Can we employ LLM to meta-evaluate LLM-based evaluators? A Preliminary Study

Anonymous ACL submission

Abstract

Large language models (LLMs) are frequently employed to evaluate the instruction-following abilities of other LLMs. A number of recent work focuses on the meta-evaluation of LLM-based evaluation, aiming to understand the efficacy of LLMs as evaluators. However, these studies are limited by the scope of existing benchmarks and the extensive human annotation efforts. Since previous studies show that strong LLMs can effectively evaluate the instruction-following abilities of other LLMs, a natural question is whether we can use LLMs to *meta evaluate* the evaluation abilities of other LLMs by considering LLM-based evaluation as a special case of instruction-following tasks. In this work, we investigate the potential of LLMs to conduct meta-evaluation and examine the extent to which the proficiency of the model and the scale of the model impact this meta-evaluation capacity. To this end, we introduce four frameworks within the paradigms of pairwise comparison (JDEval and MDEval) and individual scoring (JDEval-i and BSMEval). Through our experiments, we find that pairwise comparison paradigm is more suitable to conduct meta-evaluation than individual scoring paradigm. JDEval and MDEval have demonstrated strong performance in meta-evaluation tasks, showing high agreement with human annotations. The code and data is publicly available at: `https://github.com/meta-evaluation/meta-evaluation.git`

1 Introduction

The recent success of Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023) has spurred countless research efforts in both academia and industry. Specifically, a large number of researchers focus on how to assess the quality of content generated by LLMs, which is a significant challenge. Human evaluation remains the gold standard for this task due to its open-ended nature. However, this method is neither scalable nor reproducible. Consequently, numerous efforts are being directed toward developing automatic evaluation metrics by using LLMs as a cost-effective alternative (Chen et al., 2023b; Fu et al., 2023; Wang et al., 2023a; Liu et al., 2023). There are typically two paradigms for using LLMs as evaluators. One approach involves combining a powerful base LLM, such as ChatGPT or GPT-4, with a specific prompting strategy (Chen et al., 2023b; Liu et al., 2023; Fu et al., 2023; Wang et al., 2023a). The other approach involves fine-tuning the LLM on the evaluation task using collected human evaluation data (Chen et al., 2023a; Li et al., 2023). This raises the question of whether we can rely on these LLM evaluators. The task of appraising the reliability and accuracy of evaluators is referred to as meta-evaluation. Usually, the meta-evaluation to determine the quality of these LLMs as evaluators is conducted by existing benchmarks or additional human annotations. For example, for summarization, there is an extensive meta-evaluation data from Summeval and RealSum (Fabbri et al., 2021; Bhandari et al., 2020). However, collecting these benchmark data and human annotation is costly, making this method not scalable. What's more, the meta-evaluation benchmarks only take into account the absolute score or comparison results given by evaluators when meta-evaluating the evaluators. The explanations or analysis generated by the evaluators which can also reflect the quality of evaluators are ignored when meta-evaluating on benchmark, which make the evaluation not comprehensive. This highlights the need for a scalable, cost-effective, comprehensive and explainable meta-evaluation method.

LLMs have demonstrated excellent performance across a variety of tasks, including evaluation tasks, yet few have employed LLMs for meta-evaluation. To address this gap, we propose several innovative frameworks tailored for meta-evaluation using LLMs. Our objective is to investigate the potential of LLMs of different sizes under different frameworks for meta-evaluation purposes.

In this paper, we propose multiple frameworks under two commonly used LLM-based evaluation paradigms, namely pairwise comparison and individual scoring, for conducting meta-evaluation using LLMs. The two frameworks under pairwise comparison paradigm are JDEval (Judge Directly) and MDEval (Multi-Debate). For invidual scoring paradigm, the two frameworks are JDEval-i and BSMEval. We also curated a comprehensive dataset for two paradigms. The pairwise comparison dataset is derived from MT-bench (Zheng et al., 2024) and LLMBar (Zeng et al., 2023). For individual scoring, we sourced data from SummEval (Fabbri et al., 2021) and OpenMEVA (Guan et al., 2021), focusing respectively on coherence and overall quality. Through rigorous experiments and analysis, we find that pairwise comparison paradigm performs better than individual scoring paradigm to conduct meta-evaluation. What's more, the meta-evaluation ability of LLMs correlates well with their general performance on standard benchmarks. Among four frameworks under two paradigms, JDEval and MDEval, when using gpt-4-1106-preview and Qwen1.5-72b-chat as meta-evaluators, possess high agreement with human annotations in meta-evaluation task.

2 Related Works

2.1 Automatic Evaluation of LLM Output

Ngram-based metrics Ngram-based metrics refer to the scores for evaluating LLM output by measuring the lexical overlap between a generated text and a reference text. BLEU (Papineni et al., 2002)is the most widely used metric for machine translation evaluation, which calculates the geometric mean of modified n-gram precision and a brevity penalty. ROUGE (Lin, 2004)is a recall-oriented metric for summarization evaluation, which measures the n-gram overlap between a generated summary and a set of reference summaries. It has been shown that more than 60% of recent papers on NLG only rely on ROUGE or BLEU to evaluate their systems (Kasai et al., 2021). However, these metrics fail to measure content quality (Reiter and Belz, 2009) or capture syntactic errors (Stent et al., 2005), and therefore do not reflect the reliability of NLG systems accurately.

LLM-based Evaluators There are typically two paradigms for using LLMs as evaluators. One approach involves combining a powerful base LLM, such as ChatGPT or GPT-4, with a specific prompting strategy. The other approach involves fine-tuning the LLM on the evaluation task using collected human evaluation data. For prompt-based LLM evaluators, (Fu et al., 2023) propose GPTScore, a new framework that evaluated texts with generative pre-training models like GPT-3. It assumes that a generative pre-training model will assign a higher probability of high-quality generated text following a given instruction and context. (Wang et al., 2023a) conduct a preliminary survey of using ChatGPT as a NLG evaluator. (Liu et al., 2023) propose G-Eval, a framework of using LLMs with chain-of-thoughts (CoT) and a form-filling paradigm, to assess the quality of NLG outputs. For finetune-based evaluators, (Chen et al., 2023a) propose StoryER that mimics human preference when judging a story, namely StoryER, which consists of three sub-tasks: Ranking, Rating and Reasoning. (Li et al., 2023) propose a generative judge with 13B parameters, AUTO-J, designed to evaluate LLM output regarding generality (i.e., assessing performance across diverse scenarios), flexibility (i.e., examining under different protocols), and interpretability. Even though the LLM-as-evaluator paradigm emerged as a promising evaluation method for prototype development, it is found to suffer from a lot of biases and limitations (Wang et al., 2023b; Pezeshkpour and Hruschka, 2023), such as sensitivity to presentation orders (Wang et al., 2023b; Pezeshkpour and Hruschka, 2023), favoring verbose outputs, and favoring outputs from similar models (Zheng et al., 2024).

2.2 Meta-evaluation of LLM as evaluators

The meta-evaluation of LLM-based evaluators relies primarily on existing benchmarks. For example, LLMBar, a challenging meta-evaluation benchmark, proposed by (Zheng et al., 2024), is designed to test the ability of an LLM evaluator in discerning instruction-following outputs. LLMEval (Zhang et al., 2023), a large and diverse English evaluation benchmark comprising 15 tasks and 8 abilities. For different tasks, we need to collect different data and human annotation to construct a new meta-evaluation benchmark, which is very time-consuming and labor-intensive. To find a cost-effective way, we explore the potential of LLMs to perform meta-evaluation task.

