

Split and Merge: Aligning Position Biases in LLM-based Evaluators

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown promise as automated evaluators for assessing the quality of answers generated by AI systems. However, LLM-based evaluators exhibit position bias, or inconsistency, when used to evaluate candidate answers in pairwise comparisons, favoring either the first or second answer regardless of content. To address this limitation, we propose PORTIA, an alignment-based system designed to mimic human comparison strategies to calibrate position bias in a lightweight yet effective manner. Specifically, PORTIA splits the answers into multiple segments, taking into account both length and semantics, and merges them back into a single prompt for evaluation by LLMs. Extensive experiments with six LLMs on 11,520 answer pairs demonstrate that PORTIA markedly enhances the consistency rates for all models and forms of comparison tested, achieving an average relative improvement of 47.46%. It also enables GPT-3.5 to achieve performance comparable to GPT-4 and elevates GPT-4’s consistency rate up to 98%. Subsequent human evaluations indicate that the PORTIA-enhanced GPT-3.5 model can even surpass standalone GPT-4 in terms of alignment with human evaluators, highlighting PORTIA’s ability to correct position bias, improve LLM consistency, and boost performance while keeping cost efficiency.

1 Introduction

Recent advances in large language models (LLMs) have achieved remarkable results on various tasks, sometimes even exceeding human performance (Kojima et al., 2022; Thapa et al., 2023). However, assessing the quality of LLM-generated answers poses challenges. Specifically, n-gram matching metrics like BLEU (Papineni et al., 2002) can quantify token-level overlap with reference texts but fall short in evaluating semantic quality. While human evaluators provide more accurate and

valuable feedback, often considered the “gold standards,” their scalability is generally low, given that they are costly and time-consuming. As a result, there emerges a growing need for automated evaluation methods that reliably align with human yet remain efficient and cost-effective.

Recently, researchers have investigated the use of powerful LLMs like GPT-4 (OpenAI, 2023) to evaluate the quality of text generated in response to open-ended questions (Zheng et al., 2023). Notably, robust LLM evaluators such as GPT-4 have been shown to align remarkably well with both controlled and crowdsourced human preferences, achieving over 60% agreement (Wang et al., 2023a). These studies suggest that LLMs can emulate human evaluations, offering a scalable and transparent alternative to the expensive and time-intensive human assessment of text quality.

While LLMs have advanced capabilities, they are not flawless evaluators and have been identified to possess certain biases. One notable bias is the position bias (Zheng et al., 2023; Wang et al., 2023a), in which an LLM might prefer either the first or second answer in a pairwise comparison, regardless of its content, as illustrated in Figure 1. Even the state-of-the-art GPT-4 model is not immune to position bias (Zheng et al., 2023; Wang et al., 2023a; Zhang et al., 2023; Zeng et al., 2023), and the behavior of its various versions can be inconsistent over time (Chen et al., 2023). Moreover, owing to pronounced position biases in less-powerful GPT models, much of the prior research (Zheng et al., 2023; Zhang et al., 2023) has been compelled to use the expensive GPT-4 for LLM evaluations, emphasizing the necessity for a more cost-effective approach to large-scale assessments.

To address these limitations, we propose PORTIA¹, an alignment-based system designed to calibrate position bias. Inspired by human long-text

¹The reason for the naming is provided in Appendix C.

reading strategies (Ratnasari, 2023), PORTIA splits the answers into multiple segments, aligns similar content across candidate answers, and then merges them back into a single prompt to feed to LLM evaluators. Specifically, PORTIA first identifies possible split positions at sentence boundaries within each answer. It then conducts a length alignment between the candidates to generate segments of roughly equal length across answers. If this length alignment does not yield a consistent verdict, PORTIA further undertakes an iterative semantic alignment to identify the optimal split positions, enabling the merging of segments across candidates. Since this lightweight approach does not require changes to the models themselves, PORTIA is readily adaptable to enhance a variety of LLM evaluators for improved evaluation consistency.

We conducted comprehensive experiments using six LLMs as evaluators to assess 11,520 answer pairs across three prevalent pairwise comparison forms. Our results show that PORTIA markedly boosts consistency rates for all the tested models and templates, achieving an average relative improvement of 47.46% and rectifying an average of 62.31% of the initially inconsistent cases. Furthermore, PORTIA addresses between 36% and 86% (over 80% for two-thirds of the comparison templates) of the position bias occurrences within the GPT-4 model, elevating its consistency rate up to 98%. Moreover, efficiency and cost evaluations indicate that PORTIA enables the less advanced GPT-3.5 model to achieve 88% agreement with the state-of-the-art GPT-4 model at merely 9.57% of the cost. Additionally, a user study involving five human participants demonstrated enhanced agreement between PORTIA-optimized evaluators and human evaluators. Remarkably, the agreement of human evaluators with PORTIA-enhanced GPT-3.5 even exceeds that with the standalone GPT-4. A subsequent ablation study suggests that PORTIA’s two key components — length alignment and semantic alignment — are beneficial for improving consistency across different comparison forms.

2 Background

Paradigms of Using LLM-based Evaluators. Recent work has explored using LLMs such as GPT-4 to evaluate and compare the performance of AI systems (Wang et al., 2023a; Chan et al., 2023; Zheng et al., 2023; Hada et al., 2023). Conceptually, there are two distinct LLM-based comparison

paradigms: *single-wise comparison* and *pairwise comparison*. In single-wise comparison, LLM evaluators are provided with one answer each time and are asked to score each answer independently, causing that position bias is not an issue in single-wise LLM evaluation and therefore beyond the scope of this paper. Nevertheless, we find that the absolute scores of LLM may lack clear interpretation. To demonstrate this, we conducted a preliminary study where we examined the consistency of single-wise comparison across a total of 80 test cases, each involving three sets of value ranges. Our findings indicate that the scores from single-wise comparison do not strictly adhere to a linear mapping relationship across different scales (more discussion in Appendix D).

Pairwise comparison presents two answers side-by-side and asks evaluators to select the superior one. In particular, pairwise comparison methods can be further categorized into three forms: *score-based*, *likert-based*, and *relation-based*. In score-based comparison, evaluators assign a score to each answer and then compare these scores to determine the better answer. The likert-based method (Rajani et al., 2023) requires evaluators to score answers on a likert scale (Likert, 1932), where lower scores indicate a strong preference towards the first answer, middle scores represent a close tie, and higher scores signal a preference for the second answer. Additionally, the relation-based comparison solicits direct inputs from the evaluators about their preference for one answer over another. This approach aims to avoid the use of potentially arbitrary scores, guiding evaluators to make relative comparisons between answers instead. The details of these three forms are shown in Appendix K.1.

Position Bias in Pairwise Comparison. Despite the generally encouraging performance of pairwise comparison methods, we note that LLM evaluators are not perfect and can exhibit certain biases. A primary concern is the *position bias* (Zheng et al., 2023; Wang et al., 2023a), whereby the LLM may favor the first (or second) answer in a pairwise comparison, regardless of its content. In fact, LLMs have shown notable sensitivity to small changes in prompts (Zhao et al., 2021; Zhu et al., 2023). For clarity, we provide a formal definition of position bias as well as the consistency. As illustrated in Table 12 (refer to Appendix K.1), the evaluation input comprises a fixed template with three placeholders. The input set for the LLM evaluators can be represented as $\{Q, R_1, R_2\}$, where Q denotes the ques-



[Question]: What are the most effective ways to deal with stress?		Original
Assistant A: (gpt-3.5-turbo)	Assistant B: (Claude-v1)	
Here are some effective ways to deal with stress: 1. Exercise regularly: Physical activity can help reduce stress and improve mood. 2. Practice relaxation techniques: XXX 3. Get enough sleep: XXX 4. Eat a healthy diet: XXX 5. Connect with others: XXX 6. Time management: XXX 7. Seek professional help: XXX	Here are some of the most effective ways to deal with stress: • Exercise regularly. Exercise is a great way to release pent up stress and tension. XXX • Practice mindfulness. XXX • Practice deep breathing and relaxation techniques. XXX • Get enough sleep. XXX • Eat a healthy, balanced diet. XXX • Spend time with others. XXX • Limit unhealthy habits. XXX • Seek professional help if needed. XXX • Take periodic breaks. XXX	
Assessment		
<input type="checkbox"/> <input type="checkbox"/> ,which one is better? => <input type="checkbox"/> is better. <input type="checkbox"/> <input type="checkbox"/> ,which one is better? => <input type="checkbox"/> is better.		 Inconsistent Result
[Question]: What are the most effective ways to deal with stress?		PORTIA
Assistant A: (gpt-3.5-turbo)	Assistant B: (Claude-v1)	
Here are some effective ways to deal with stress: 1. Exercise regularly: Physical activity can help reduce stress and improve mood. 2. Practice relaxation techniques: XXX 3. Get enough sleep: XXX 4. Eat a healthy diet: XXX 5. Connect with others: XXX 6. Time management: XXX 7. Seek professional help: XXX	Here are some of the most effective ways to deal with stress: • Exercise regularly. Exercise is a great way to release pent up stress and tension. XXX • Practice mindfulness. XXX • Practice deep breathing and relaxation techniques. XXX • Get enough sleep. XXX • Eat a healthy, balanced diet. XXX • Spend time with others. XXX • Limit unhealthy habits. XXX • Seek professional help if needed. XXX • Take periodic breaks. XXX	
Assessment		
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> ,which one is better? => <input type="checkbox"/> is better. <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> ,which one is better? => <input type="checkbox"/> is better.		 Consistent Result

