Enhancing Objectivity in LLM-as-a-Judge through Perturbation Injection

Zhihao Zhu, Haoran Liao, Yaohui Jin*
{zzh2021,liaohaoran,jinyh}@sjtu.edu.cn
MoE Key Lab of Artificial Intelligence,
Shanghai Jiao Tong University, Shanghai, China

Abstract

LLM-as-a-judge is considered a potential substitute for human evaluation due to its efficiency and cost-effectiveness. However, recent studies indicate that LLM-as-a-judge exhibits systematic biases when comparing candidate answers, including contextual, verbosity, and positional bias. These biases, as we find, mirror human cognitive biases like the anchoring effect and availability heuristic, where intuitive decisions prioritize superficial features over deeper analysis. Inspired by the Dual Process Theory, we propose that LLM evaluations often resemble system 1 thinking, leading to biased judgments. To address this, we introduce PeBC, a Perturbation-Based Calibration framework that shifts LLM evaluations from system 1 to system 2 reasoning through perturbation injection, bias analysis, and rule calibration. Our experiments on the meta-evaluation benchmarks LLMBar-Natural and LLMBar-Adversarial demonstrate that PeBC successfully mitigates evaluation biases, outperforming existing state-of-the-art (SOTA) methods across various test scenarios and achieving better alignment with human judgments.

Keywords: Large language models; Evaluation bias; Dual process theory; Perturbation injection

Introduction

Evaluating the responses generated by models or content written by humans is a challenging task. Traditional methods primarily relied on human annotation and scoring for evaluation. While this method can provide high-quality results, it is exceedingly time-consuming and costly. To address these issues, various automated evaluation techniques have been proposed. For instance, automated evaluation metrics based on n-grams, such as Rouge (Lin, 2004), BLEU (Papineni, Roukos, Ward, & Zhu, 2002), and METEOR (Banerjee & Lavie, 2005), have been widely used. However, these methods often show a weak correlation with human judgments (He et al., 2022), particularly in tasks requiring open-ended generation or domain-specific expertise.

The emergence of LLMs (Achiam et al., 2023) with strong zero-shot reasoning (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022) and instruction-following capabilities (Ouyang et al., 2022) has opened up new opportunities for using LLMs as annotators (Xu, Guo, Duan, & McAuley, 2023) and evaluators (Peng, Li, He, Galley, & Gao, 2023; Zheng et al., 2023). By inputting the content to be evaluated along with the corresponding evaluation criteria as prompts, these models can score and compare candidate responses and provide explanations. This efficient and cost-effective approach has gradually been adopted by many works (Y. Liu et al., 2023; Zhou et al., 2024; Y. Wang et al., 2023; Chen et al., 2024; Hashemi, Eisner, Rosset, Van Durme, & Kedzie, 2024; Huang et al., 2024; Verga et al., 2024). However, the sensitivity of LLMs to prompt design has raised concerns about the reliability of using LLMs as judges. For example, (Zheng et al., 2023; Zeng et al., 2023; P. Wang et al., 2023) have shown that LLMs are highly sensitive to the order and position of candidate answers during evaluation. Additionally, LLMs exhibit a tendency to favor more verbose responses over more accurate ones (Zheng et al., 2023; Saito, Wachi, Wataoka, & Akimoto, 2023). Therefore, eliminating biases in LLM evaluations and improving consistency with human preferences have emerged as critical issues (Li et al., 2023; Y. Liu et al., 2024).

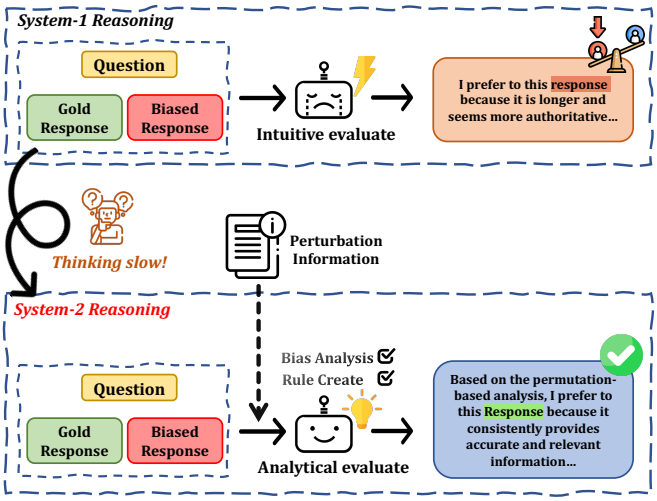

Figure 1: An overview of the transition in LLM evaluation paradigms: from system 1 (intuitive, instinctive, and prone to biases) reasoning to system 2 (analytical, thoughtful, and more reliable) reasoning, guided by perturbation-driven analysis.

