on the original synopses, but their performance degraded after anonymization and summarization. Although the main plots remain the same, with only slight differences in recognizable details, this greatly affected the results. Both experiments highlight that memorization introduces significant bias in LLM-based evaluation models, rendering them unreliable for both general and personalized evaluations.

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

B Constructing Personalized Alignment Datasets

To alleviate contamination issues, we create a less biased personalized evaluation dataset by anonymizing famous characters and summarizing existing plots. Our pipeline is illustrated in Figure 8. We use oasst-30b (Köpf et al., 2023), a 30B LLaMA-based model fine-tuned on OpenAssistant Conversations for alignment, to anonymize and summarize the plots.

Specifically, we anonymize the raw plot by asking the LLMs to identify character and local names and then create new names for them. Based on the JSON mapping generated, we replace those names with new ones. We do not directly ask LLMs to replace names to avoid potential hallucinations during the replacement. For characters with the same family names, LLMs can create new character names that still share the same last names (though not the original last names). For example, 'Glenn Holland' and 'Iris Holland' are mapped to 'William Thompson' and 'Emily Thompson'.

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 synopsis 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.

B.1 Variation of Score Preference across Reviewers

We present the variation of score preferences in Per-MPST in Table 10. We computed the count of reviewers for each query, along with the average (mean) and standard deviation (std) of the reviewers' scores. The table shows that while the average scores for queries are almost identical, there is a considerable standard deviation, highlighting the differences in reviewer opinions. This underscores the need for an evaluation method that accounts for varying preferences.

Table 10: Score variance between reviewers in Per-MPST

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

Amazon Book dataset Similar to Per-MPST, we preprocess the Amazon book dataset to create a personalized version. We use the 5-core subset of the book domain, where every user and item has a minimum of 5 reviews. The original instances in the Amazon book dataset only include the book

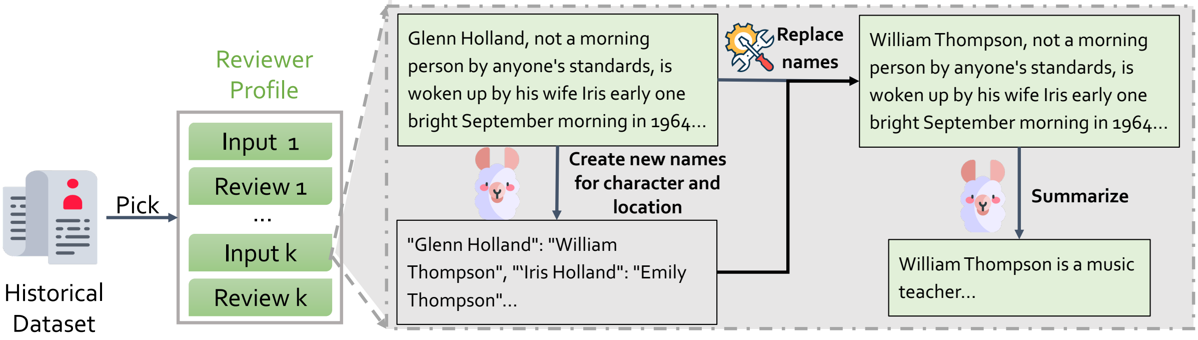

Reviewer Profile
Historical Dataset
Pick
Input 1
Review 1
...
Input k
Review k
Glenn Holland, not a morning person by anyone's standards, is woken up by his wife Iris early one bright September morning in 1964...
Create new names for character and location
"Glenn Holland": "William Thompson", "'Iris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

Figure 8: 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.

title without the content. Therefore, we use LLM-based retrieval to add a brief book description for each instance. Each example features one annotated review serving as the user profile ($K = 1$). Ultimately, we create a personalized version of the Amazon book dataset, consisting of 120 evaluation examples.

C Prompts

We demonstrate the framework and prompts of **PERSE** in Figure 9. The prompts used in Appendix A for addressing the contamination issue, as well as the prompts for anonymization and summarization, are listed in Table 11.