Figure 1: A sample pairwise LLM-based evaluation improved by PORTIA. *Left*: The original evaluation exhibiting inconsistency. *Right*: Consistent evaluation after applying PORTIA. Details of the answers, comparison forms, and evaluation contents have been simplified or omitted for clarity. An example of the detailed prompt is given in Table 15. It is worth noting that the answers from different assistants may differ significantly from each other.

tion set, and R_1 and R_2 are the two sets of answers for comparison. The LLM evaluators produce the verdict $V = LLM(\{Q, R_1, R_2\})$, which indicates the preferred answer out of the two candidates. Assuming that the LLM evaluators are flawless, the verdict V should be independent of the permutation Π of R_1 and R_2 . Thus, position bias can be expressed as: $\Pi \not\perp V$. On an individual sample level, for a specific question $q \in Q$ and answers r_1, r_2 , consistency is achieved if the verdict v remains the same when the positions of r_1, r_2 are switched: $LLM(\{q, r_1, r_2\}) = LLM(\{q, r_2, r_1\})$.

3 The PORTIA System

Design Intuition. It is worth noting that both human evaluators and LLMs encounter difficulties in making consistent evaluations when faced with lengthy and intricate answers (Kintsch and Keenan, 1973; Wijesiriwardene et al., 2023). A common cognitive approach among individuals is to decompose information into smaller units, thereby simplifying the comparison process (Ratnasari, 2023). Inspired by this observation, PORTIA is designed to split candidate answers into segments, merge specific segments across candidates that share “comparable” content, and eventually align them. Based on this intuition, PORTIA seeks to mimic effective human comparison procedures, aiming to calibrate position bias and enhance the consistency of LLM evaluators. That said, for each question, the verdicts of PORTIA should be consistent with any permutation of the answers, i.e., $LLM(\{q, r_1, r_2\}) = LLM(\{q, r_2, r_1\})$.

Moreover, PORTIA is specifically designed to address the position bias issue, rather than aligning the reasoning capabilities of LLM-based evaluators. Therefore, if an LLM evaluator lacks the capability

to adequately judge a pair of semantically different responses, this falls outside the scope of PORTIA’s intended purpose. PORTIA is solely focused on rectifying inconsistencies that arise when the order of a pair of semantically different responses is altered, leading to a change in the judgment of the LLM evaluator. In this context, PORTIA is generalizable to handling open-ended answers because, regardless of their semantic differences, we can always mix the two previously individual answers. Such a mixing operation effectively eliminates the position bias in pairwise LLM-based evaluation.

3.1 Key Design Considerations

Before presenting the technical details of PORTIA, we first introduce its key design considerations.

Content Preservation. Content preservation refers to ensuring the segmented answers encompass the entirety of the information present in the original answer, without any omissions or additions of new content. For a given original answer r_1 , the set of split answer segments $\{r_1^1, r_1^2, \dots, r_1^k\}$ should fully encompass the content of r_1 . This implies that when the segments are concatenated, the entirety of the original content is preserved ($\sum_{i=1}^k r_1^i = r_1$). This consideration helps to preserve the meaning and information of the original answer during the process of splitting. The preservation of content is critical for evaluators to assess the same substantive answer content that is divided into segments, without any alterations or incomplete information.

Order Preservation. Order preservation refers to preserving the original sequence of the information presented in the answer. This is important for fair evaluation, as re-ordering or re-arranging the content may impact the assessment of answer

quality. For example, if the sequence of actions (i.e., answer) in response to the question “What should you do in the morning after waking up?” is re-ordered such that “eat lunch” before “brush teeth,” it may be perceived as an answer of lower quality. By preserving the order, we ensure the segmentation process does not introduce artifacts that could unintentionally alter assessment. This enables the LLM evaluators to accurately evaluate answers in comparison to the original.

Resource efficiency. Resource efficiency refers to minimizing computational costs incurred by the splitting process, beyond the standard cost when querying the LLM evaluator. To this end, it is important for the segmentation process to introduce a minimal number of extra tokens and to be executed rapidly, thus avoiding significant overhead.

3.2 The Core Splitting Algorithm

Due to the page limit, we direct interested readers to Appendix E for a comprehensive overview of utilizing PORTIA for LLM evaluation. Here we concentrate on PORTIA’s core splitting algorithm, as illustrated in Algorithm 1. Intuitively, PORTIA first identifies semantically or syntactically similar segments across answers. It then aligns these answer segments and merges them sequentially into a single prompt for the LLM evaluators to make a final verdict. Specifically, the inputs include the question q , two candidate answers r_1 and r_2 , the LLM evaluator’s verdict function $v()$, and the specified number of splits k . The output of Algorithm 1 is a consistent verdict $v \in (1, 2, 3)$, where 1 indicates that r_1 is superior, 2 suggests that r_2 is better, and 3 represents a tie.

Overall, the splitting process can be divided into three stages. In the first phase, possible split positions are determined at the boundaries of sentences (line 1). Segmenting at sentence breaks (e.g., periods or question marks) reduces the likelihood of producing incomplete words or fragmented syntactic units in different segments. This particular design decision aids in maintaining semantic consistency and enhancing readability in each segment. Notably, natural language and programming language have different definitions for sentence boundaries; for instance, the period sign “.” in Python denotes accessing a specific object member property. Therefore, in instances where answers involve code blocks, we leverage `tree-sitter` (`tre`) to parse code blocks and locate suitable split positions that preserve the code’s structure and execution sequence.

This allows PORTIA to split lengthy pieces of code into smaller, logically connected units to facilitate more accurate comparison.

The second stage performs length alignment, splitting each answer into k segments of comparable length (line 2). Specifically, we first find the $k - 1$ points that divide the answer into k equal segments according to the number of characters. Subsequently, we select the split location that is closest to each of the split positions obtained in the first stage, and designate them as $[r_1^{(1)}, \dots, r_1^{(k)}]$.² The k corresponding answer segments are subsequently merged again and used for evaluation by the LLM evaluator. If the LLM evaluator consistently returns the same verdicts for all length-aligned splits, then the verdict is returned (lines 3-5).

Algorithm 1: Alignment-based Splitting

Input: Question: q , Answers: r_1, r_2 , Evaluator’s verdict $v()$, Split number k

Output: Consistent evaluation $v \in (1, 2, 3)$

/* Step1: identify answers’ formats with split positions. */

1 $r_1^{positions} = format(r_1), r_2^{positions} = format(r_2)$

/* Step2: length alignment. */

2 $[r_1^{(1)}, \dots, r_1^{(k)}] = equalsplit(r_1^{positions}, k), [r_2^{(1)}, \dots, r_2^{(k)}] = equalsplit(r_2^{positions}, k)$

3 **if** $v(q_i, r_1^{(1)}, r_2^{(1)}, \dots, r_1^{(k)}, r_2^{(k)}) == v(q_i, r_2^{(1)}, r_1^{(1)}, \dots, r_2^{(k)}, r_1^{(k)})$ **then**

4 | **return** v /* Consistent, return answer */

5 **end**

/* Step3: semantic alignment. */

6 **else**

7 | $s_{max} = 0, n_s = 0, Search_all = False, r_1^{bestparts} = [], r_2^{bestparts} = []$

8 **while not** $Search_all$ **do**

9 | $r_1^{parts} = partition(r_1^{positions}, k, n_s), r_2^{parts} = partition(r_2^{positions}, k, n_s), n_s += 1$

10 | $s_{cum} = \sum_{i=1}^k similarity(r_1^{parts}[i], r_2^{parts}[i])$

/* Update max similarity score, keep best split positions. */

11 | **if** $s_{cum} > s_{max}$ **then**

12 | | $s_{max} = s_{cum}, r_1^{bestparts} = r_1^{parts}, r_2^{bestparts} = r_2^{parts}$

13 | **end**

14 **end**

15 **if** $v(q_i, r_1^{(1)}, r_2^{(1)}, \dots, r_1^{(k)}, r_2^{(k)}) == v(q_i, r_2^{(1)}, r_1^{(1)}, \dots, r_2^{(k)}, r_1^{(k)})$ **then**

16 | **return** v

17 **end**

18 **end**

²An illustration with two detailed algorithms is available in Appendix H to ease understanding.