In this work, we propose a novel evaluation framework, named PeBC (**Pe**rturbation-**B**ased **C**alibration), designed to mitigate biases in LLM evaluations. Inspired by the Dual Process Theory (Evans, 2008; Kahneman, 2011), which delineates two modes of thinking—system 1 (fast, intuitive) and

*Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

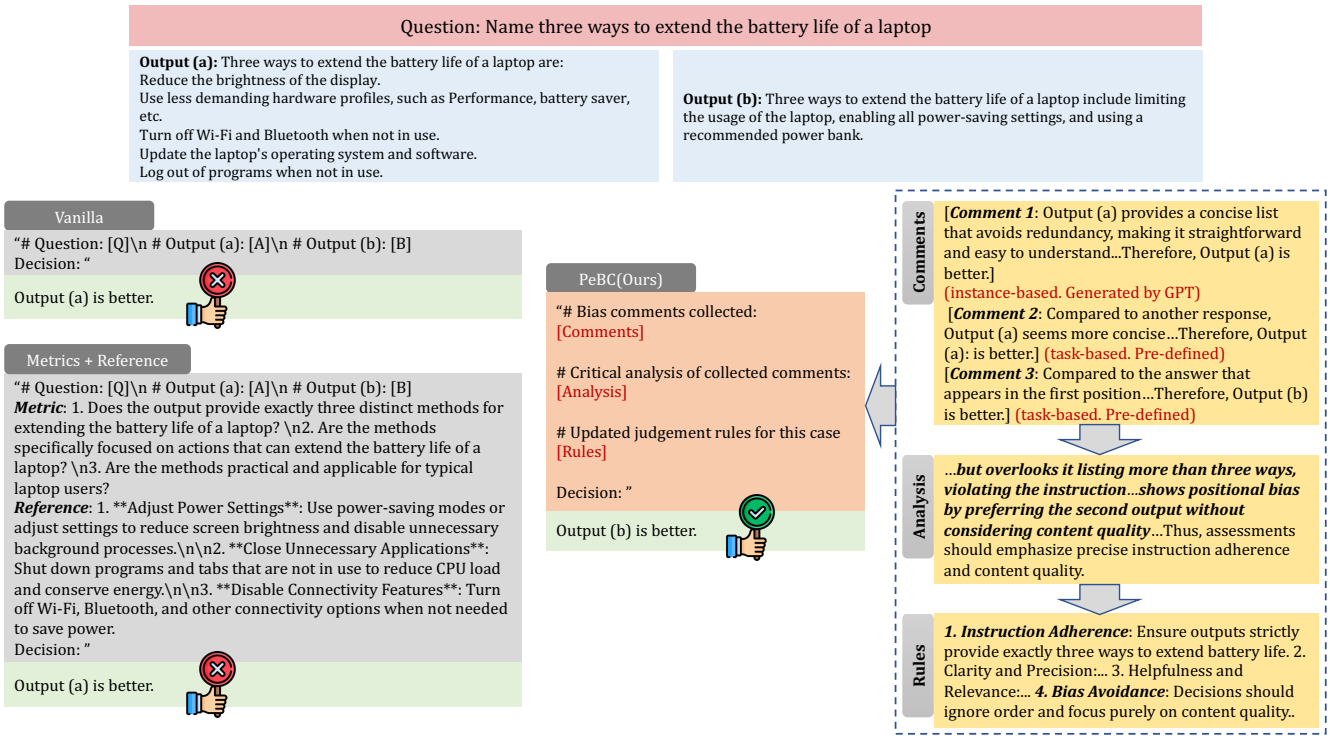

Figure 2: An example of **PeBC**. The left side compares the failed attempts of *Vanilla* and *Metrics+Reference* methods in evaluation, while the right side shows the main components of our perturbation-based calibration evaluation framework: analysis extracted from injected perturbation information and updated rules, which allow the LLM evaluator to accurately select Output (b) for adhering to the instruction of providing exactly three methods, rather than being influenced by positional or verbosity bias.