D Training Details

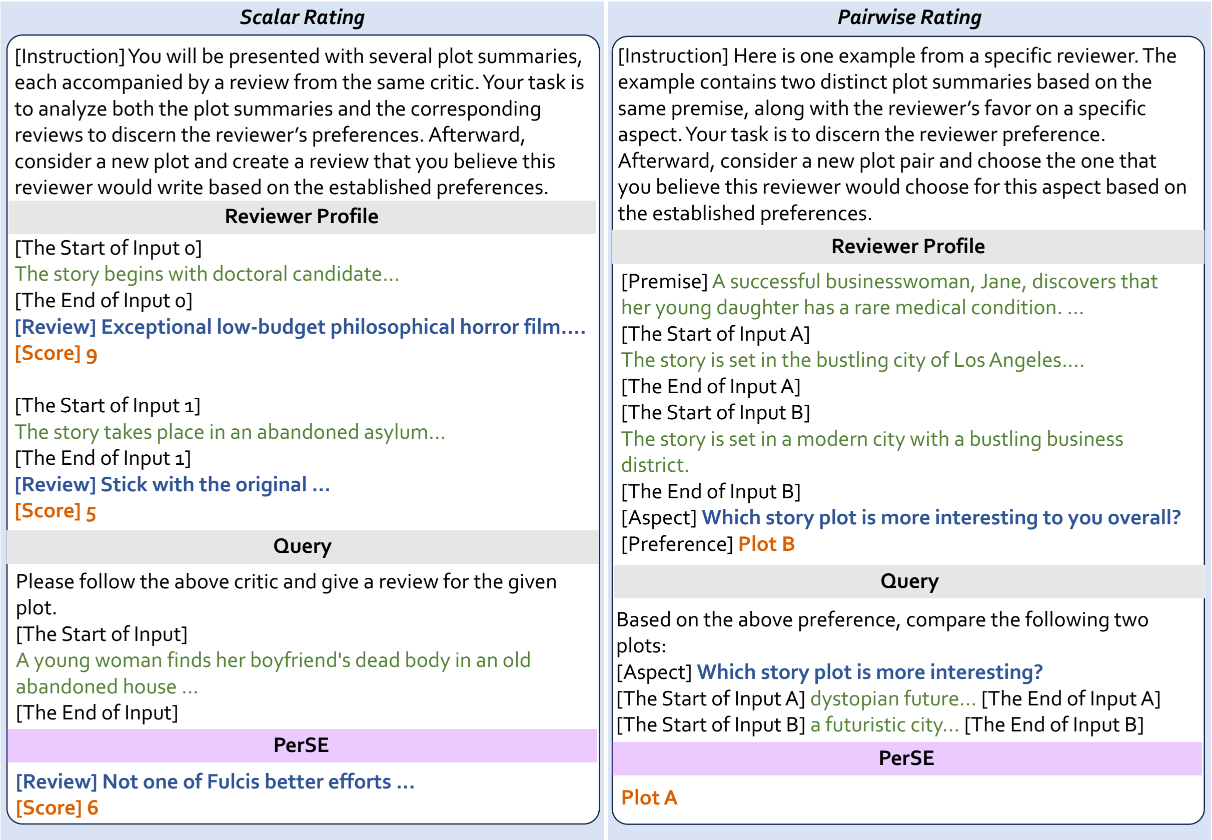
Each model in our experiments was trained on 8 x 80G A100 GPUs with a learning rate of 1e-5. We set the batch size to 4 for **PERSE**-7b and 2 for **PERSE**-13b. The models converged after 2k and 6k steps on Per-MPST, respectively. We trained two unified models on Per-DOC for all aspects by fine-tuning 7b and 13b LLaMA-2-chat. The models after 1k and 2k steps. It takes about 10 hours in total for these two models. For the ablation study, we also trained one model for each aspect on Per-DOC. Each model converged after 500 steps for 7b and 2k steps for 13b. The total training time was approximately 5 x 5 hours.

E More Case Studies

PERSE Infers Preferences Rather Than Copying Scores from Context. In Figure 12, we present another example from Per-MPST. From the reviews, we can see that the reviewer enjoys horror elements. However, the new plot and its level of terror are unsatisfactory, leading the reviewer to 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 suspect they are influenced by the high review scores in the reviewer's preference, overlooking the analysis of the new plot. In contrast, **PERSE** focuses on the dull aspects of the plot, aligning more closely with what the reviewer is concerned about. It assigns a score of 5, which differs from the existing review scores but is closer to the actual score the reviewer gave this plot.