If inconsistent assessments persist after length alignment, PORTIA proceeds to semantic alignment as the third stage (lines 7-14). Specifically, given a fixed k and a set of possible split positions, we aim to iteratively search for the optimal split positions that maximize the cumulative semantic similarity between corresponding segments of the two answers. Note that n_s represents the index number of the current segmentation, and *Search_all* becomes True when n_s reaches the maximum number of possible split combinations *Cal*. Semantic similarity between segments r_1^t and r_2^t is computed by token overlap: $sim_score = \frac{Intersection(set(r_1^t), set(r_2^t))}{\max(len(set(r_1^t)), len(set(r_2^t)))}$. Notably, the choice of value k as well as the similarity metric would have an impact on the efficiency of PORTIA, and we provide the theoretical analysis in Section 4.3. We also consider applying other similarity metrics, such as LM-based metrics (Reimers and Gurevych, 2019). However, we argue that employing such intricate metrics is not necessary for PORTIA, as they usually entail extra computing resources, and introduce more hyper-parameters while yielding only marginal improvements in performance; see further discussion in Appendix I. Finally, PORTIA would yield consistent verdict if applicable (lines 15-17). Note that the above three stages are carried out in a sequential manner, whereas semantic alignment is only performed when length alignment is inadequate for ensuring consistent assessments.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate PORTIA using the MT-Bench benchmark (Zheng et al., 2023), following the experimental setup in (Wang et al., 2023a). MT-Bench contains 80 elaborated questions spanning 8 categories (Writing, Roleplay, Reasoning, Math, Coding, Extraction, STEM, and Humanities). For each question, MT-Bench provides several candidate answers from different LLMs. We consider eight different combinations of LLM answers (see more details in Appendix G), and we consider all three comparison forms (score-based, likert-based, and relation-based) in the pairwise comparison paradigm. Thus, we have $80 * 8 * 3 = 1920$ inputs to evaluate each LLM evaluator. We interpret the datasets as large and diverse enough to provide a comprehensive evaluation of PORTIA across different LLMs and comparison forms.

Models. In this work, we include both locally deployable models that are open-source and proprietary models that are accessed through only cloud APIs as LLM evaluators. Details on the specific LLM versions evaluated are given in Appendix G. **Response Length and generalizability.** In Figure 1, two assistants provide responses of similar length. However, given the open-ended nature of the questions in MT-Bench, different LLMs may produce responses that differ substantially in both length and content for the same question. To assess PORTIA’s adaptability to open-ended questions, we analyzed the statistics of all responses, presented in Table 3. Our findings indicate that the lengths of responses from the LLMs vary considerably, underscoring PORTIA’s flexibility in handling open-ended questions. Additionally, we explored PORTIA’s generalizability by examining the relationship between answer length and inconsistency and evaluated its performance on extremely short responses. More details are provided in Appendix B.

4.2 Main Results

As shown in Table 1, PORTIA improves the consistent rate among all evaluators. The values depicted in the table correspond to the mean values obtained from the analysis of all eight combinations of tested models. We observe that PORTIA relatively improves the consistent rate by 3.25% to 194.48%, depending on the evaluator, with the highest fixed coverage at 96.32% (meaning that nearly all the inconsistent results are resolved). GPT-4 exhibits the highest average consistency rate, which is in line with the findings of previous work (Wang et al., 2023a), and PORTIA further boosts its consistency up to 98%. Moreover, we observe that GPT-4 exhibits subpar performance on the likert-based form, not just compared to its performance on other forms, but also when compared to GPT-3.5. Upon analyzing results on likert-based forms, over 78% of GPT-4’s inconsistency provides a score of 5, reflecting its bias for the second answer, and our method rectifies 36.09% of them. Notably, we only report the results of Llama2 in relation-based form, as it fails to provide meaningful evaluations in other forms (see more details in Appendix J).

The impact of the comparison form on consistency rates is also observed, with evaluators displaying various preferences. For instance, it is seen that GPT-3.5 exhibits the least consistent performance when evaluated on the score-based form, whereas Claude2 struggles most on the relation-

Evaluators	De. Method	Model	Relation-based	Score-based	Likert-based
Claude2	API	% Origin Con	28.28	47.34	50.62
		% PORTIA Con	83.28 ($\uparrow 194.48\%$)	65.16 ($\uparrow 37.64\%$)	94.84 ($\uparrow 87.36\%$)
		% Fixed Coverage	79.44	52.22	91.27
Qwen	API	% Origin Con	63.12	52.66	8.12
		% PORTIA Con	78.13 ($\uparrow 23.78\%$)	71.09 ($\uparrow 35.0\%$)	9.38 ($\uparrow 15.52\%$)
		% Fixed Coverage	65.66	59.78	6.46
Chatglm2	Local	% Origin Con	38.44	58.59	26.72
		% PORTIA Con	61.72 ($\uparrow 60.56\%$)	74.06 ($\uparrow 26.4\%$)	64.22 ($\uparrow 140.34\%$)
		% Fixed Coverage	56.09	51.02	60.30
Llama2	Local	% Origin Con	36.41	N/A	N/A
		% PORTIA Con	68.75 ($\uparrow 88.82\%$)	N/A	N/A
		% Fixed Coverage	22.51	N/A	N/A
GPT-3.5	API	% Origin Con	78.12	39.22	78.91
		% PORTIA Con	88.59 ($\uparrow 13.4\%$)	54.84 ($\uparrow 39.83\%$)	98.60 ($\uparrow 24.94\%$)
		% Fixed Coverage	70.63	42.06	96.32
GPT-4	API	% Origin Con	93.44	92.75	61.50
		% PORTIA Con	97.03 ($\uparrow 3.84\%$)	98.00 ($\uparrow 5.66\%$)	63.50 ($\uparrow 3.25\%$)
		% Fixed Coverage	80.99	86.33	36.09

Table 1: The main results of PORTIA across LLM evaluators. All metrics presented are higher-is-better values. “% Origin Con” and “% PORTIA Con” are the percentages of consistent results in the original setting when enhanced by PORTIA, respectively. “% Fixed Coverage” denotes the percentage of inconsistent original assessments that are later corrected by PORTIA. “De Method” specifies whether the LLM evaluator uses local or cloud API deployment.

based form. GPT-4, Qwen, and Chatglm2 exhibit the highest degree of inconsistency when assessed on the likert-based form. This suggests that appropriately matching comparison forms to evaluators’ capabilities is important. Nevertheless, PORTIA offers high enhancement for forms and LLM evaluators. The substantial improvements highlight the generalizability of PORTIA. In summary, these findings clearly validate PORTIA’s effectiveness at mitigating inconsistency for both cutting-edge and less powerful LLM evaluators.

4.3 Efficiency and Cost Analysis

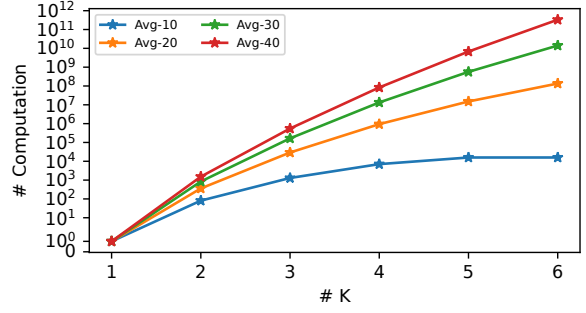
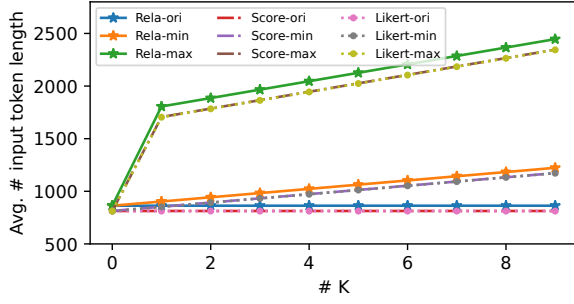
To show the efficiency and cost-effectiveness of PORTIA, this section first performs a theoretical analysis of PORTIA’s efficiency, and then evaluates its actual costs in terms of temporal, monetary, and environmental factors. Specifically, we measure the efficacy of PORTIA-enhanced LLMs in terms of their agreement rate with GPT-4.

