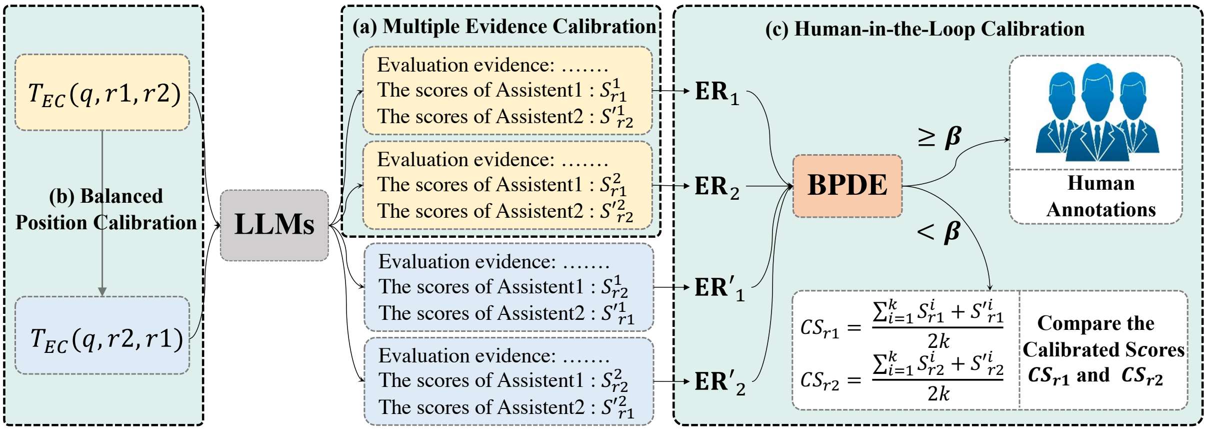

$T_{EC}(q, r1, r2)$
(b) Balanced Position Calibration
$T_{EC}(q, r2, r1)$
LLMs
(a) Multiple Evidence Calibration
Evaluation evidence: ....
The scores of Assistent1 : S^1_{r1}
The scores of Assistent2 : S'^1_{r2}
Evaluation evidence: ....
The scores of Assistent1 : S^2_{r1}
The scores of Assistent2 : S'^2_{r2}
Evaluation evidence: ....
The scores of Assistent1 : S^1_{r2}
The scores of Assistent2 : S'^1_{r1}
Evaluation evidence: ....
The scores of Assistent1 : S^2_{r2}
The scores of Assistent2 : S'^2_{r1}
(c) Human-in-the-Loop Calibration
ER_1
ER_2
ER'_1
ER'_2
BPDE
$\geq \beta$
$< \beta$
Human Annotations
$CS_{r1} = \frac{\sum_{i=1}^{k} S^i_{r1} + S'^i_{r1}}{2k}$
$CS_{r2} = \frac{\sum_{i=1}^{k} S^i_{r2} + S'^i_{r2}}{2k}$
Compare the Calibrated Scores CS_{r1} and CS_{r2}

Figure 3: Demonstration of our calibration framework with three calibration methods. S_r and $S_r^{'}$ denotes scores of the response r in the first and second positions, respectively. BPDE is short for Balanced Position Diversity Entropy score, which is calculated based on the evaluation results (ER) of MEC and BPC.

[Question]
{Q}
[The Start of Assistant 1's response]
{R1}
[The End of Assistant 1's response]
[The Start of Assistant 2's response]
{R2}
[The End of Assistant 2'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, and level of detail 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 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. Then, output two lines indicating the scores for Assistant 1 and 2, respectively.
Output with the following format:
Evaluation evidence: <evaluation explanation here>
The score of Assistant 1: <score>
The score of Assistant 2: <score>

Table 3: The evidence calibration evaluation template that prompts LLMs to generate the evaluation evidence first (red text), and then evaluate two responses.

judgment in such a situation. To further investigate this issue, we grouped all the examples based on the score difference between the two responses. As shown in Figure 2, we found that when the score difference between the two responses is small (e.g., score gap ≤ 1), the evaluation results of GPT-4 are significantly affected by the position of the responses. On the other hand, when the score difference between the two responses is large (e.g., score gap ≥ 3), GPT-4's evaluation results are relatively stable.

3 Calibrating the Positional Bias

We have identified that positional bias can significantly impact the evaluation results of LLMs, making them unfair evaluators. In this section, we propose a calibration framework with three simple yet effective strategies to alleviate this bias to achieve a more reliable and fair evaluation result.