PERSE Provides Diverse Reviews for the Same Plot Based on Different Preferences. In Figure 13, we illustrate the reviews of the same plot from two reviewers, A and B, each with different preferences. Both reviewers have read the book. Reviewer A is critical and has a high standard for good movies, leading to low scores in the annotated reviews. Consequently, he gives a score of 2 due to his disappointment with the movie adaptation. In contrast, Reviewer B is relatively tolerant and tends to give high scores. Although the movie is much worse than the book, he still gives a score of 6. However, GPT-4 and LLaMA-2-70b assign similarly high scores in both cases, disregarding the reviewers' preferences. However, **PERSE** is capable of providing personalized scores for different reviewers, predicting a score of 1 for Reviewer A and 8 for Reviewer B. Although the predicted score for Reviewer B is not as close as GPT-4's, it reflects the positive attitude captured by **PERSE**.

PERSE Achieves Better Performance on Fine-Grained Pairwise Rating. We illustrate an example from Per-DOC in Figure 14. **PERSE** successfully predicts the preference on 4 out of 5 aspects, while GPT-4 correctly predicts 3 aspects and LLaMA-2-70b succeeds on only 2 aspects. GPT-4 predicts Plot A for all aspects, ignoring the differences between aspects and providing an overall



Scalar Rating

[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 Profile

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

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

Query

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

PerSE

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

Pairwise Rating

[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 Profile

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

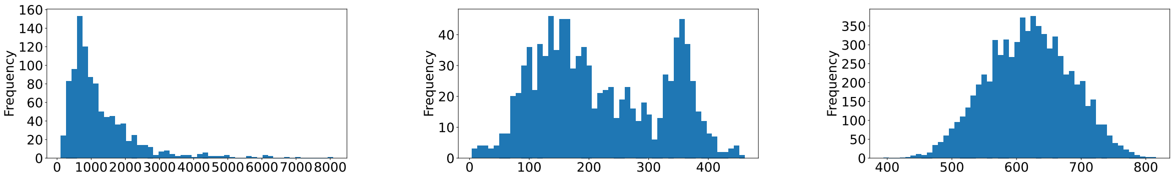
Query

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

PerSE

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.



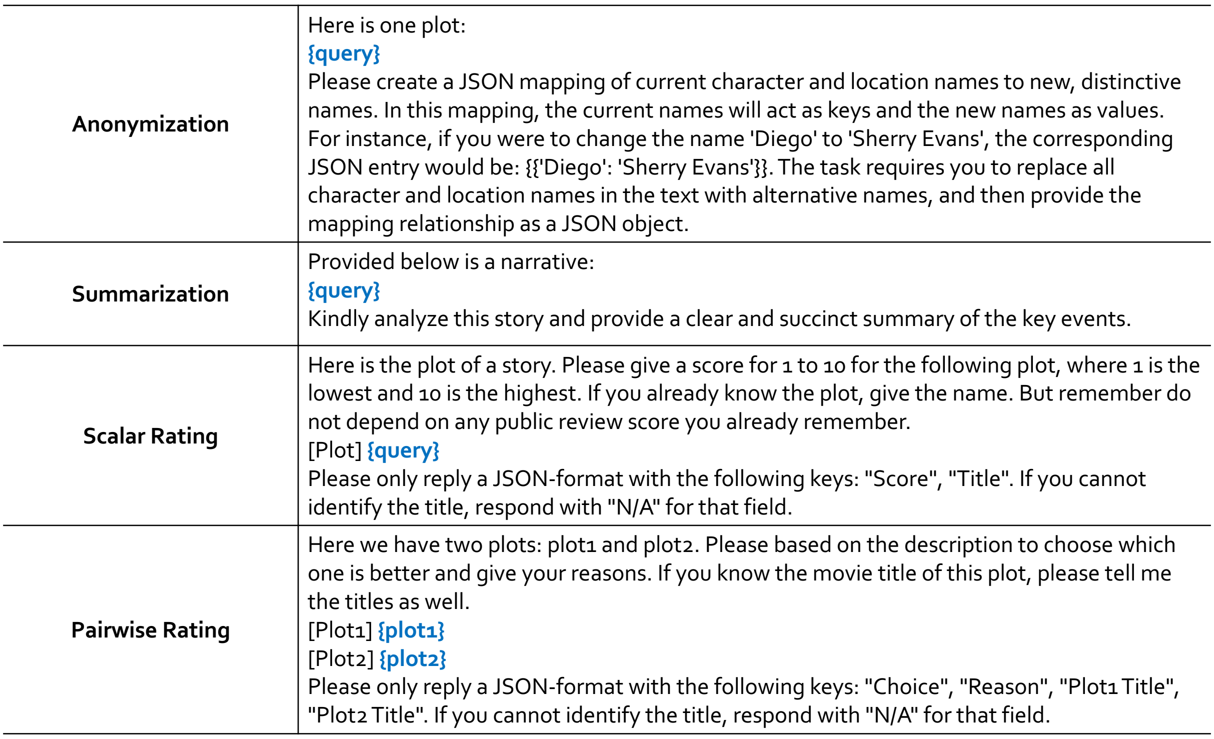
(a) Raw movie length in MPST v2.