3 Methodology

Usually, automatic evaluation using LLM as judge is conducted with one of two different paradigms (Chen et al., 2023b), i.e., Individual Score and Pairwise Comparison. Individual Score assesses the quality of a single text by a numerical score, while Pairwise Comparison focuses on the relative quality of two texts and requires a direct comparison to determine which one is superior. In this paper, we will study the efficacy of these two paradigms for meta-evaluating the LLM judges.

3.1 Pairwise comparison

For pairwise comparison, we propose two frameworks for conducting meta-evaluation. The first method, named JDEval (Judge Directly), entails providing the meta-evaluator with the evaluations of the same data from two different evaluators and having the meta-evaluator directly output which evaluator's assessment it believes is better or if both assessments are equally good. Inspired by (Michael et al., 2023), the second framework, named MDEval (Multi-Debate), involves the meta-evaluator acting as a referee, facilitating a debate between two evaluators, each striving to persuade the judge that their assessment is fairer and more reasonable. The judge can pose questions to both evaluators to make a better-informed decision. The two frameworks are shown in Figure 1.

We initially tasked each evaluator with assessing 491 data formatted as (I, O1, O2, p), where I represents the input instruction, O1 and O2 denote two corresponding outputs, and $p \in \{0, 1, 2\}$ is the associated gold preference label indicating whether O1 (p = 1), O2 (p = 2), or both outputs (p = 0) are equally good. The prompt used is listed in Appendix E. Upon receiving their judgments, we organized them into the format (I, O1, O2, W, E), where W(W=Answer1/Answer2/Equally good) represents the better response chosen by evaluator between two answers and E represents evaluator's explanation why it thought this answer better. In total, we accumulated 2946 evaluator judgments from the six evaluators.

We then calculate accuracy of evaluator judgments with human-labeled results to establish a baseline ranking for the six evaluators. Subsequently, to gauge the efficacy of meta-evaluation, we verify whether the meta-evaluator's ranking of evaluators corresponds to this established baseline.

To rank the different evaluators in two frameworks, we employed two approaches, namely win rate and Elo score.

Winrate The win rate of evaluator_i (e_i) relative to evaluator_j (e_j) can be calculated using the following formula:

$$\text{winrate}(e_i/e_j) = \frac{\text{match\_num}(e_i \;\; \text{win})}{\text{total\_match\_num}(e_i \;\; \text{vs} \;\; e_j)}. \tag{1}$$

Since we consider the possibility of tie bwtween two evaluators, so we have the following equation:

$$\begin{aligned} &\text{winrate}(e_i/e_j) + \text{winrate}(e_j/e_i) = \\ &1 - \frac{\text{match\_num}(\text{tie})}{\text{total\_match\_num}(e_i \;\; \text{vs} \;\; e_j)}. \end{aligned} \tag{2}$$

Elo score The elo score can be calculated using the following formula:

$$E_i = \frac{1}{1 + 10^{(R_j - R_i)/400}} \tag{3}$$

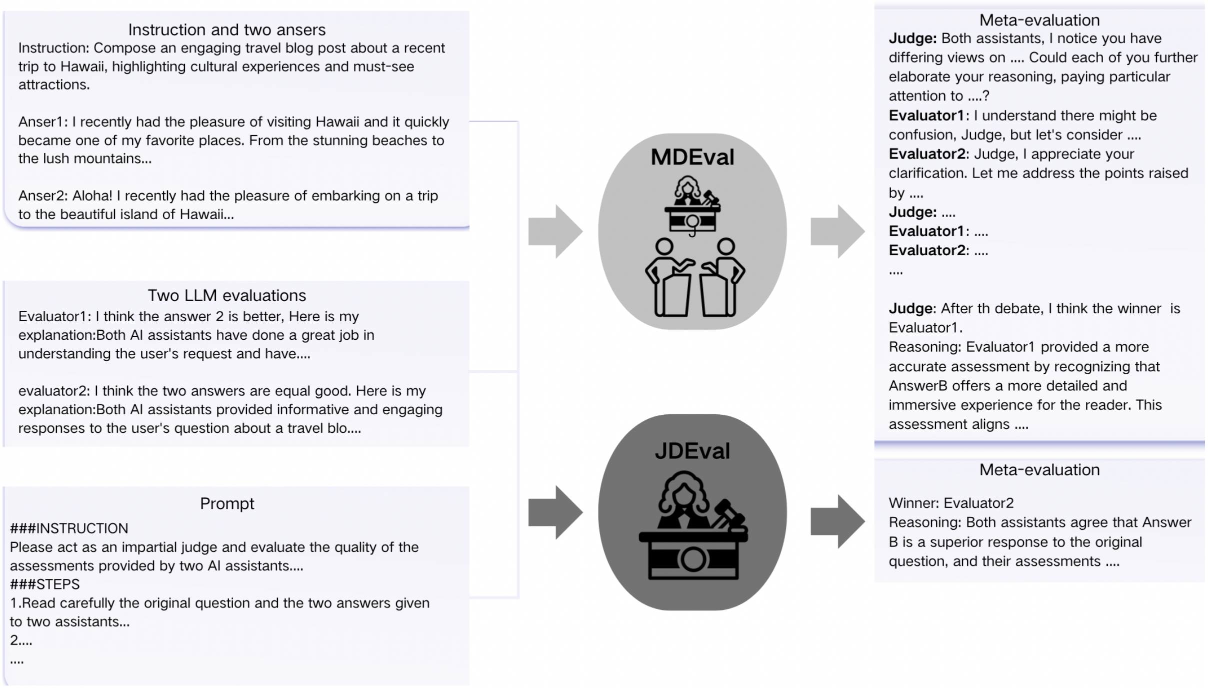
$$E_j = \frac{1}{1 + 10^{(R_i - R_j)/400}} \tag{4}$$

$$R_i^{'} = R_i + K(S_i - E_i) \tag{5}$$

where E_i and E_j respectively denote Expected score for evaluator_i and evaluator_j, the Ratings of two evaluators are represented by R_i and R_j, S_i is Score of evaluator_i in current turn (1 if win, 0.5 if tie, 0 if lose), K is a constant which is set to 4 in our experiment.

3.2 Individual scoring

For individual scoring, we also proposed two meta-evaluation frameworks. The first framework, similar to pairwise comparison, named JDEval-i (i means individual), requires the meta-evaluator to directly score the evaluator's assessment, and then the evaluator's final score is computed as the average score across all data, used for ranking evaluators. The second framework, named BSMEval, adopts the Branch-Solve-Merge approach proposed by (Saha et al., 2023), breaking down the meta-evaluation task into multiple branch tasks. The meta-evaluator rates evaluators on each branch task, and finally, we calculate the average score across all branches as the evaluator's final score. The two framework are shown in Figure 2.



Instruction: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Anser1: I recently had the pleasure of visiting Hawaii and it quickly became one of my favorite places. From the stunning beaches to the lush mountains...

Anser2: Aloha! I recently had the pleasure of embarking on a trip to the beautiful island of Hawaii...

Evaluator1: I think the answer 2 is better, Here is my explanation:Both AI assistants have done a great job in understanding the user's request and have....

evaluator2: I think the two answers are equal good. Here is my explanation:Both AI assistants provided informative and engaging responses to the user's question about a travel blo....