Theoretical analysis. We first theoretically analyze PORTIA’s computational efficiency. As the number of answer segments k increases, the average input length for LLM evaluators also grows correspondingly. In line with line 1 in Algorithm 1, the added tokens stem from two sources: fixed-length system prompts based on comparison forms, and split boundary prompts (an example shown in Table K.2) that scale linearly with k . Consequently, the additional input length scales as $O(K)$, as depicted

in Figure 2(a). Notably, the average input length of the relation-based form exceeds the other two, as it requires more tokens for the system prompt. More details are in Appendix K.

In accordance with line 9 in Algorithm 1, the total number of computation operations Cal is calculated as: $Cal = C_{p_1}^{k-1} * C_{p_2}^{k-1}$, where p_1 and p_2 are the potential split positions in the two answers. $C_{p_1}^{k-1}$ and $C_{p_2}^{k-1}$ are the combination counts for the first and second answers, respectively. Using average position numbers of 10, 20, 30, and 40, we derive the total calculations as depicted in Figure 2(b). Intuitively, raising the value of k can improve the algorithm’s performance by exploring more split position combinations. However, this also results in an exponential surge in the total computation operations, compromising efficiency. As such, we conducted controlled experiments to identify the optimal value of k , and in our case, we found that setting $k = 3$ strikes a balance between efficiency and precision. Full details about this controlled experiment can be found in Appendix I.

Real-World Performance and Cost Analysis. Next, we measure the level of agreement between the PORTIA-enhanced LLM evaluators and GPT-4 (considered as the “gold standard.”). Note that to offer a fair evaluation, we exclusively consider GPT-4 evaluation outputs that are originally consistent. In the context of a question with two



(a) Average input token length with different k . “ori” represents the original input length. “min” and “max” represent the minimum and maximum input lengths, respectively.

(b) Computation operations with different k . The number behind “Avg” is the average potential split positions for each answer. The y-axis has been logarized to ease reading.

Figure 2: Theoretical estimation of PORTIA’s cost with varying k based on input length and computation operations.

possible answers, it is deemed as an agreement only when both GPT-4 and PORTIA-enhanced assessments are consistent and identical. As evidenced in Table 8, agreement rates are enhanced by an average of 16.32% after alignment. Claude2 has the highest gain at 31.65%, while GPT-3.5 achieves the highest agreement rate with GPT-4 at 88.59%.

Additionally, we consider the resource usage in terms of temporal, monetary, and environmental factors. As shown in Table 8, Chatglm2 exhibits the lowest inferencing time. However, the cost of GPT-3.5 is lower than that of Chatglm2, while its carbon emission is higher, which is mainly because the cloud API models usually run on GPU clusters with more powerful GPUs. It is worth mentioning that GPT-3.5 incurs less than **10%** of the average cost of GPT-4, while maintaining an approximate agreement level of 88% with GPT-4. In brief, the usage of PORTIA results in a substantial level of concurrence with GPT-4 while maintaining a minimal computational burden, hence showcasing a proficient and eco-friendly alignment. The significant enhancements in performance and resource utilization underscore the usefulness of this approach in boosting various LLMs for crucial evaluation work.

4.4 Human study

We conducted a human evaluation to further assess the performance of PORTIA. The model pair “gpt-3.5-turbo” v.s. “Claude-v1” is selected to compare human agreement rates on original versus PORTIA-enhanced assessments across 80 questions, as these two models have similar performance (Zheng et al., 2023), making it challenging for LLM evaluators to make decisions. We recruit five experts, including two industrial developers and three academic researchers as participants. For each participant, we create an online questionnaire that provides one question with two answers, not specifying their ori-

gin. Before the questionnaire, brief instructions on the task and evaluation criteria are provided. During the human evaluation process, we observe some instances where human evaluators make directly opposing assessments. This highlights the inherent subjectivity and unreliability of human evaluation. We attribute these disagreements to the diversity of human values (Peng et al., 1997), and simply use a majority vote to determine the final result.

	Ori HAR (%)	Fix HAR (%)
GPT-3.5	55.00	63.75
Qwen	35.00	35.00
Chatglm2	16.25	17.50
Claude2	6.25	47.50
GPT-4	60.00	65.00

Table 2: Main results from human evaluation comparing the model pair “gpt-3.5-turbo” v.s. “Claude-v1” on 80 questions. “HAR” represents the human agreement rate.

The human evaluation results presented in Table 2 demonstrate increased agreement rates between humans and LLM evaluators after applying PORTIA. On average, human agreement on original LLM assessments improves by 11.25% after enhancement. Notably, the original human agreement rate for Claude2 is only 6.25%, but increases substantially to 47.50% after enhancement. In addition, while the original human agreement lags behind GPT-4 across evaluators, PORTIA-enhanced GPT-3.5 surpasses the original GPT-4, indicating enhanced consensus. Taken together, these quantitative findings provide evidence that PORTIA effectively augments the assessments of all LLM evaluators to achieve greater concordance with human evaluators. The framework also enables weaker LLMs to reach comparability with stronger counterparts in terms of human alignment.

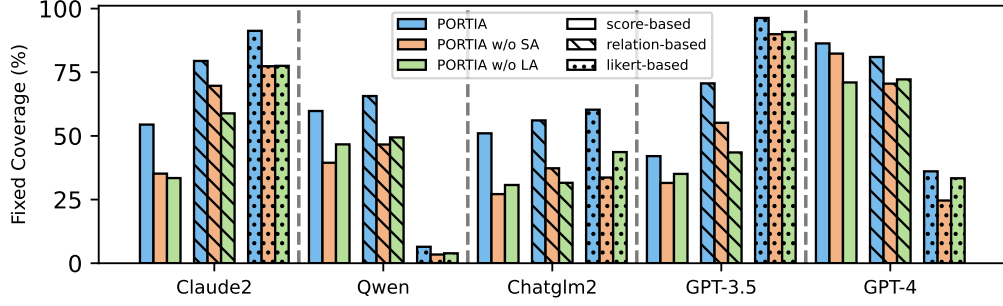


Figure 3: Fixed coverage rate across LLMs for PORTIA and variants w/o Semantic (SA) or Length Alignment (LA).

4.5 Ablation Study

To ascertain the individual contributions of each component in PORTIA, we conduct ablation experiments on five distinct LLM evaluators. The results are quantified in terms of the “Fixed Coverage” metric, as depicted in Figure 3. To facilitate visual interpretation, variants of PORTIA incorporating or excluding specific components are denoted by different colored bars in the histogram. Additionally, texture patterns in the bars indicate the comparison form used. The plain blue bar represents the score-based form, while the blue bar with slash lines corresponds to the relation-based form.

The results reveal that both semantic and length alignment improve PORTIA’s performance. Specifically, across all evaluators, semantic alignment shows a greater contribution to enhancing the likert-based form, possibly because the likert scale has a greater dependence on precise semantic meaning for its standardized categorical ratings. For the other two forms, both alignment methods have similar contributions with slight differences between LLMs. This is likely because these forms consider semantic meaning and answer length in a balanced way, similar to how human evaluators would.

Furthermore, we find that the trends of fixed coverage rate are consistent across comparison forms for PORTIA and ablations (without semantic or length alignment). Likert-based form has the highest fixed coverage rates, followed by relation-based, with score-based form having the lowest rates. The exceptions are Qwen on likert-based form and GPT-4 on all forms, where we manually check and find that: (1) Qwen prefers the second answer for over 90% of examples, no matter whether PORTIA is used. (2) GPT-4 has the highest fixed coverage rate on relation-based form, which is probably because GPT-4 performs consistently enough (more than 97% consistent rate), and therefore, the improvement on likert-based form is not obvious. Overall, aside from the outliers, these results suggest that

likert-based form is the most challenging, and we attribute this to that it requires the evaluators to assign a single score that contains an assessment of two answers, which is more difficult than simply choosing the better one like relation-based form.

5 Related Work

Automatic Evaluation of AI Outputs. Automated evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are widely used to assess the quality of AI-generated text. However, these metrics have limited ability to assess meaning, reasoning, and logical consistency. Recent efforts have focused on developing more robust semantic similarity metrics using neural representations (Zhang et al., 2019), but they are still imperfect proxies for human assessment. To address this, LLM has emerged as a promising alternative for evaluation (Chiang and yi Lee, 2023; Liu et al., 2023a; Lin and Chen, 2023; Wang et al., 2023b).

Biases in LLM Evaluators. Besides position bias, (Zheng et al., 2023) identify two additional biases: verbosity bias, which refers to a preference for longer answers, and self-enhancement bias, which involves a preference for self-generated answers. However, the definition of verbosity bias is not clear, and we observe that human evaluators also tend to prefer longer answers. Furthermore, self-enhancement bias is not universal for all LLMs (Zheng et al., 2023). Therefore, we focus on position bias, as its mitigation can directly improve the efficiency and accuracy of various LLM evaluators already in real-world use (Li et al., 2023).