3.1 Multiple Evidence Calibration

Previous studies (Zheng et al., 2023; Wang et al., 2023b) utilize the evaluation template that draws the conclusion first and then makes an explanation, e.g., the template used in Table 1. However, due to the nature of the auto-regressive model, the conclusions generated by the model are not supported by the explanation generated afterward. To this end, as shown in Table 3, we design an evidence calibration (EC) evaluation template $T_{EC}(Q, R1, R2)$ that requires the model to generate the explanation (evaluation evidence) first and then give the score. In this way, the score can be calibrated with the evaluation evidence. To further improve the reliability of the evaluation, rather than generating only a single EC score for each response, we perform a multiple evidence calibration (MEC, Figure 3(a)) that samples k EC scores $\{S^1_{r1}, \ldots, S^k_{r1}\}$ and $\{S'^1_{r2}, \ldots, S'^k_{r2}\}$ for responses $r1$ and $r2$, where S_r and $S_r^{'}$ denotes scores of the response r at the first and second positions, respectively.

3.2 Balanced Position Calibration

We further employ a balanced position calibration (BPC) strategy to alleviate the previously identified positional bias of LLMs. As shown in Figure 3(b), for each example $(q, r1, r2)$, BPC additionally creates a query prompt $T_{EC}(q, r2, r1)$ by swapping the position of two responses in the original query prompt $T_{EC}(q, r1, r2)$. Combined with MEC, we can achieve $2k$ scores $\{S_{r1}^1, \ldots, S_{r1}^k, \ldots, S_{r1}^{\prime 1}, \ldots, S_{r1}^{\prime k}\}$ and $\{S_{r2}^{\prime 1}, \ldots, S_{r2}^{\prime k}, \ldots, S_{r2}^1, \ldots, S_{r2}^k\}$ for $r1$ and $r2$, respectively. The final calibrated scores of two responses (CS_{r1} and CS_{r2}) are the average of the $2k$ scores:

$$CS_R = \sum_{i=1}^{k} \frac{S_R^i + S_R^{\prime i}}{2k}, R = r1, r2 \tag{2}$$

and we regard the response with the higher average score as the better response.

3.3 Human-in-the-Loop Calibration

In addition to the automatic calibration strategies, another interesting question we want to explore is whether Human-In-The-Loop Calibration (HITLC) which performs the cooperation of humans and LLMs as evaluators, could stabilize the evaluation result. The key point of human-in-the-loop calibration is when humans should be involved in the evaluation and calibrate the evaluation result on which LLM evaluators do not perform well.

To target the "when" problem, inspired by Cai, Chang, and Han (2023), we introduce a **Balanced Position Diversity Entropy (BPDE)** score to find examples requiring auxiliary human calibration based on the evaluation results of MEC and BPC. Specifically, as shown in Figure 3(c), we first compute $2k$ evaluation results $\{\mathbf{ER}_i\}_{i=1}^{2k}$ based on the $2k$ pairs of scores.

$$\underset{1 \le i \le k}{\mathbf{ER}_i} = \begin{cases} \textbf{win}, S_{r1}^i > S_{r2}^{\prime i} \\ \textbf{tie}, S_{r1}^i = S_{r2}^{\prime i} \\ \textbf{lose}, S_{r1}^i < S_{r2}^{\prime i} \end{cases}, \underset{1 \le i \le k}{\mathbf{ER}'_i} = \begin{cases} \textbf{win}, S_{r1}^{\prime i} > S_{r2}^i \\ \textbf{tie}, S_{r1}^{\prime i} = S_{r2}^i \\ \textbf{lose}, S_{r1}^{\prime i} < S_{r2}^i \end{cases}, \tag{3}$$

and BPDE is defined as the entropy of the evaluation results:

$$\mathbf{BPDE} = \sum_{\mathbf{er} \in \{\textbf{win}, \textbf{tie}, \textbf{lose}\}} -\mathbf{p_{er}} \log \mathbf{p_{er}} \tag{4}$$

$$\mathbf{p_{er}} = \frac{\sum_{i=1}^{k} \mathbb{I}(\mathbf{ER}_i = \mathbf{er}) + \mathbb{I}(\mathbf{ER}'_i = \mathbf{er})}{2k}. \tag{5}$$

A higher BPDE score indicates that it is more likely the evaluation requires manual correction. A threshold is needed for BPDE as the hyper-parameter to select the top-β most likely biased evaluations. After selection based on the BPDE score, the annotators will evaluate the selected examples and integrate the human annotations based on the majority opinion as described in Section 4.1.

4 Experiments

4.1 Human Annotation

To assess the effectiveness of our proposed strategies, three of the authors manually annotate the "win/tie/lose" outcomes of responses from ChatGPT and Vicuna-13B independently in all 80 Vicuna Benchmark questions. All of the annotators are researchers familiar with Artificial Intelligence and are well-equipped to assess the quality of the responses. Following the same template as the original Vicuna, the annotators are instructed to assess the responses provided by Vicuna-13B and ChatGPT from four different perspectives: helpfulness, relevance, accuracy, and level of detail. The responses of Vicuna and ChatGPT are presented to the annotators in random order. The evaluation process for each example took an average of three minutes. **The final result is based on the majority opinion among three annotators.**

4.2 Experimental Setup and Metric

We use the OpenAI API to conduct our experiments ("gpt-3.5-turbo-0301" for ChatGPT, and "gpt-4" for GPT-4). For the methods that do not need to sample multiple generation results, we set the generated temperature to 0 for deterministic generation results. For the multiple evidence strategy, we set the temperature to 1 and sample three generation results ($k = 3$). We use the accuracy and kappa correlation coefficient (McHugh, 2012) with the final majority of human annotation results to measure the performance of different evaluators and evaluation methods. When calculating the results for methods that do not utilize BPC, we randomize the order of the two responses from the assistants and calculate the average results of 100 runs to ensure stable results.