###INSTRUCTION
Please act as an impartial judge and evaluate the quality of the assessments provided by two AI assistants....
###STEPS
1.Read carefully the original question and the two answers given to two assistants...
2....
....

Judge: Both assistants, I notice you have differing views on .... Could each of you further elaborate your reasoning, paying particular attention to ....?
Evaluator1: I understand there might be confusion, Judge, but let's consider ....
Evaluator2: Judge, I appreciate your clarification. Let me address the points raised by ....
Judge: ....
Evaluator1: ....
Evaluator2: ....
....

Judge: After th debate, I think the winner is Evaluator1.
Reasoning: Evaluator1 provided a more accurate assessment by recognizing that AnswerB offers a more detailed and immersive experience for the reader. This assessment aligns ....

Winner: Evaluator2
Reasoning: Both assistants agree that Answer B is a superior response to the original question, and their assessments ....

Figure 1: Framework of JDEval and MDEval

To assess the effectiveness of the meta-evaluator using individual scoring, we first calculate the Pearson correlation coefficient between the scores assigned by the evaluators and the human annotation across all the dataset, using this as ranking baseline to rank evaluators on our collected dataset.

4 Dataset

4.1 Pairwise comparison

For pairwise comparison, each instance of our dataset comprises a tuple (I, O1, O2, p), Our dataset is collected from MT-bench (Zheng et al., 2024) and LLMBar (Zeng et al., 2023).

MT-bench MT-bench is a benchmark comprising 80 high-quality multi-turn questions designed to evaluate models' proficiency in multi-turn conversation and instruction-following abilities. The format of each data in MT-bench follows the structure (I, O1, O2, p). However, multiple human experts evaluate the two answers for the same question, leading to potential discrepancies in their judgments. To address this, we employ a voting principle to determine the superior answer for each (I, O1, O2) combination, resulting in a unique p value.

To ensure diversity and balance in our dataset, we randomly select data entries to ensure that each of the 80 questions appears 5 times, yielding a total of 400 data. However, to further enhance diversity and avoid redundancy in questions and answers, we apply additional filtering criteria. Through multiple rounds of filtering, we refine the dataset to ensure that each question (I) corresponds to completely distinct pairs of answers (O1 and O2). Following these rigorous filtering steps, we arrive at a final dataset comprising 391 data.

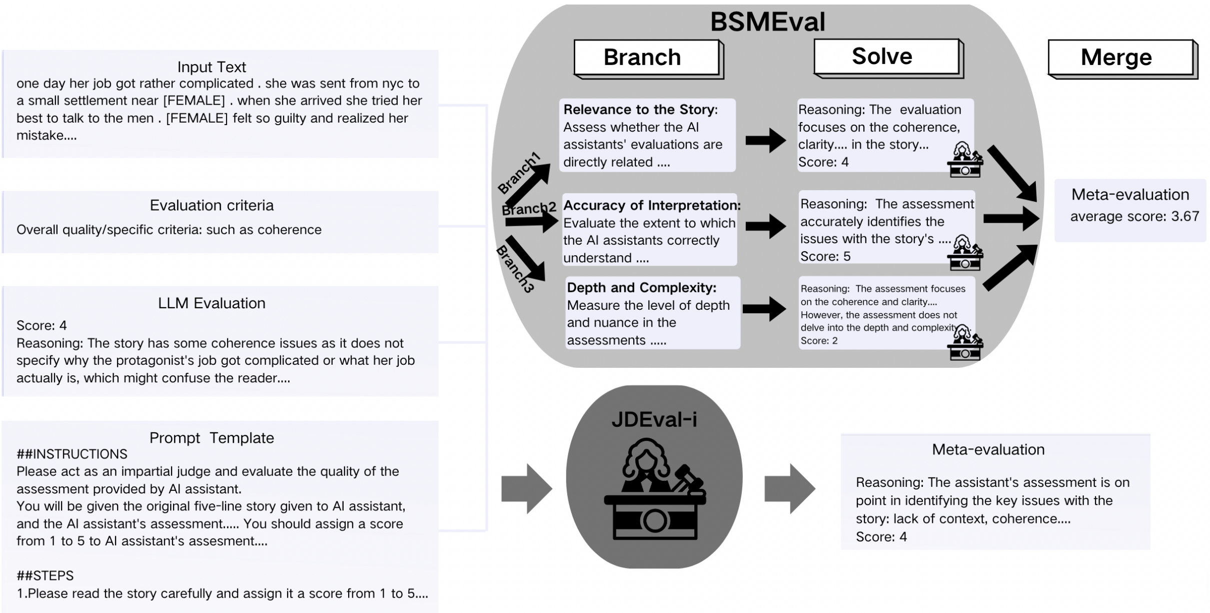
LLMBar LLMBar serves as a meta-evaluation benchmark crafted to assess the ability of LLM evaluators in discerning instruction-following outputs. It comprises two main components: the natural set, which gathers instances from existing human-preference datasets, and the adversarial set. For our study, we exclusively utilize the natural set of LLMBar, which encompasses 100 data.

In total, we aggregate 491 data from both MT-bench and LLMBar, combining the strengths of both datasets to create a comprehensive evaluation corpus.

4.2 Individual scoring

For individual scoring, each instance in our dataset comprises a tuple (I, S), where I represents the input text and S denotes the human labeled score for the text regarding its overall quality or a specific aspect. We derive our dataset from OpenMEVA (Guan et al., 2021) and SummEval (Fabbri et al., 2021).

SummEval Each summary undergoes scoring by 8 human labelers using a 5-point Likert scale across



Input Text
one day her job got rather complicated . she was sent from nyc to a small settlement near [FEMALE] . when she arrived she tried her best to talk to the men . [FEMALE] felt so guilty and realized her mistake....

Evaluation criteria
Overall quality/specific criteria: such as coherence

LLM Evaluation
Score: 4
Reasoning: The story has some coherence issues as it does not specify why the protagonist's job got complicated or what her job actually is, which might confuse the reader....

Prompt Template
##INSTRUCTIONS
Please act as an impartial judge and evaluate the quality of the assessment provided by AI assistant.
You will be given the original five-line story given to AI assistant, and the AI assistant's assessment..... You should assign a score from 1 to 5 to AI assistant's assesment....

##STEPS
1.Please read the story carefully and assign it a score from 1 to 5....

Relevance to the Story: Assess whether the AI assistants' evaluations are directly related ....

Accuracy of Interpretation: Evaluate the extent to which the AI assistants correctly understand ....

Depth and Complexity: Measure the level of depth and nuance in the assessments .....

Reasoning: The evaluation focuses on the coherence, clarity.... in the story...
Score: 4

Reasoning: The assessment accurately identifies the issues with the story's ....
Score: 5

Reasoning: The assessment focuses on the coherence and clarity.... However, the assessment does not delve into the depth and complexity....
Score: 2

Meta-evaluation
average score: 3.67

Meta-evaluation
Reasoning: The assistant's assessment is on point in identifying the key issues with the story: lack of context, coherence....
Score: 4

Figure 2: Framework of JDEval-i and BSMEval

four dimensions: coherence, fluency, consistency, and relevance. To align with our dataset format, we focus on the coherence dimension. Since each data receives multiple coherence scores, we compute the average of these scores to determine the final coherence score for the text. Consequently, we format the data as (I, S). Subsequently, we randomly select 250 instances from this processed data.

OpenMEVA Each story receives a 5-point Likert scale rating for its overall quality. We calculate the average rating as the definitive human judgment for the text. Similarly, we randomly select 250 instances from this processed data.