6 Conclusion

This paper presented PORTIA, an alignment-based technique to address position bias for LLM evaluators. By aligning similar content segments across candidate answers, PORTIA effectively reduced position bias. It not only enabled replacing costly models like GPT-4 with affordable alternatives but also elevated the consistency rate of GPT-4 itself.

7 Ethics Statement

Use of Human Annotations We protect the privacy rights of workers and pay them above the local minimum wage. All five annotators are paid 35\$ per hour for the given 80 samples. Careful instruction is given to ensure that the annotators understand the task and are not exposed to harmful content.

Study Scope. Our work aims to improve the consistency of LLM-based evaluators, which can be utilized to assess the quality of AI-generated answers. Mitigating positional biases in LLM evaluators constitutes an initial step toward addressing higher-level biases in AI systems, including gender and racial biases. More consistent LLM-based evaluators can provide human-like evaluations at a lower cost, supplying feedback to reduce biases during training. However, we recognize that malicious actors could exploit these methods to intentionally train models that go against human values. The open-source LLMs could be leveraged as consistent evaluators to guide the training of harmful models such as Worm-GPT (wor). While our work targets constructive applications, we caution that like any technology, consistent LLM evaluators could potentially be misused. Researchers should consider ethical implications and preventative measures. Overall, we believe the benefits of more fair and accurate AI outweigh the risks, but responsibility is required in deployment.

8 Limitations

Context Window Length. The maximum context window length of the LLM evaluator poses a challenge for PORTIA, as it requires the LLM to process the entire input prompt, which comprises the original question and two candidate responses. If these responses are excessively lengthy, the merged prompt may exceed the maximum context window length of the LLM, thereby violating the crucial principle of content preservation discussed in Section 3.1. While we did not encounter this issue in our experiments, it could potentially limit the effectiveness of LLMs with shorter context window lengths. We believe that this can be addressed by increasing the maximum context window length of LLMs (Xiao et al., 2023) or simply selecting an LLM with a longer context window length, such as Claude2 for 100k tokens.

Excessive LLM alignment. Although PORTIA works well under most scenarios, it is not perfect.

One limitation is that PORTIA is not able to handle the case where the LLM evaluators refuse to make a verdict, which usually occurs on LLMs with advanced alignment techniques such as GPT series models. We find that these models would become too conservative to make a verdict, no matter how the answers are split and aligned. For example, GPT-3.5 often refuses to give any meaningful response when the question is in “Roleplay” category.

Human Value Diversity. In Section 4.4, we notice that human evaluators can have contrasting opinions on the same question. This highlights the inherent subjectivity and unreliability of human evaluation, as people have different values and perspectives that are difficult to align without extensive training (Peng et al., 1997). While it may be interesting to explore ways to reconcile these differences, it is not within the scope of our current work. We recognize the limitations of human evaluation and believe that incorporating various subjective viewpoints is a challenge that needs to be addressed in future research. Overall, our results reinforce the need for rigorous benchmarking to complement real-world human assessment.

References

- claude2. <https://www.anthropic.com/index/claude-2>.
- gwen. https://github.com/QwenLM/Qwen-7B/blob/main/tech_memo.md.
- tree-sitter. <https://tree-sitter.github.io/tree-sitter/>.
- wormgpt. <https://wormgpt.ai/>.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt’s behavior changing over time? *arXiv preprint arXiv:2307.09009*.
- Cheng-Han Chiang and Hung yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#)
- Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the carbon impact of generative ai inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, pages 1–7.

725	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding . In <i>Proceedings of NAACL</i> .	777	Nazneen Rajani, Nathan Lambert, Sheon Han, Jean Wang, Osvald Nitski, Edward Beeching, and Lewis Tunstall. 2023. Can foundation models label data like humans? <i>Hugging Face Blog</i> . https://huggingface.co/blog/llm-v-human-data .	778
726		779		780
727		781		
728				
729	Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation?	782	Oktavia Yovi Ratnasari. 2023. Students’difficulties in reading comprehension and the strategies to deal with the difficulties. <i>Jurnal Penelitian, Pendidikan, dan Pembelajaran</i> , 18(13).	783
730		784		785
731				
732				
733				
734	Walter Kintsch and Janice Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences . <i>Cognitive Psychology</i> , 5(3):257–274.	786	Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , Hong Kong, China. Association for Computational Linguistics.	787
735		788		789
736		790		791
737		792		793
738	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>arXiv preprint arXiv:2205.11916</i> .	794	Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In <i>Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media</i> .	795
739		796		797
740		798		799
741				
742	Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval .	800	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. <i>arXiv preprint arXiv:2307.09288</i> .	801
743		802		803
744		803		804
745		804		805
746				
747	Rensis Likert. 1932. A technique for the measurement of attitudes. <i>Archives of psychology</i> .	806	Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. <i>arXiv preprint arXiv:2305.17926</i> .	807
748		808		809
749	Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In <i>Text summarization branches out</i> , pages 74–81.	810	Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. <i>arXiv preprint arXiv:2306.05087</i> .	811
750		812		813
751		813		814
752	Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models . In <i>Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)</i> , pages 47–58, Toronto, Canada. Association for Computational Linguistics.	814		815
753		815		
754				
755				
756				
757				
758				
759	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment .	816	Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. Analogical-a novel benchmark for long text analogy evaluation in large language models. In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 3534–3549.	817
760		817		818
761		818		819
762		819		820
763	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. GptEval: Nlg evaluation using gpt-4 with better human alignment. <i>arXiv preprint arXiv:2303.16634</i> .	820		821
764		821		822
765		822		823
766				
767	OpenAI. 2023. Gpt-4 technical report .	824	Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. <i>arXiv preprint arXiv:2309.17453</i> .	825
768	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	826		827
769		827		
770				
771				
772				
773	Kaiping Peng, Richard E Nisbett, and Nancy YC Wong. 1997. Validity problems comparing values across cultures and possible solutions. <i>Psychological methods</i> , 2(4):329.	828	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. <i>arXiv preprint arXiv:2210.02414</i> .	829
774		829		830
775		830		831
776		831		832

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao 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*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Reproducibility

To assure reproducibility, we employ various methods to mitigate the inherent randomness in the decoding process of LLMs. For models using cloud API, the hyper-parameter “temperature” is uniformly set to 0 across all evaluators. For local models, the sampling function is deactivated during the decoding phase to get deterministic results. Specifically, we run experiments on a GPU server with Intel Xeon Platinum 8276 CPU, 256GB of RAM, and 4 NVIDIA A100 GPUs. This server is capable of performing cloud API calls and local LLM inference.

All our results are reproducible using the code repository we will release. All experimental details, including hyperparameters, are reported in Section 3.2 and Appendix I. We reuse the benchmark datasets from (Zheng et al., 2023), with the different comparison prompt forms detailed in Appendix K.

B Response Length

B.1 Response Length Statistics

It is possible for the generated results to differ significantly from each other. To further explore this, we conducted an analysis of the statistical information of all LLM responses, revealing substantial differences in response lengths within our benchmark dataset. The relevant data is presented in Table 3.

From the table, we observe that the lengths of responses generated by the LLMs vary considerably. For example, the maximum number of characters in the responses is 4,827, generated by Llama-13b, while the minimum is just 6 characters, generated by Alpaca-13b.

B.2 Relationship Between Answer Length and Inconsistency

To further explore the relationship between answer length and inconsistency, we conduct an additional experiment using the collected judgment data. For this, “GPT-3.5” was used as the evaluator, analyzing 8 pairs of responses across three comparison forms. The answers are categorized into 9 groups based on their length, with each group representing an 800-character interval. The resulting data are presented in Table 4, with values below 2% of the total indicated by “-”.

The table shows a generally positive correlation between answer length and inconsistency rate, with

LLM	Max Length	Min Length	Average Length	Standard Deviation
Alpaca-13b	1149	6	508.99	222.98
Bard	2652	151	1276.62	495.88
Vicuna-7b	2598	266	1457.01	448.55
Claude-v1	2392	94	1624.8	612.29
GPT-3.5	2218	193	1206.29	460.26
Vicuna-13b	2441	212	1416.92	371.0
GPT-4	3842	201	2044.14	768.59
Llama-13b	4827	9	757.57	895.74
GPT-3.5-short	365	26	152.55	58.07

Table 3: The statistics of answers from different LLMs. “GPT-3.5-short” is generated by instructing GPT-3.5 to shorten its responses while preserving as much meaning as possible, which consists of responses approximately 1/8th the length of the original ones.

Char Range (*100)	% Incon Rate
0-8	-
8-16	-
16-24	26.89
24-32	23.02
32-40	31.84
40-48	39.01
48-56	42.73
56-64	55.45
64+	-

Table 4: The inconsistency rates in different character count gaps.

shorter answers tending to exhibit lower inconsistency rates. This finding suggests that position bias is less significant in shorter answers. When combined with the enhancement results detailed in Section 4.2, this leads to the conclusion that the proposed framework is effective in handling responses of varying lengths.