system 2 (slow, analytical)—we posit that LLM evaluations often exhibit system 1-like behavior, leading to biased judgments. PeBC addresses this limitation by shifting LLM evaluations from system 1 to system 2 thinking through the strategic injection of perturbation information (Yang, Klein, Celikyilmaz, Peng, & Tian, 2024; A. Liu et al., 2024), as shown in Figure 1. This shift enables the model to engage in more analytical and reflective decision-making, thereby reducing biases. The framework operates in three key stages: (1) Perturbation Injection, where a biased response set is generated by introducing perturbations based on the original instructions and candidate responses; (2) Bias Identification and Refinement, where the LLM evaluator conducts a preliminary assessment to identify potential biases and refines the evaluation criteria accordingly; and (3) Final Assessment, where the updated criteria and evaluation results are provided to the LLM evaluator for a final, bias-corrected judgment. By leveraging LLM's self-refinement capabilities, PeBC enhances the model's ability to detect and correct biases, resulting in more objective, consistent, and human-aligned evaluations.

In summary, the main contributions of our work are:

1. We propose PeBC, a perturbation-based calibration framework that mitigates biases in LLM evaluations by shifting from intuitive, system 1-like thinking to analytical, system 2-driven reasoning.

2. We introduce a novel zero-shot calibration strategy composed of three composable modules: Perturbation, Analysis, and Rules, which leverage LLMs' self-refinement capabilities to systematically reduce biases.

3. We demonstrate the effectiveness of PeBC through extensive experiments, including comparisons with state-of-the-art baselines and ablation studies, showcasing its robustness and alignment with human judgment.

Problem Formulation

Let an instruction I consist of a question Q and a target T. Given two candidate responses A and B, a large language model employs its evaluation strategy π to determine which response better aligns with I. Formally, the evaluation process can be expressed as:

$$R = \pi(I, A, B),$$

where $R.y \in \{A, B\}$ denotes the selected response. However, the LLM's strategy π is often influenced by systematic biases, which can be modeled as follows:

1. Positional Bias: The LLM's preference for a response based on its position (e.g., first or last) can be represented as:

$$\pi_{\text{pos}}(I,A,B) = \begin{cases} A & \text{if } A \text{ is in the preferred position,} \\ B & \text{otherwise,} \end{cases}$$

where the preferred position is a heuristic-driven choice, analogous to the anchoring effect in human cognition.

2. Verbosity Bias: The LLM's tendency to favor longer or more detailed responses, regardless of their relevance or accuracy, can be modeled as:

$$\pi_{\text{verb}}(I,A,B) = \begin{cases} A & \text{if len}(A) > \text{len}(B), \\ B & \text{otherwise,} \end{cases}$$

where $\text{len}(\cdot)$ denotes the length of a response. This behavior mirrors the availability heuristic, where salient features (e.g., response length) override content quality.

3. Contextual Bias: The LLM's inclination to prefer responses that align with specific contextual cues or keywords can be expressed as:

$$\pi_{\text{ctx}}(I,A,B) = \begin{cases} A & \text{if } A \text{ matches a preferred context } C, \\ B & \text{otherwise,} \end{cases}$$

where C represents a set of contextually preferred features.

These biases collectively indicate that the LLM's evaluation strategy π primarily relies on system 1-like reasoning, which is characterized by fast, intuitive, and heuristic-driven decision-making. This mode of thinking, while efficient, often leads to systematic errors over deeper semantic alignment with the instruction. Such behavior mirrors well-documented cognitive biases in human decision-making (Evans, 1989). Formally, the biased evaluation can be represented as:

$$\pi(I,A,B) = \pi_{\text{pos}}(I,A,B) \oplus \pi_{\text{verb}}(I,A,B) \oplus \pi_{\text{ctx}}(I,A,B),$$

where $\oplus$ denotes a combination of biases that may interact in complex ways.