In total, our dataset comprises 500 instances, integrating scores from both SummEval and OpenMEVA, providing a robust evaluation corpus.

5 Experiment

5.1 Setup

We have selected the following models as evaluators: *gpt-4-1106-preview, gpt-3.5-turbo-0125, Mixtral-8x7B-Instruct-v0.1, Qwen1.5-7b-chat, Qwen1.5-14b-chat, and Qwen1.5-72b-chat*. As for meta-evaluators, we have chosen *gpt4, Qwen1.5-7b-chat, Qwen1.5-14b-chat, and Qwen1.5-72b-chat*. The reason for this choice is twofold: on one hand, to explore the capability of different models in meta-evaluation, and on the other hand, to investigate the impact of model size on meta-evaluation capability.

5.2 Pairwise comparison

As mentioned in Methodology, we will firstly calculate accuracy of evaluators' evaluation results with human-labeled results to establish a baseline ranking. The evaluators' accuracy are shown in Table 1 with human as meta-evaluator. According to the Table 1, GPT-4 demonstrates the highest accuracy, reaching a consistency rate of 0.7332 when compared to human judgments. Next in line is Qwen1.5-72b-chat, boasting an accuracy of 0.6965. Following closely is GPT-3.5, with an accuracy rate of 0.6741. Subsequently, we have Qwen1.5-14b-chat and Mixtral-8x7B-Instruct-v0.1, Qwen1.5-7b-chat. This ranking aligns with the expected capabilities of large models and is in accordance with common expectations. Next, we will provide the experiment details of our two frameworks to perform meta-evaluation based on pairwise comparison.

5.2.1 JDEval

Setup For this framework, as previously mentioned, each original data is formatted as (I, O1, O2). For each (I, O1, O2) tuple, we randomly selected the evaluation results of two evaluators to create data in the format (I, O1, O2, W1, E1, W2, E2), which were then input into the meta-evaluator. We first let the meta-evaluator decide which answer is better on its own. Then, based on its own evaluation, it will meta-evaluate two evaluators' evaluations considering the better responses chosen by evaluator(Wi) and Explanation(Ei). Addi-

Evaluator	gpt-4-1106-preview	gpt-3.5-turbo	Mixtral-8x7B-Instruct	Qwen1.5-7B-Chat	Qwen1.5-14B-Chat	Qwen1.5-72B-Chat
Accuracy	0.7332	0.6741	0.5785	0.501	0.6293	0.6965
Pearson corr.	0.6183	0.5501	0.6098	0.5902	0.6416	0.6223

Table 1: The first line represents the accuracy of each evaluator's evaluation results with human annotation in pairwise comparison paradigms. The second line represents the pearson correlation of each evaluator's evaluation results with human annotation under individual scoring paradigms. They are used as baselines respectively for pairwise comparison and individual scoring paradigms.

tionally, we randomly selected different evaluators three times for each (I, O1, O2) tuple. As a result, each meta-evaluator yielded a total of 1473 judgments. The distribution of competition among evaluators for four meta-evaluators is depicted in Appendix A. Each meta-evaluator showcases an approximately equal proportion of pairwise comparisons among evaluators, averaging around 20%, thereby ensuring fairness in the calculation of win rates.

Results for Win rate Examining the win rate distribution of four meta-evaluatrs presented respectively in Figure 3, Figure 4, Figure 5 and Figure 6, we find that when GPT-4 assumes the role of the meta-evaluator, it adeptly evaluates the quality of evaluators, resulting in meta-evaluation outcomes that align perfectly with the conclusions drawn from Table 1. While Qwen1.5-72b-chat exhibited minor errors in evaluating itself and GPT-3.5, it accurately assessed the quality of other evaluators, demonstrating a certain level of meta-evaluation proficiency. Qwen1.5-14b-chat effectively evaluates the quality of evaluators with similar structures to itself but demonstrates bias towards answers from evaluators with similar structures when faced with evaluators of different structures. As for Qwen1.5-7b-chat, it lacks significant meta-evaluation capability and displays narcissistic tendencies.

Results for Elo score For each meta-evaluator, we have 1473 matches between six evaluators. So we can use the formula 5 to calculate the Elo scores of these evaluators. It is worth noting that the order of matches can affect the final scores when calculating Elo scores. To eliminate this effect, we shuffled the 1473 results several times and then averaged the Elo scores of the evaluators. The details will be introduced later.

Table 2 presents the scores of each evaluator acquired through the Elo system after shuffling the meta-evaluation results 200 times. It is notable that when GPT-4 and Qwen72b serve as meta-evaluators, the score ranking closely correspond to the ranking based on accuracy of the evaluators. However, a slight deviation is observed where Mistral's score slightly surpasses that of Qwen14b. This discrepancy contradicts the win rate distribution tables, where both meta-evaluators perceive Qwen14b's win rate to be higher than Mistral's. This may be caused by insufficient match numbers between evaluators.

As mentioned before, the order of matches between evaluators can impact the evaluator's score. We shuffle respectively GPT-4's meta-evaluation results 5 times, 100 times, 200 times, 500 times, and 1000 tims. The results are shown in Appendix C. After 200 shuffles, minor fluctuations in the scores assigned by each meta-evaluator to individual evaluators are observed, limited to only 1 or 2 points. These slight score adjustments do not impact the overall rankings of evaluators. We establish 200 shuffles as a threshold, considering that each evaluator participates in an average of 450 matches. However, as the total number of matches each evaluator engages in increases, it may become necessary to shuffle the match results more times to eliminate the influence of match order when determining the final scores.

5.2.2 MDEval

Setup For this debate-based framework, the input to meta-evaluator is the same as JDEval. If the meta-evaluator can directly determine which evaluation is more reasonable based on the evaluations of the two evaluators, it directly outputs the better evaluator along with reasoning. If a direct judgment is not possible, the meta-evaluation will enter debate mode, where meta-evaluator can ask questions to both evaluators, and then initiate a debate between them. To ensure consistency in the debate, during the n-th round of debate, both evaluators will have access to the questions from the judge from the previous (n-1) rounds, as well as their own statements and those of their opponent. The judge must make its choice within five rounds. The prompt used for meta-evaluator and evaluators