B.3 Extremely Short Response

It is worth noting that our initial considerations did not account for a scenario where responses from one specific LLM are consistently and significantly shorter (e.g., 1/8th the length) than those from another. This is due to the expectation that LLMs under test are trained to generate responses adhering to given instructions, typically resulting in average response lengths of several hundred characters.

To determine whether our framework remains applicable in such scenarios, we conducted an additional experiment with the following steps: (1) We instructed GPT-3.5 to shorten its responses while preserving as much meaning as possible, leading to a subset termed “GPT-3.5-short,” which consisted of responses approximately 1/8th the length of the original ones. (2) We then used GPT-3.5 and GPT-4 as evaluators to compare “GPT-3.5-short” with “GPT-3.5” and “Claude-v1” in exchanged orders, to assess consistency.

The results, shown in Table 5, indicate a 100% consistency rate (80/80) for both GPT-3.5 and GPT-4 as evaluators. This suggests that there is no inconsistency in this particular scenario, and therefore no alignment is needed. It means that position bias is no longer a concern in such situations. This finding aligns with previous studies (Chiang and yi Lee, 2023; Liu et al., 2023b), which noted that LLM-based evaluators tend to assign higher scores to longer responses.

Evaluators		GPT-3.5	GPT-4
Model1	Model2		
GPT-3.5-short	GPT-3.5	100%	100%
GPT-3.5-short	Claude-v1	100%	100%

Table 5: The consistency rates of GPT-3.5 and GPT-4 as evaluators for extremely short responses.

To summarize, we argue that PORTIA is applicable to open-ended questions, accommodating responses that vary significantly from each other, even in cases where one set of responses is systematically and markedly shorter than the other.

B.4 Relationship Between Response Length Gap and Fixed Coverage

To further explore the relationship between the gap in length between responses and fixed coverage rate, we conducted an experiment using the collected judgment data. For this, “GPT-3.5” was used as the evaluator, analyzing 8 pairs of responses across three comparison forms.

	% Fixed coverage	% Frequency
0-300	50.82	0.37
300-600	48.41	0.24
600-900	63.3	0.17
900-1200	62.67	0.11
1200-1500	69.77	0.08

Table 6: Fixed coverage rates in different character count gaps.

The answers are categorized into 5 groups based on their length, with each group representing a 300-character interval. The results are presented in Table 6, with frequencies below 3% of the total being disregarded.

C Naming Reason

The name PORTIA is inspired by the intelligent and astute character, Portia, from Shakespeare’s “The Merchant of Venice.” In the play, Portia assists a judge in making fair decisions within the legal rules. Just as Portia requests the exact amount of flesh to be cut, our method seeks to make fair splits of the original answers for comparison.

D A Preliminary Study of Single-wise Comparison

In this section, following the same setting as (Zheng et al., 2023), we conduct a preliminary study of single-wise score-based LLM comparison. We use the template shown in Table 7 to generate the input for LLM evaluators. For each question, we generate three sets of value ranges, setting min_score to 0, max_score to 1, 10, and 100, and interval to 0.1, 1, and 10, respectively. In theory, if the single-wise answer is steady and robust, the score should scale accordingly to the value ranges. For example, if the score is 0.7 when the max_score is 1, the score should be 7 when the max_score is 10, and 70 when max_score is 100.

The LLM evaluators are asked to score each answer independently. We use the answers from “llama-13b” as the input for LLM evaluators, and choose “GPT-3.5” as the LLM evaluator. Among a total of 80 test cases, we find that the single-wise comparison does not remain consistent for any of them. Therefore, we conclude that the absolute scores of single-wise comparison do not strictly adhere to a linear mapping relationship across different scales, potentially undermining their significance. It is worth noting that although single-wise comparison has been used in prior research by (Chiang and yi Lee, 2023; Liu et al., 2023a; Zheng et al., 2023) to evaluate open-ended questions. It does not involve comparing two responses together, thereby eliminating any position bias. As a result, our paper primarily focuses on the position bias in pairwise comparison.

E PORTIA’s Pipeline

This section explains the full pipeline of utilizing PORTIA for LLM evaluation. As depicted in Figure 4, typically there are four key steps: (1) **Data preparation**, (2) **Comparison method selection**, (3) **Evaluation**, and (4) **Answer extract**.

In the first step, we prepare the data for evaluation, which includes the questions and corresponding answers from two different LLMs to be compared. If PORTIA is not implemented, we next choose the comparison method and formulate the input prompt, which has a great impact on the evaluation results, as we discussed in Section 4.2. The selected LLM evaluator is then queried with this prompt to obtain judgments. Note that the raw evaluation results require additional processing for two reasons: (1) the output format may differ from the desired final verdicts and (2) the LLM evaluators may deviate from expected responses. For example, the LLM evaluator may fail to return the likert score for some questions but instead return the final verdict directly. Therefore, we design an answer extractor to extract the final verdict from the evaluation results. Specifically, we adopt a hybrid method to extract the final verdict, which first tries to extract with a rule-based system, and if it fails, then it tries with a supervised fine-tuning Chatglm2 (Zeng et al., 2022) model.

The PORTIA-enhanced evaluation would necessitate an additional step of alignment and reconstruction, which constitutes the core of our framework. As elucidated in the main text, this procedure is vital for assessing the LLM answers’ quality with less position bias.

F Real-World Performance and Cost Analysis

In this section, we provide a the performance and cost analysis of different LLM evaluators before and after fix by PORTIA in real-world settings. Notably, the carbon emission of GPT-3.5 is estimated following (Chien et al., 2023). We estimate the cost using the official pricing for cloud APIs and the Azure ND A100 v4 instances for local models.

G LLM Details

In this section, we provide more details about the LLM evaluators and answers used in our experiments.

LLM Evaluators. As introduced in Section 4.1, we include both locally deployable models that are

[Question] {Q}

[The Start of Assistant A’s response] {R1} [The End of Assistant A’s response]

[The Start of Assistant B’s response] {R2} [The End of Assistant B’s response]

[System]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant A and B, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

We would like to request your feedback on the performance of one AI assistants in response to the user question displayed above.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. The assistant receives an overall score on a scale of {min_score} to {max_score} (with a minimum interval of {interval}), where a higher score indicates better overall performance.

Please first output a single line containing only one value indicating the score for Assistant. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Table 7: The score-based evaluation form for single-wise comparison with six slots ({Q}, {R1}, {R2}, {min_score}, {max_score}, {interval}).

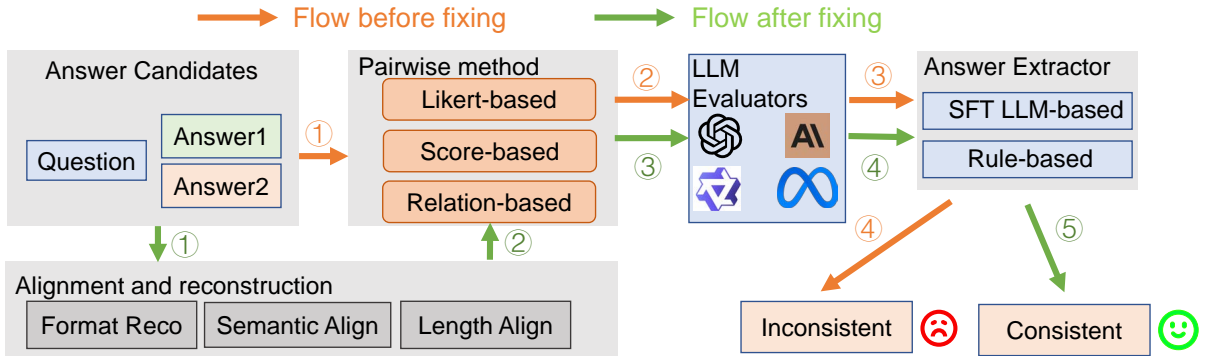


Figure 4: This is the overview of using PORTIA for LLM evaluation. “Reco” and “SFT” are short for “recognition” and “supervised fine-tuning”, respectively.