(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.

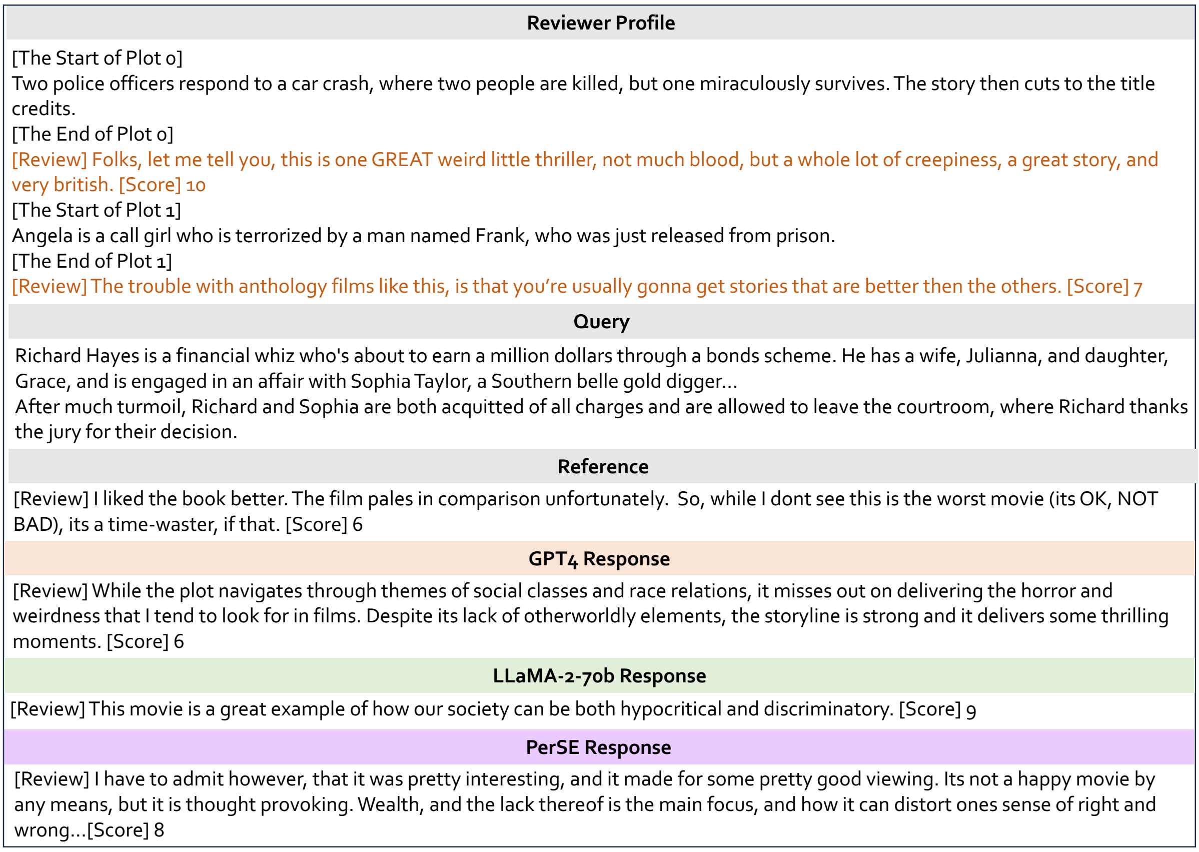
evaluation. In contrast, **PerSE** focuses more on the distinctive attributes of each aspect and makes judgments accordingly.



Anonymization	Here is one plot: **{query}** 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: **{query}** Kindly analyze this story and provide a clear and succinct summary of the key events.
Scalar Rating	Here is the plot of a story. 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] **{query}** 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 Rating	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 the data processing and investigation of contamination. The blue text is the placeholder for plots.