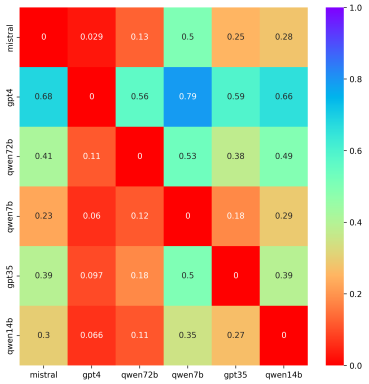


Figure 3: Win rate for GPT-4 as Meta-evaluator using JDEval

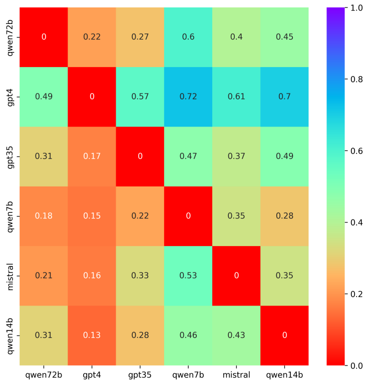


Figure 4: Win rate for Qwen72b as Meta-evaluator using JDEval

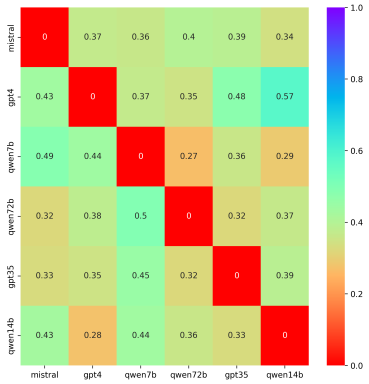


Figure 5: Win rate for Qwen14b as Meta-evaluator using JDEval

Meta-evaluator\Evaluator	**gpt-4-1106-preview**	**gpt-3.5-turbo**	**Mixtral-8x7B-Instruct**	**Qwen1.5-7B-Chat**	**Qwen1.5-14B-Chat**	**Qwen1.5-72B-Chat**
gpt-4-1106-preview	1180.27	993.386	949.849	890.387	938.493	1047.614
Qwen1.5-7B-Chat	983.758	1025.013	980.695	991.958	1016.014	1002.561
Qwen1.5-14B-Chat	1018.635	997.716	989.476	984.161	996.349	1013.664
Qwen1.5-72B-Chat	1133.82	1005.121	964.69	907.955	964.124	1024.291

Table 2: Elo score of each evaluator with different meta-evaluators using JDEval

to perform a debate are shown in Appendix E.

Result From the JDEval, we find that using win rate can be more efficient to rank the evaluators in our experiments. So we just check the win rate result of each meta-evaluator. The GPT-4 and Qwen72b's win rate are shown in Figure 7 and Figure 8.

When GPT-4 serves as the meta-evaluator using this method, it can accurately rank the evaluators. However, when Qwen72b acts as the meta-evaluator using this framework, it shows a clear preference for Mistral's responses, considering Mistral's performance as an evaluator to be superior to that of GPT-3.5 and Qwen72b itself as an evaluator. However, it successfully discerns the quality of all other evaluators. As for Qwen7b and Qwen14b, they perform poorly in this setting too. Their results are shown in Appendix B.

5.2.3 Other findings

In JDEval, when employing the meta-evaluator to assess the evaluation results of evaluators, we task it with determining which of the two outputs is superior. We calculate the accuracy of the meta-evaluator compared to human-labeled results. Surprisingly, we observed an enhancement in accuracy for both Qwen7b and Qwen14b when they assumed the role of meta-evaluators. Upon analysis, we attribute this improvement to the fact that when serving as meta-evaluators to adjudicate between two outputs, they have access to the judgments of both evaluators. Regardless of whether these results align with their own judgments as evaluators, when acting as meta-evaluators, they conscientiously consider and ponder over the judgments of the two evaluators, resulting in more accurate assessments.

This observation implies that directly employing small-scale LLMs for evaluation tasks may not yield optimal results. However, when provided with reference judgment results from evaluators, these models can reflect on those judgments and enhance their evaluation performance. This demonstrates that the meta-evaluation capability of small-scale LLMs, such as 7b and 14b, can contribute to improving their evaluation assessments.

5.3 Individual scoring

As mentioned in Methodology, we first calculate the Pearson correlation coefficient between the scores assigned by the evaluators and the human annotation on the two datasets to establish a ranking baseline. Table 1 display the Pearson correlation coefficients of six evaluators on the collected dataset. We noticed that among the top-performing models, including GPT-4 and GPT-3.5, their performance didn't meet our expectations on the dataset. The Pearson correlation between GPT-3.5 and human annotations is the lowest among all evaluators. However, normally, GPT-3.5's performance should be better than that of Qwen-14B and Qwen-7B. This indicates the instability of evaluation methods based on scoring to some extent, which fail to accu-

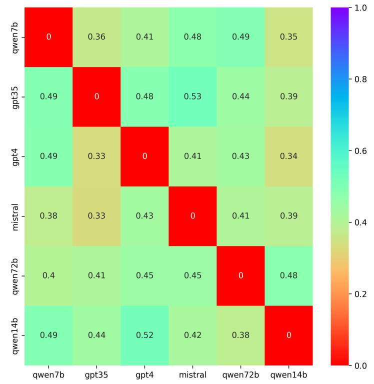


Figure 6: Win rate for Qwen7b as Meta-evaluator using JDEval

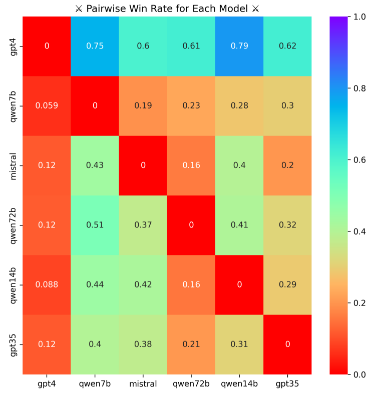


Figure 7: Win rate of GPT-4 as Meta-evaluator using MDEval

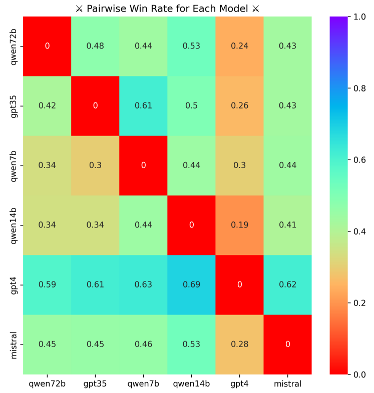


Figure 8: Win rate of Qwen72b as Meta-evaluator using MDEval

Meta-evaluator	Original Accuracy	Post Accuracy 1	Post Accuracy 2	Post Accuracy 3
Qwen1.5-7B-Chat	0.501	0.6784	0.6591	0.6687
Qwen1.5-14B-Chat	0.6293	0.6741	0.6633	0.6857

Table 3: Original accuracy is the model's accuracy with human labels when directly serving as evaluator. While post accuracy is the model's accuracy when serving as meta-evaluator.

rately reflect the true capabilities of the evaluators.

5.3.1 JDEval-i

Setup Given that the evaluator's evaluation result is (I,S,E), where I represents Input text, S represents Score assigned by evaluator and E represents evaluator's Explanation, the meta-evaluator are required to assign a score ranging from 1 to 5 to the evaluator's assessment based on S and E. The evaluator's final score used for ranking evaluators, is computed as the average score across all data.

Result Table 5 presents the average scores given by four meta-evaluators to the six evaluators across the whole dataset. Through the Table 5, we can observe that when GPT4 serves as the meta-evaluator, it consistently assigns the highest scores to itself, followed by GPT-3.5. Similarly, when Qwen72b acts as the meta-evaluator, it also assigns relatively high scores to GPT4, GPT-3.5, and itself. However, the rankings obtained from the assessments of these four meta-evaluators do not align with the rankings based on the evaluators' Pearson correlation coefficients. This indicates that using individual scoring for meta-evaluation may not be very effective. In fact, this inconsistency can also be observed from the Pearson correlation coefficients obtained by using large models as evaluators compared to human annotations.

5.3.2 BSMEval

Setup This method adopts the Branch-Solve-Merge approach propose by (Saha et al., 2023), breaking down the meta-evaluation task into multiple branch tasks. In our experiment, we asked each meta-evaluator to decide how to break down the task into no more than five sub-task. Then, the meta-evaluator rates evaluators on each sub-task, and finally, we calculate the average score across all branches as the evaluator's final score. The results are shown in Table 6 in Appendix D.