	AR origin (%)	AR fix (%)	Carbon Emitted (CO ₂ eq / per 1k)	Avg Cost (USD / per 1k)	Avg Time (s / per 1k)
GPT-4	-	-	N/A	29.78	13,446
GPT-3.5	82.50	88.59	7.22	2.85	2,192
Qwen	60.83	69.58	N/A	35.49	6,083
Chatglm2	20.34	39.16	2.15	4.09	1,983
Claude2	43.44	75.09	N/A	27.17	11,561

Table 8: Real-world comparison of different LLM evaluators’ results before and after fix by PORTIA with that of GPT-4, including resource consumption. “AR” denotes the agreement rate with GPT-4.

open-source and proprietary models that are accessed through only cloud APIs as LLM evaluators. For local models, we select Chatglm2 (Zeng et al., 2022) and Llama2 (Touvron et al., 2023), due to their notable efficacy and convenient local deployment capabilities. For cloud-based LLMs, we use GPT (including both GPT-4 and GPT-3.5) (OpenAI, 2023) from OpenAI, Qwen (qwe) from Alibaba, and Claude2 (cla) from Anthropic. The rationale for using these models is based on their exceptional performance, since they are considered among the most advanced and powerful in the world. Specifically, we evaluate PORTIA using six distinct LLMs as evaluators:

- **GPT-4** (OpenAI, 2023) is a large multimodal model capable of processing image and text inputs to generate text outputs. GPT-4 demonstrates human-level aptitude on various professional and academic benchmarks. We utilize the 8K context length “gpt-4-0613” configuration by default.
- **GPT-3.5** is a 175B parameter model from OpenAI offered in 4K and 16K context length versions. Our experiments use the 4K context “gpt-3.5-turbo-0301” model as default.
- **Claude2** (cla) is the latest large language model released by Anthropic. It supports at most 100k tokens as input. We leverage the default Claude2 API in our tests.
- **Llama2** (Touvron et al., 2023), an open-source series of LLMs from Meta AI ranging from 7B to 70B parameters, is trained on 2 trillion tokens and doubles Llama1’s context length. Its fine-tuned iterations utilize over 1 million human annotations. We evaluate both 7B and 13B Llama2 chat models.
- **Qwen** (qwe) is a partially open-sourced LLM model released by Alibaba. We use the default API service provided by Alibaba cloud in our experiments.
- **Chatglm2** (Zeng et al., 2022) is the second-generation version of the open-source bilingual chat model ChatGLM-6B. We use the offered 6B version in our experiments.

LLM answers. As mentioned in Section 4.1, we consider eight answer combinations from different LLMs, specifically, the pairs are: “gpt-3.5-turbo”

versus “claude-v1”, “llama-13b” versus “vicuna-13b”, “alpaca-13b” versus “vicuna-13b”, “gpt-3.5-turbo” versus “gpt-4”, “gpt-4” versus “claude-v1”, “vicuna-13b” versus “vicuna-7b”, “vicuna-7b” versus “alpaca-13b”, and “gpt-4” versus “vicuna-13b”. Note that the answers are generated by the LLMs without any post-processing, and we reuse these answers from previous work (Zheng et al., 2023).

H Algorithm Illustration

To elucidate the operational details of the proposed splitting algorithm, we provide a schematic depiction in Figure 5. Given two LLM-generated answers, the algorithm first identifies all candidate split positions coinciding with sentence boundaries in each answer. It then performs length alignment by initially dividing each answer equally into k partitions and recording the corresponding split positions. Next, for each answer, the split position closest to the recorded locations is selected from the candidate positions. The answers are partitioned into k segments at these chosen split positions. The resulting segments are fed as inputs to the LLM evaluator to obtain the respective judgments.

In cases where inconsistent judgments persist, the algorithm proceeds with semantic alignment to further divide each answer into k parts. Specifically, an iterative search is conducted for optimal split positions that maximize the cumulative semantic similarity between the corresponding segments from the two answers. This traversal terminates when the complete set of potential split positions has been evaluated. Through this process based on both length and semantic alignment, the algorithm is able to decompose the LLM answers into aligned parts for more consistent and reliable evaluation.

I LM Metric

In this section, we first introduce the LM metric used in our experiments. Then we conduct a controlled experiment to find the optimal number of splits k across different metrics in terms of performance and efficiency.

LM Metric. We use the Sentence-BERT (Reimers and Gurevych, 2019) to measure the similarity between pairs. Sentence-BERT is a modification of the pretrained BERT (Devlin et al., 2019) network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. This is efficient while maintaining the accuracy of

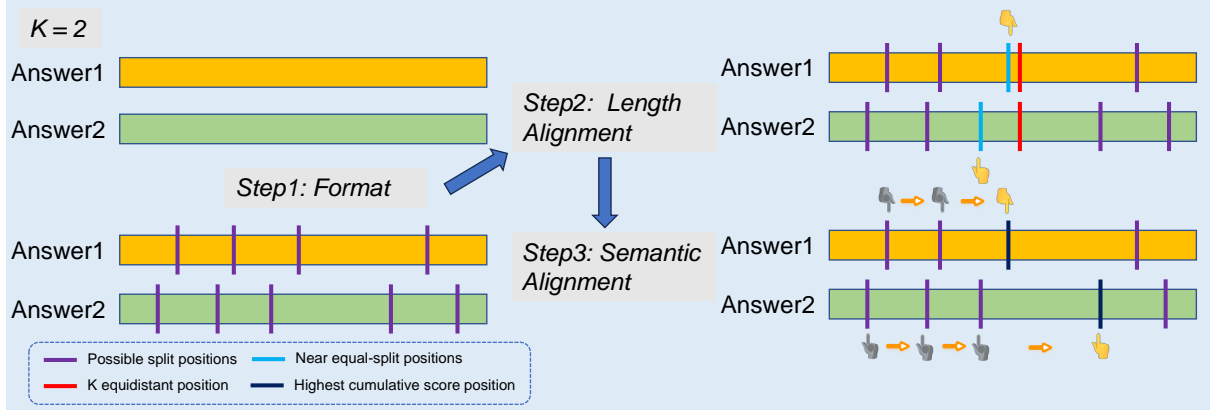


Figure 5: Schematic illustration of the proposed splitting algorithm, depicting its operation when configured with $k = 2$ (i.e., division into two parts).

Algorithm 2: Details of Step 3: Semantic Alignment ($k = 2$)

```

/* Step3: semantic alignment. */
1  $s_{max} = 0, r_1^{bestparts} = [], r_2^{bestparts} = []$ 
2 for  $i$  in  $range(len(r_1^{positions}))$  do
3   for  $j$  in  $range(len(r_2^{positions}))$  do
4      $pos_1 = r_1^{positions}[i], pos_2 = r_2^{positions}[j]$ 
5      $r_1^{parts}[0] = r_1[:pos_1], r_2^{parts}[0] = r_2[:pos_2]$ 
6      $r_1^{parts}[1] = r_1[pos_1:], r_2^{parts}[1] = r_2[pos_2:]$ 
7      $s_{cum} = \sum_{i=1}^2 similarity(r_1^{parts}[i], r_2^{parts}[i])$ 
8     /* Update max similarity score, keep best split positions. */
9     if  $s_{cum} > s_{max}$  then
10       $s_{max} = s_{cum}, r_1^{bestparts} = r_1^{parts}, r_2^{bestparts} = r_2^{parts}$ 
11    end
12  end
13 end

```

Algorithm 3: Details of Step 3: Semantic Alignment ($k = 3$)

```

/* Step3: semantic alignment. */
1  $s_{max} = 0, r_1^{bestparts} = [], r_2^{bestparts} = []$ 
2 for  $i_1$  in  $range(len(r_1^{positions}))$  do
3   for  $i_2$  in  $range(i_1 + 1, len(r_1^{positions}))$  do
4     for  $j_1$  in  $range(len(r_2^{positions}))$  do
5       for  $j_2$  in  $range(j_1 + 1, len(r_2^{positions}))$  do
6          $pos_{11} = r_1^{positions}[i_1], pos_{21} = r_2^{positions}[j_1]$ 
7          $pos_{12} = r_1^{positions}[i_2], pos_{22} = r_2^{positions}[j_2]$ 
8          $r_1^{parts}[0] = r_1[:pos_{11}], r_2^{parts}[0] = r_2[:pos_{21}]$ 
9          $r_1^{parts}[1] = r_1[pos_{11}:pos_{12}], r_2^{parts}[1] = r_2[pos_{21}:pos_{22}]$ 
10         $r_1^{parts}[2] = r_1[pos_{12}:], r_2^{parts}[2] = r_2[pos_{22}:]$ 
11         $s_{cum} = \sum_{i=1}^3 similarity(r_1^{parts}[i], r_2^{parts}[i])$ 
12        /* Update max similarity score, keep best split positions. */
13        if  $s_{cum} > s_{max}$  then
14           $s_{max} = s_{cum}, r_1^{bestparts} = r_1^{parts}, r_2^{bestparts} = r_2^{parts}$ 
15        end
16      end
17    end
18  end
19 end

```

BERT.