To address these limitations, we propose shifting the LLM's evaluation process toward system 2 thinking, which is slow, analytical, and reflective. In human cognition, system 2 enables individuals to override intuitive biases by engaging in deliberate reasoning and self-correction. Similarly, the LLM can be guided to systematically identify and correct its biases, moving beyond heuristic-driven evaluations to achieve more accurate and human-aligned judgments π'.

Proposed Method

In this section, we elaborate on the main components of PeBC, including three distinct calibration strategies: Perturbation, Analysis, and Rules. To provide a clearer understanding of the calibration strategies, we also present the overall prompting design in Table 1.

Perturbation Information Injection

To address the biases in LLM evaluations, we introduce perturbation information to trigger the model's awareness of biased content. Specifically, we generate a response set containing perturbation information, becoming part of the evaluation input for the LLM evaluators. Here, we focus on three common biases: *positional bias, verbosity bias, and contextual bias*.

Table 1: The prompt template of PeBC. ({Question}, {O_a}, and {O_b}) are from **LLMBar**. {Comment}, {Analysis} and {Rules} are generated by LLM evaluators.

[System] You are a helpful assistant in evaluating the quality of the outputs for a given instruction. Your goal is to select the best output for the given instruction.

[User] Select the Output (a) or Output (b) that is better for the given instruction. The two outputs are generated by two different AI chatbots respectively.
Instruction:
{Question}
Output (a):
{O_a}
Output (b):
{O_b}

Bias and error comments collected:(Avoid biased judgment)
$\pi_{\text{pos}}(I,A,B) \rightarrow$ {*Bias_Comment_1*}
$\pi_{\text{verb}}(I,A,B) \rightarrow$ {*Bias_Comment_2*}
$\pi_{\text{ctx}}(I,A,B) \rightarrow$ {*Bias_Comment_3*}

Critical analysis of collected comments:(critically analyze the biased comments above, and explain how you would avoid similar biases to make a correct and unbiased assessment. Do not contain your final evaluation result here)
{Analysis}

Updated judgement rules for this case:(Do not contain your final evaluation result here)
{Rules}

Decision(Give a brief explanation of your evaluation followed by either "Therefore, Output (a) is better." or "Therefore, Output (b) is better." verbatim. Always claim which is better at the end. In your explanation, you should always use "Output (a)" or "Output (b)" to refer to the two outputs respectively.):

When generating perturbation responses for different types of biases, we adopt two distinct and complementary perturbation injection strategies: task-based and instance-based. For

positional bias and verbosity bias, we design general response templates (as shown in Figure 2). For contextual bias, we pre-generate perturbation responses using an advanced LLM (GPT-4o) based on specific instances. We controlled the parameter (temperature=0.7 and top-p=0.9) to balance response diversity and semantic relevance. Given that it is a highly aligned model with human preferences, directly generating biased responses is impractical. Instead, we introduce contextual bias by randomly shuffling the pros and cons generated by GPT-4o.

Bias Analysis and Extraction

Existing works (Madaan et al., 2024; Yao et al., 2022; Wei et al., 2022) have shown that LLMs significantly improve response quality by generating concise analyses and reasoning before producing final answers. Based on this finding, `PeBC` requires the LLM evaluator to conduct a detailed analysis of the content of biased comments present and how to avoid them. This "evaluation experience" serves as the basis for subsequent evaluations, enhancing the model's ability to perceive and correct biases.

Calibration of Assessment Rules

To achieve more consistent and fair outputs, we calibrate the assessment criteria based on the results of the previous analysis steps. Specifically, the LLM judge dynamically updates the evaluation rules according to the identified biases and performs the final assessment using the refined rules. This calibration process aligns the model's outputs with unbiased judgment standards, enhancing the reliability of LLM-as-a-judge evaluations.

Experiments

We evaluated `PeBC` on the meta-evaluation benchmark LLMBar, which consists of two components: the Natural set and the Adversarial set, representing typical evaluation scenarios and more complex, adversarial challenges, respectively.