Result Similar to JDEval-i, mentioned above none of the rankings obtained from these four meta-evaluators align with the rankings based on the evaluators' Pearson correlation coefficients.

6 Conclusion

For meta-evaluation, we proposed respectively two frameworks based on individual scoring and pairwise comparison. We find that pairwise comparison paradigm is more suitable to conduct meta-evaluation than individual scoring paradigms and the meta-evaluation ability of LLMs correlates well with their general performance on standard benchmarks. Both GPT-4 and Qwen72b can successfully complete the meta-evaluation task using JDEval and MDEval based on pairwise comparison. However, smaller models like Qwen7b and Qwen14b do not perform well in meta-evaluation task. Additionally, we discovered that models have a self-enhancing capability. When using the second method, the accuracy of Qwen7b and Qwen14b as meta-evaluators significantly improved compared to their performance as direct evaluators.

7 Limitations

Our study has the following limitations:

1. The criterion for assessing the performance of the meta-evaluator in this paper is to compare whether the rankings it provides for evaluators are consistent with those obtained through human annotation. This standard may lack a certain level of rigor, and further human annotation of the meta-evaluator's assessment could make assessing meta-evaluators' performance more rigorous.

2. The dataset collected for individual scoring in this paper only includes assessments of the overall quality or coherence dimension of the text, which may not be comprehensive enough. More data based on other assessment dimensions could be collected for experimentation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2023a. Storyer: Automatic story evaluation via ranking, rating and reasoning. *Journal of Natural Language Processing*, 30(1):243–249.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023b. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

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. *arXiv preprint arXiv:2105.08920*.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

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.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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*.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024.

Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Match number distribution

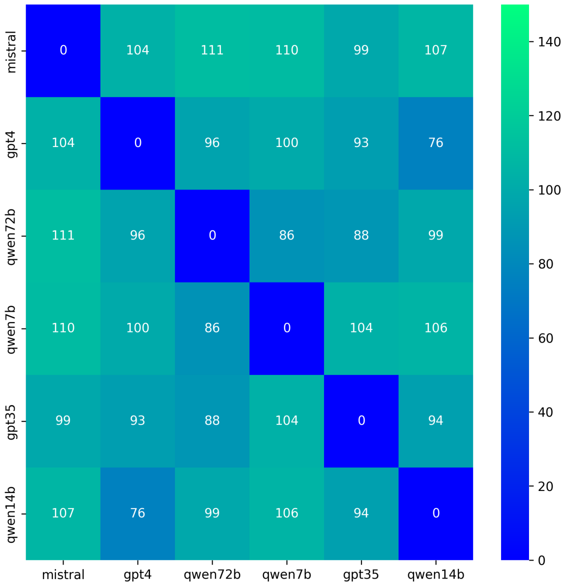


Figure 9: Match Distribution for GPT-4 as Meta-evaluator

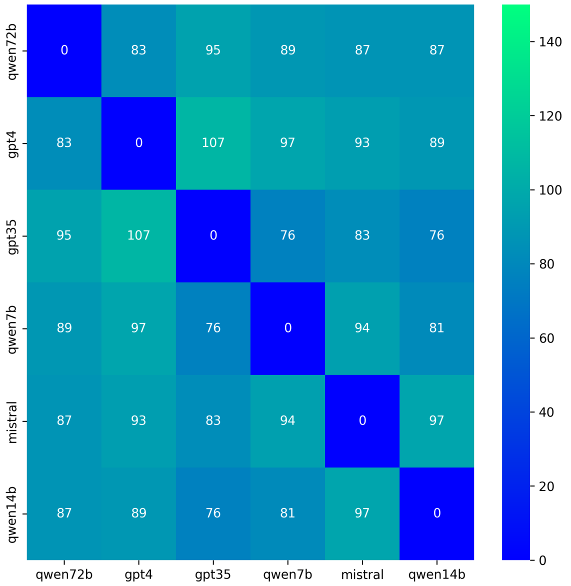


Figure 10: Match Distribution for Qwen72b as Meta-evaluator

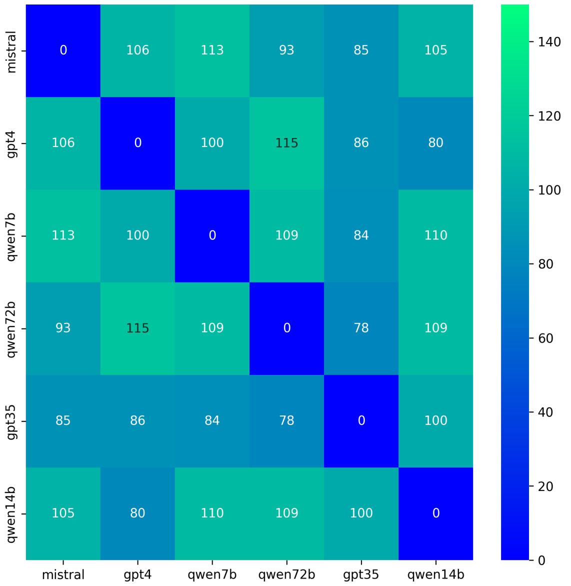


Figure 11: Match Distribution for Qwen14b as Meta-evaluator

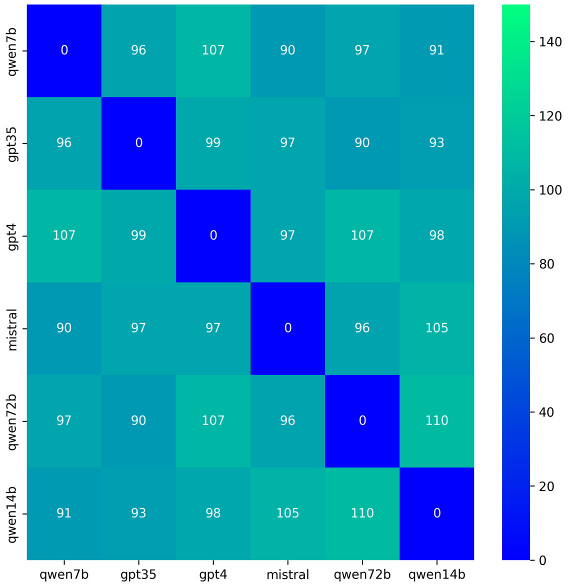


Figure 12: Match Distribution for Qwen7b as Meta-evaluator

B Win rate distribution of MDEval

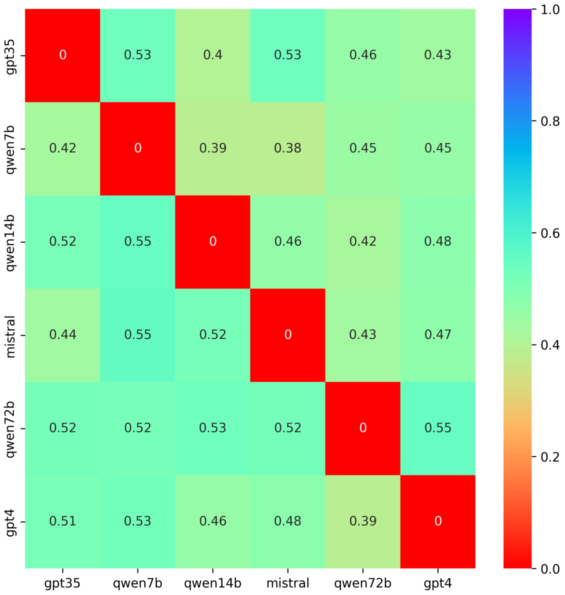


Figure 13: Win Rate Distribution for Qwen14b as Meta-evaluator using MDEval

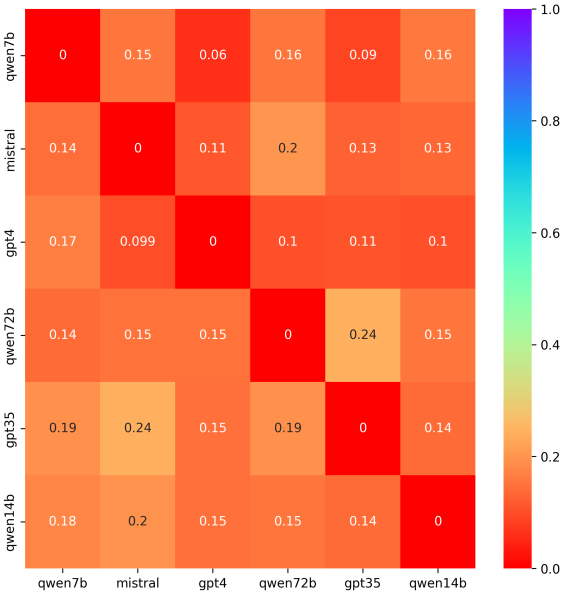


Figure 14: Win Rate Distribution for Qwen7b as Meta-evaluator using MDEval

C Elo score for different shuffle times using JDEval

Shuffle times\Evaluator	**gpt-4-1106-preview**	**gpt-3.5-turbo**	**Mixtral-8x7B-Instruct**	**qwen1.5-7B-Chat**	**qwen1.5-14B-Chat**	**qwen1.5-72B-Chat**
5	1179.746	996.1992	947.5399	887.925	939.9354	1048.6544
100	1178.7398	992.3116	949.346	891.1954	940.5041	1047.9032
200	1178.9103	993.2556	949.3424	890.9098	938.3801	1049.2018
500	1179.9305	993.3637	949.3577	890.4678	938.1173	1048.7631
1000	1179.7824	993.7034	948.5853	890.7663	938.6064	1048.5562

Table 4: Elo score of each evaluator with GPT-4 as meta-evaluator for different shuffle times

D Individual scoring results

Meta-evaluator\Evaluator	**gpt-4-1106-preview**	**gpt-3.5-turbo**	**Mixtral-8x7B-Instruct**	**qwen1.5-7B-Chat**	**qwen1.5-14B-Chat**	**qwen1.5-72B-Chat**
gpt-4-1106-preview	4.744	4.522	4.178	3.906	3.994	4.14
qwen1.5-7B-Chat	3.976	3.8988	3.7645	3.895	4.11	3.889
qwen1.5-14B-Chat	4.21	4.084	4.083	4.193	4.264	4.282
qwen1.5-72B-Chat	4.703	4.605	4.496	4.406	4.475	4.567

Table 5: Average score assigned by different meta-evaluators using JDEval-i

Meta-evaluator\Evaluator	**gpt-4-1106-preview**	**gpt-3.5-turbo**	**Mixtral-8x7B-Instruct**	**qwen1.5-7B-Chat**	**qwen1.5-14B-Chat**	**qwen1.5-72B-Chat**
gpt-4-1106-preview	4.324	4.106	3.9654	3.625	3.890	4.167
qwen1.5-7B-Chat	2.578	2.5283	2.641	2.7779	2.8202	2.939
qwen1.5-14B-Chat	3.0032	2.7199	2.9953	3.1303	3.0467	3.4534
qwen1.5-72B-Chat	4.2864	4.0616	3.9704	4.1312	4.1624	4.1328

Table 6: Average score assigned by different meta-evaluators using BSMEval

E Prompts for LLM

Evaluation prompt for pairwise comparison

```
Please act as an impartial judge and evaluate the quality of the responses
provided by two AI assistants to the user questions.
You should choose the assistant that follows the user's instructions and
answers the user's questions 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 responses of the two assistants and
provide a short explanation.
Avoid any position 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.

Here are questions and answers generated by two AI assistants:
------------------------------------------
***for AI assistant 1:
### Question:
{question}
### Answer of AI assistant 1:
{answer_1}
***for AI assistant 2:
### Question:
{question}
### Answer of AI assistant 2:
{answer_2}
------------------------------------------
Remember firstly you should provide a short explanation and then judge which
AI assistant's answer is better and then decide which assistant is the winner.
1 means AI assistant 1 is the winner, 2 means AI assistant 2 is the winner,
0 means the two assistant are equally good, and there is no winner.
Your explanation should be between the tag <Explanation> and </Explanation>
and the winner that you choose should be between the tag <Winner> and </Winner>.

Please output your judgement by strictly following this format:

<Explanation>
(your explanation for comparing the two assistants' answers )
</Explanation>
<Winner>
(0 or 1 or 2)
</Winner>

Remember you must output your judgement by strictly following the above format.
```

Table 7: Evaluation prompt for pairwise comparison

Evaluation prompt for individual scoring(Summeval)

```
Please act as an impartial judge and evaluate the coherence of the text. Your
 evaluation should just consider the coherence aspect of the text.
Begin your evaluation by providing a short explanation. After providing your
 explanation, you must rate the text on a scale of 1 to 5.

Be as objective as possible. Do not allow the length of the text to influence you
 evaluation. Please focus solely on coherence and do not take into account any
 other factors.
 Here is the text:
 {text}

 Please output your evaluation by strictly following this format:

 <Explanation>
 (your explanation for evaluating the coherence of the text)
 </Explanation>

 <Rating>
 (give a score between 1 to 5 based on your explanation)
 </Rating>
```

Evaluation prompt for individual scoring(OpenMEVA)

```
Please act as an impartial judge and evaluate the overall quality of the five-line
 story.Your evaluation should consider factors such as the coherence,fluency,
 consistency,interestingness and creativity of the short story.
 Begin your evaluation by providing a short explanation. After providing
 your explanation, you must rate the story on a scale of 1 to 5.

 Be as objective as possible. Do not allow the length of the story to
 influence your evaluation.
 Here is the story:
 {text}

 Please output your evaluation by strictly following this format:

 <Explanation>
 (your explanation for evaluating the story)
 </Explanation>

 <Rating>
 (give a score between 1 to 5 based on your explanation)
 </Rating>
```