4.3 Main Results

Table 4 illustrates the performance of different methods on our manually annotated 80 annotated examples. As is shown: **1)** There is a good correla-

EVALUATORS	METHODS	ACCURACY	KAPPA	COST
Human 1	-	68.8%	0.50	\$30.0
Human 2	-	76.3%	0.62	\$30.0
Human 3	-	70.0%	0.50	\$30.0
Human Average	-	71.7%	0.54	\$30.0
GPT-4	VANILLA	52.7%	0.24	\$2.00
GPT-4	EC ($k = 1$)	56.5%	0.29	\$2.00
GPT-4	MEC ($k = 3$)	58.7%	0.30	\$3.19
GPT-4	MEC ($k = 6$)	60.9%	0.33	\$6.38
GPT-4	MEC ($k = 3$) + BPC ($k = 3$)	62.5%	0.37	\$6.38
GPT-4	MEC ($k = 3$) + BPC ($k = 3$) + HITLC ($\beta = 20\%$)	73.8%	0.56	\$23.1
ChatGPT	VANILLA	44.4%	0.06	\$0.10
ChatGPT	EC ($k = 1$)	52.6%	0.23	\$0.10
ChatGPT	MEC ($k = 3$)	53.2%	0.24	\$0.17
ChatGPT	MEC ($k = 6$)	55.6%	0.27	\$0.34
ChatGPT	MEC ($k = 3$) + BPC ($k = 3$)	58.7%	0.31	\$0.34
ChatGPT	MEC ($k = 3$) + BPC ($k = 3$) + HITLC ($\beta = 20\%$)	71.3%	0.52	\$18.3

Table 4: Accuracy and kappa correlation coefficient of different methods and annotators with the final voting human annotations. The VANILLA evaluation method was commonly used in previous works, which provided the conclusion first and then followed with the explanation. (M)EC, BPC, and HITLC denote our proposed (multiple) evidence calibration, balanced position calibration, and human-in-the-loop calibration respectively. $\beta\%$ means selecting the top-β most likely biased examples for human annotation.

tion coefficient between the annotations provided by each human annotator and the final voting results. In detail, the average accuracy and the kappa correlation coefficient of human annotations are 71.7% and 0.54, respectively; **2)** Overall, GPT-4 achieves higher alignment with human judgments compared with ChatGPT, showing its powerful alignment ability with humans; **3)** Compared to the commonly used VANILLA evaluation method, our proposed automatic calibration strategies (i.e., EC, MEC and BPC) significantly enhance the alignment between GPT-4 and ChatGPT with human judgments; For instance, by employing the MEC and BPC calibration strategies, ChatGPT demonstrates a notable improvement in both accuracy and the kappa correlation coefficient. Specifically, the accuracy is improved by 14.3%, and the kappa correlation coefficient is increased from 0.06 to 0.31; **4)** "MEC ($k = 3$) + BPC ($k = 3$)" outperforms "MEC ($k = 6$)", demonstrating that LLMs are affected by positional bias, and BPC effectively ensures that LLMs serve as fair evaluators; **5)** Our proposed HITLC can effectively enhance the alignment between GPT-4 and ChatGPT with human judgments, requiring only a small amount of human labor. For example, by incorporating just 20% ($\beta = 20\%$) human assistance, ChatGPT attains comparable Human Average accuracy, while reducing the annotation cost from \$30 to \$18.3, a 39% reduction.[1]

In conclusion, our proposed calibration methods are simple yet very effective in improving the evaluation performance with LLM as evaluators, while maintaining low costs.

5 Analysis

5.1 Ablation on Evidence Number *k* and Temperature *t*

In the MEC and BPC strategy, we sample k evaluation results for each query prompt and ensemble them to enhance the evaluation process. We conduct an analysis to examine the influence of the number of evidence k, on the model's evaluation performance. As illustrated in Figure 4(a), we compared the performance of ChatGPT with different values of k, namely 1, 3, 5, and 7. The

[1] The minimum hourly wage in the United States is near \$7.5, which can be found at https://www.worker.gov/. On average, annotating an example takes 3 minutes, and the Vicuna evaluation benchmark comprises 80 examples in total. Consequently, the cost per annotator amounts to \$30.