Reviewer Profile

[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

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

[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

[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

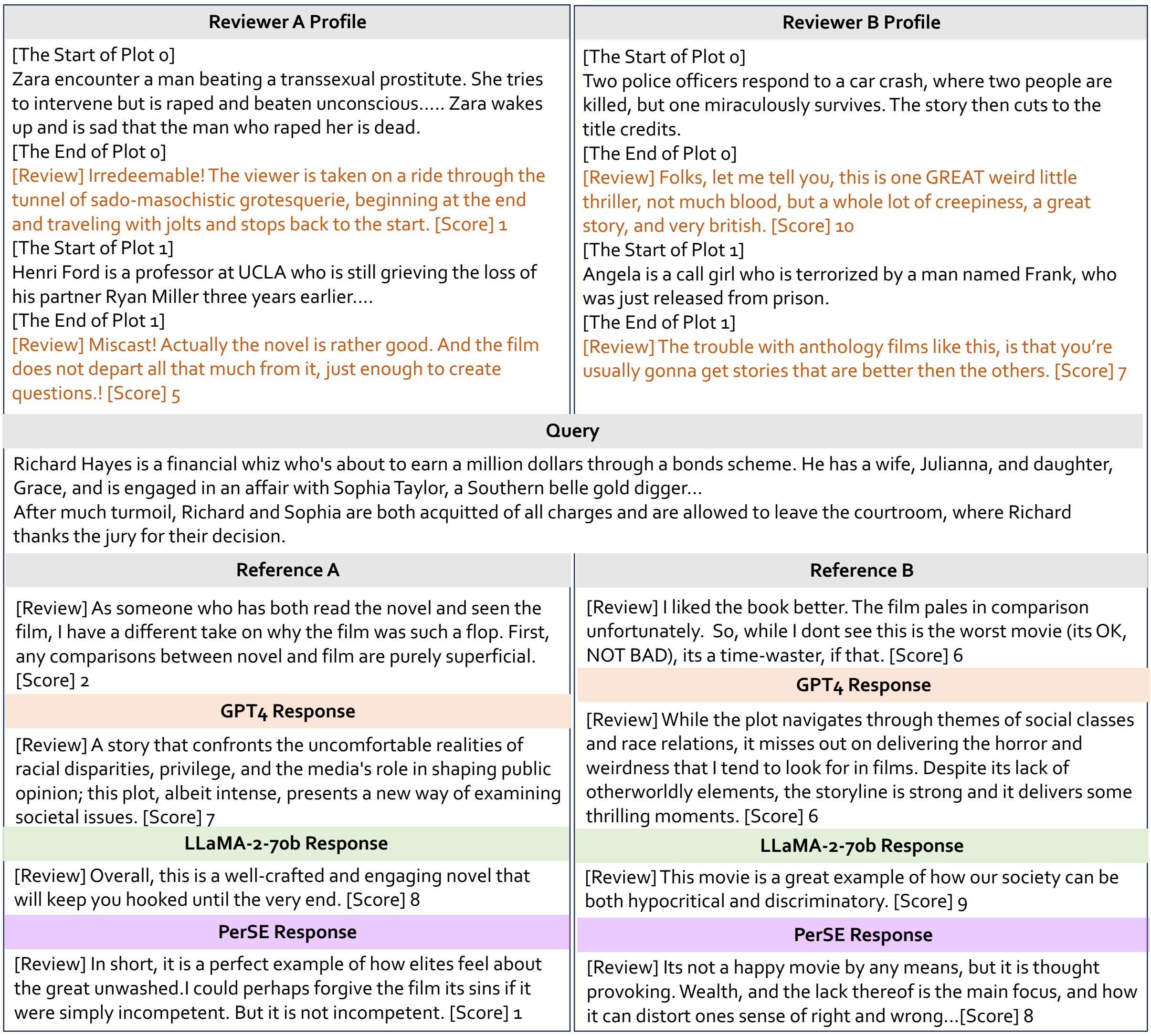
LLaMA-2-70b Response

[Review] This movie is a great example of how our society can be both hypocritical and discriminatory. [Score] 9

PerSE Response

[Review] I have to admit however, that it was pretty interesting, and it made for some pretty good viewing. 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 12: 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.



Reviewer A Profile

[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

Reviewer B Profile

[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

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

[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

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

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

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

Reference B

[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

[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

[Review] This movie is a great example of how our society can be both hypocritical and discriminatory. [Score] 9

PerSE Response

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

Reviewer Profile	[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**
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 14: 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.