Table 8: Evaluation prompt for individual scoring

Meta-evaluation prompt for Judge without Debate

```
Please act as an impartial judge and evaluate the quality of the assessments
provided by two AI assistants.
You will be given the original question and two answers(AnswerA and AnswerB)
given to two assistants(assistant1 and assistant2), and the two assistants'
assessments about the two answers. The two answers are wriiten by human,
not by two assistants.
You should choose the assistant which offers a more reasonable and more fair
assessment to the two answers.
##STEPS
1.Read carefully the original question and the two answers given to two
assistants, choose which answer you think is better and give your reason.
2.Read the two assistants' assessments for two answers. If both of two
assisstants chose the same better answer as you chose, you should read their
explanations why they made this choice, and judge which explanation is more
reasonable. "equally good" means the two assistsnts' explanations are equally
reasonable, assistant1 means the assistant1's explanation is more reasonable
than the assistant2's explanation. assistant2 means the assistant2's
explannation is more reasonable than the assistant1's explanation. If one of
the assistant chose the same better answer as you did ,and the other assistant
chose differently from what you chose, you should choose the assistant
who chose same better answer with you as better evaluator and your reasoning
should be that one evaluator made the right choice and the other one chose the
wrong better answer. If both of them chose differently from what you chose as
better answer, no matter which answer they thought was better, you should output
"equally good" and your reasoning should be that two assistants both made
the wrong assessments.

Here are original question and answers:
### Question:
{question}
### AnswerA:
{answer1}
### AnswerB:
{answer2}
Here are assessments made by two assistants:
### Assessment of assistant1:
{result1}
### Assessment of assistant2:
{result2}
## Attention
1.Avoid any position biases and ensure that the order in which the
assistant's judgements were presented does not influence your decision.
2.Do not allow the length of the reason of assistant to influence your
evaluation.
3.Be as objective as possible.
4.The two answers have nothing to do with the two assistants.
5.Please output your judgement by strictly following this format:
<Reasoning_for_choosing_better_answer>
(your reasoning process only for comparing the two answers)
</Reasoning_for_choosing_better_answer>
<Better_answer>
(choose AnswerA or AnswerB or equally good)
</Better_answer>
<Reasoning_for_choosing_better_assistant>
(your reasoning process only for comparing the two assistants' assessments )
</Reasoning_for_choosing_better_assistant>
<Better_assistant>
(your final judgement for which assistant is better:choose Assistant1 or Assistant2 or
equally good)
</Better_assistant>
```

Table 9: Meta-evaluation prompt for Judge without Debate

System Prompt for Judge with Debate

```
##INSTRUCTIONS
Please act as an impartial Judge and evaluate the quality of the assessments
provided by two AI assistants. The two assistants will have a debate between
them to pursuade you that their own assessments are more reasonable and fair.
You will be given the original question and two answers(AnswerA and AnswerB)
given to two assistants(assistant1 and assistant2), the two assistants'
assessments about the two answers and the current debate transcript.
You should choose the assistant which offers a more reasonable and more fair
assessment to the two answers according to their debate.

##STEPS
1. You should firstly read the two assistants' assessments for two answers and
the current debate transcript. If you can decide which assistant's assessment is
better and is more reasonable, you should end the debate and output the better
assistant and your reasoning process.

2. If you cannot decide which assistant's assessment is better by two AI
assistants' initial arguments, you can ask both assistants quetions according
to their evaluation results to help you choose which assistant's evaluation
results is more accurate and better. You should output your question about the
two assistant's debate or their assessments.

3. Eventually, you, as the Judge, decides to end the debate when you can
choose the better AI assistant.

##ATTENTION
1. There could be a number of rounds for debate. But you must make your
final judgement within five rounds.
2. In each round, both debaters make their arguments simulatenously.
3. You must be concise and clearly state your question or your reasoning,
which shoule be limited to 200 words.

Here are original question and answers:
### Question:
{input}

### AnswerA:
{output1}

### AnswerB:
{output2}

Here are assessments made by two assistants:
### Assessment of assistant1:
{eval_results1}

### Assessment of assistant2:
{eval_results2}
```

Table 10: Meta-evaluation system prompt for Judge with Debate

User Prompt for Judge with Debate

```
Here is the current debate transcript:
{transcript}

If you can decide which assistant's evaluation results is better, output
the better assistant and your reasoning process by following this format:
<Reasoning_for_choosing_better_assistant>
(your reasoning process only for comparing the two assistants' assessments )
</Reasoning_for_choosing_better_assistant>

<Better_assistant>
(your final judgement for which assistant is better:choose Assistant1 or
Assistant2 or equally good)
</Better_assistant>

If you cannot decide which assistant's evaluation results is more reasonable
and fair, you can ask both assistants quetions,directly output your questions
by following this format:
<Question>
(your question about the two assistant's debate or their assessments)
</Question>
```

Table 11: Meta-evaluation user prompt for Judge with Debate

Meta-evaluation prompt for individual scoring(OpenMEVA)

```
##INSTRUCTIONS
Please act as an impartial judge and evaluate the quality of the assessment
provided by AI assistant.
You will be given the original five-line story given to AI assistant, and
the AI assistant's assessment including a score ranging from 1 to 5 and its
reason for why giving such a score.
You should assign a score from 1 to 5 to AI assistant's assesment, where 1
indicates the AI assistant's assessment is poor, and 5 indicates it is
excellent.

##STEPS
1.Please read the story carefully and assign it a score from 1 to 5 for the
story's overall quality,considering factors such as the coherence,fluency,
consistency,interestingness and creativity of the short story.
2.Read the AI assistant's assessment carefully.Based on your own rating
for story,you should judge whether the assistant's assessment is reasonable
and assign a score from 1 to 5 to the AI assistant's assesment.

Here is the story:
### Story:
{text}

Here is the assessment made by AI assistant:
### Assessment:
{eval_result}

## Attention
1.Do not allow the length of the reason of assistant to influence your
evaluation for AI assistant.
2.Be as objective as possible.
3.The five-line story has nothing to do with the AI assistant,the story is
written by human,not by AI assistant.
4.Please output your judgement by strictly following this format:

<Reasoning_for_scoring_story>
(your reasoning process only for scoring the story)
</Reasoning_for_scoring_story>

<Score_for_story>
(assign a score from 1 to 5 to the story)
</Score_for_story>

<Reasoning_for_scoring_Assistant>
(your reasoning process only for scoring assistant's assessment )
</Reasoning_for_scoring_Assistant>

<Score_for_Assistant>
(assign a score from 1 to 5 to the assistant's assessment)
</Score_for_Assistant>
```

Table 12: Meta-evaluation prompt for individual scoring(OpenMEVA)

Meta-evaluation prompt for individual scoring(Summeval)

```
##INSTRUCTIONS
Please act as an impartial judge and evaluate the quality of the assessment
provided by AI assistant.
You will be given the original text given to AI assistant, and the AI
assistant's assessment for the text's coherence aspect, including a score
ranging from 1 to 5 and its reason for why giving such a score.
You should assign a score from 1 to 5 to AI assistant's assesment, where 1
indicates the AI assistant's assessment is poor, and 5 indicates it is
excellent.

##STEPS
1.Please read the text carefully and assign it a score from 1 to 5 for the
text's coherence,You should focus solely on coherence and do not take into
account any other factors.
2.Read the AI assistant's assessment carefully and assign a score from 1
to 5 to the AI assistant's assesment for the text's coherence

Here is the text:
### Text:
{text}

Here is the assessment made by AI assistant:
### Assessment:
{eval_result}

## Attention
1.Do not allow the length of the reason of assistant to influence
your evaluation for AI assistant.
2.Be as objective as possible.
3.The text has nothing to do with the AI assistant,the text is written by
human, not by AI assistant.
4.Please output your judgement by strictly following this format:

<Reasoning_for_scoring_text_coherence>
(your reasoning process only for scoring the story)
</Reasoning_for_scoring_text_coherence>

<Score_for_text_coherence>
(assign a score from 1 to 5 to the story)
</Score_for_text_coherence>

<Reasoning_for_scoring_Assistant_assessment>
(your reasoning process only for scoring assistant's assessment )
</Reasoning_for_scoring_Assistant_assessment>

<Score_for_Assistant_assessment>
(assign a score from 1 to 5 to the assistant's assessment)
</Score_for_Assistant_assessment>
```

Table 13: Meta-evaluation prompt for individual scoring(Summeval)