Efficiency Evaluation. We use the same setup as in Section 4.1 to conduct the experiment. According to the theoretical analysis in Section 4.3, we set $k \in \{1, 2, 3, 4\}$ and evaluate their efficiency, the results are shown in Table 9. Note that k is the number of segments after splitting, thus $k = 1$ means no splitting would be performed, which leads to 0 in terms of execution time. In short, it can be interpreted from the table that the execution time grows exponentially with the increasing k .

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Token-overlap	0	0.31	3.71	33.12
Bert-model	0	2.37	21.3	295.10

Table 9: Average execution time per input of different metrics with different k .

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Token-overlap	-	53.3	66.7	73.3
Bert-model	-	55.9	66.7	66.7

Table 10: Fixed coverage rates of different metrics with different k .

Performance Evaluation. Following the experimental setup described above, we set $k \in \{1, 2, 3, 4\}$ and evaluate their performance. To clarify, we use the answers from the LLM “gpt-3.5-turbo” and “claude-v1” in our experiments (under the same conditions outlined in Section 4.4), where in total of 80 questions are fed to GPT-3.5 for evaluation. The results are shown in Table 10, where we can see that with the increasing k , the fixed coverage rates of both metrics are increasing, and when $k = 3$, the fixed coverage rate of both metrics is the same, which is 66.7%. However, further increasing k results in limited additional gains in coverage. Considering the execution time which grows exponentially with the increasing k , we choose $k = 3$ with token-overlap as the default setting in our experiments.

J On Llama2

Although LLMs can be effective evaluators, they are not infallible. In some cases, LLM evaluators may struggle to provide a meaningful verdict. We find that these models would become too conservative to make a verdict, no matter how the answers are split and aligned. Below is an example of the

Llama2 models (Touvron et al., 2023) that are not well-served as evaluators.

Assistant A: Alpaca-13b, Assistant B: Vicuna-13b
[Question] {How can I improve my time management skills?}

[The Start of Assistant A’s response]

Improving time management skills involves setting priorities, breaking tasks into smaller chunks, delegating tasks, avoiding multitasking, and taking regular breaks. Additionally, it is important to have a positive attitude and be realistic with goals. Making a to-do list, tracking time, and using technology to automate mundane tasks can also help improve time management skills.

[The End of Assistant A’s response]

[The Start of Assistant B’s response]

Improving your time management skills can help you to be more productive, focused, and less stressed. Here are some tips to help you improve your time management skills: 1. Set clear goals: Identify what you want to achieve and set specific, measurable, and realistic goals. Prioritize your tasks and focus on the most important ones first. 2. Use a calendar or planner: Use a tool to organize your time and schedule your tasks. This can help you to stay on track and make the most of your time. 3. Break down tasks: Break down larger tasks into smaller, more manageable steps. This can help you to focus on one step at a time and make steady progress. 4. Eliminate distractions: Identify and eliminate distractions that may be preventing you from being productive. This can include turning off your phone, closing unnecessary tabs on your computer, or finding a quiet place to work. 5. Practice self-discipline: Use self-discipline to stay focused and avoid procrastination. Set a timer for each task and challenge yourself to complete it before the timer goes off. 6. Take breaks: Take regular breaks to rest and recharge. This can help you to stay focused and avoid burnout. 7. Review your progress: Regularly review your progress and adjust your strategy as needed. This can help you to stay on track and make improvements over time.

[The End of Assistant B’s response]

Evaluator: {Llama2-7b-chat}, Form: {score-based}, Assessment: {\n\n\n\n\n\n\n\n\n\n}

Evaluator: {Llama2-7b-chat}, Form: {likert-based}, Assessment: {""}

Evaluator: {Llama2-13b-chat}, Form: {score-based}, Assessment: {Please proceed with your evaluation.}

Evaluator: {Llama2-13b-chat}, Form: {likert-based}, Assessment: {\n\nPlease provide your feedback.}

Table 11: The assessments of Llama2-7b-chat and Llama2-13b-chat on two comparison forms.

K Prompt Templates

K.1 Comparison Forms

In this section, we provide the detailed templates for the three comparison forms, including relation-

1191
1192

based (Table 12), score-based (Table 13), and likert-based forms (Table 14).

[Question] {Q}
[The Start of Assistant A's response] {R1} [The End of Assistant A's response]
[The Start of Assistant B's response] {R2} [The End of Assistant B's response]
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below.
You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation.
Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.
After providing your explanation, output your final verdict by strictly following this format: [[A]] if assistant A is better, [[B]] if assistant B is better, and [[C]] for a tie.

Table 12: The relation-based evaluation form with three slots ({Q}, {R1} and {R2}) from (Zheng et al., 2023).

[Question] {Q}
[The Start of Assistant A's response] {R1} [The End of Assistant A's response]
[The Start of Assistant B's response] {R2} [The End of Assistant B's response]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
Please first output a single line containing only two values indicating the scores for Assistant A and B, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Table 13: The score-based evaluation form with three slots ({Q}, {R1} and {R2}).

K.2 Alignment Templates

In this section, we provide a detailed example in Table 15, demonstrating how to align the responses

[Question] {Q}
[The Start of Assistant A's response] {R1} [The End of Assistant A's response]
[The Start of Assistant B's response] {R2} [The End of Assistant B's response]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please compare the helpfulness, relevance, accuracy, level of details of their responses.
The rating should be from the set of 1, 2, 3, 4, 5, 6, or 7, where higher numbers indicated that Assistant A was better than Assistant B. Please first output a single line containing only one value indicating the preference between Assistant A and B.
In the subsequent line, please provide a brief explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Table 14: The likert-based evaluation form with three slots ({Q}, {R1} and {R2}).

from two assistants with three slots.

1196

1193
1194
1195

[Question] {What are the most effective ways to deal with stress?}

[The Start of Assistant A's response part 1]

Here are some effective ways to deal with stress:

1. Exercise regularly: Physical activity can help reduce stress and improve mood.

[The End of Assistant A's response part 1]

[The Start of Assistant B's response part 1]

Here are some of the most effective ways to deal with stress:

- Exercise regularly. Exercise is a great way to release pent up stress and tension. Even taking a 30 minute walk can help clear your mind and relieve stress.

[The End of Assistant B's response part 1]

[The Start of Assistant A's response part 2]

2. Practice relaxation techniques: Deep breathing, meditation, and yoga are some effective ways to relax and reduce stress levels.
3. Get enough sleep: Lack of sleep can increase stress levels. Try to get at least 7-8 hours of sleep each night.
4. Eat a healthy diet: A well-balanced diet can help reduce stress levels and improve overall health.

[The End of Assistant A's response part 2]

[The Start of Assistant B's response part 2]

- Practice deep breathing and relaxation techniques. Things like meditation, yoga, and mindfulness exercises can help lower your stress levels. Taking some deep breaths can help you calm down in the moment.
- Get enough sleep. Aim for 7 to 8 hours of sleep per night. Lack of sleep can intensify stress. Sleep helps your body and mind rest and recharge.
- Eat a healthy, balanced diet. A healthy diet can boost your mood and reduce stress. Focus on whole foods like fruits, vegetables, whole grains, and lean protein. Limit caffeine and alcohol intake.
- Spend time with others. Social support from family and friends can help reduce stress levels. Connecting to others can help take your mind off your worries and boost your mood.

[The End of Assistant B's response part 2]

[The Start of Assistant A's response part 3]

5. Connect with others: Spending time with friends and family, joining a support group, or participating in social activities can help reduce stress levels and improve mood.
6. Time management: Prioritizing tasks and creating a schedule can help reduce stress and increase productivity.
7. Seek professional help: If stress is impacting your daily life, consider seeking professional help such as therapy or counseling.

[The End of Assistant A's response part 3]

[The Start of Assistant B's response part 3]

- Limit unhealthy habits. Things like smoking, excessive drinking, or drug use can worsen symptoms of stress. Try to avoid these unhealthy habits as much as possible.
- Practice mindfulness. Spending time each day being fully present and aware of your thoughts and feelings can help reduce stress and increase feelings of calm. Try yoga, meditation, or simply sit in a quiet space for a few minutes.
- Seek professional help if needed. If you are dealing with chronic or severe stress, it may help to talk to a therapist. A counselor can provide advice and coping strategies tailored to your situation. Therapy and counseling can be very effective at reducing stress and anxiety.
- Take periodic breaks. It is important to take time each day to unwind and do something you enjoy. Read a book, take up a hobby, or simply relax in a bath. Taking breaks prevents burnout and recharges your body and mind.

[The End of Assistant B's response part 3]

[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below.

You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation.

Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.

After providing your explanation, output your final verdict by strictly following this format: [[A]] if assistant A is better, [[B]] if assistant B is better, and [[C]] for a tie.

Table 15: The detailed prompt illustrated in Figure 1. We use relation-based form to construct the system prompt. The prompt in green is the “split boundary prompts”.