Implementation Details

We utilized GPT-4o and ChatGPT* as the LLM evaluator in `PeBC`, which is among the most advanced proprietary models available. During the experiment, we set the temperature to 0 to ensure reproducibility. We also utilized GPT-4o to generate perturbation responses with contextual bias, ensuring that the responses closely align with the provided instructions.

Benchmark

In our experiment, we employ the LLMBar (Zeng et al., 2023) to assess how well LLM evaluators can mitigate biased outputs and align with human judgments. The LLMBar benchmark comprises two sets: Natural, which includes 100 instances from existing datasets to reflect performance in routine evaluation scenarios, and Adversarial includes 319 instances, categorized into NEIGHBOR, GPTINST, GPTOUT, and MANUAL. It is designed with deceptive adversarial instances to test the judgment of LLM evaluators under complex conditions. The adversarial set also contains instances that may seem superficially appealing, such as those with persuasive language or longer text lengths, but which deviate from correct instruction execution.

Baselines

We compare `PeBC` with state-of-the-art and widely applied evaluation strategies†:

1. **Vanilla** (Dubois et al., 2024): The LLM judge selects the better output between two options without any explanation.
2. **Vanilla+rule**: This strategy enhances the Vanilla method with predefined evaluation rules.
3. **Chain-of-Thoughts** (Wei et al., 2022): The LLM judge generates concise reasoning before choosing a preferred output among options.
4. **Metrics+Reference** (Zeng et al., 2023): The LLM uses instruction-specific metrics it generates to guide comparisons, integrating self-generated reference responses.

Main Results

The main results of our study are presented in Tables 2 and 3. Table 2 reports the results of different methods using GPT-4o-based evaluators, while Table 3 presents the results using GPT-3.5-based evaluators. The results demonstrate the effectiveness of `PeBC` in enhancing the performance of evaluators across various settings.

For GPT-4o-based evaluators, `PeBC` outperforms the other baseline methods. In the Adversarial setting, we find that `PeBC(Pert+Ana+Rules)` achieves the highest average accuracy of 89.3%, which is 1.2% higher than the current best strategy. In MANU, where adversarially challenging instances are manually constructed, `PeBC(Pert+Ana+Rules)` stands out with an accuracy of 91.3%, a 6.5% improvement. This significant gain can be attributed to our method's ability to systematically identify and correct biases. The perturbation injection and rule calibration steps enable the model to focus on semantic alignment, helping it navigate complex instruction patterns more effectively, rather than relying on superficial cues like verbosity or position.

For GPT-3.5-based evaluators, `PeBC` demonstrates significant improvements. `PeBC(Pert+Ana+Rules)` achieves an average accuracy of 58.2%, which is 18.2% higher than the Vanilla method and 9.4% higher than the Metrics+Reference strategy, with improvements observed across all datasets. This substantial improvement underscores the effectiveness of `PeBC` in mitigating biases that are especially prevalent in less capable models. Weaker evaluators, such as GPT-3.5,

*By default, we use gpt-4o-2024-08-06 and gpt-3.5-turbo-0613 for GPT-4o and ChatGPT respectively.

†The implementation of these four baselines primarily follows (Zeng et al., 2023). All strategies, including `PeBC`, are configured in a zero-shot prompting.

Table 2: Results of GPT-4o-based evaluators on baselines and PeBC. The highest average accuracy is marked by **bold**. Random guess would achieve an Acc. of 50%. Abbreviations: NAT: LLMBar-Natural, ADV: LLMBar-Adversarial, NEI: NEIGHBOR, GPTI: GPTINST, GPTO: GPTOUT, MANU: MANUAL, Avg: Average.

Method	NAT	ADV					Avg
	Acc.(↑)	NEI(↑)	GPTI(↑)	GPTO(↑)	MANU(↑)	Avg(↑)	Acc.(↑)
Vanilla	92.0	69.4	84.8	74.5	76.1	75.5	79.5
Vanilla+rule	95.0	77.6	89.1	72.3	76.1	79.9	83.5
COT	96.0	83.6	91.3	72.3	73.9	82.8	85.9
Metrics+Reference	97.0	**84.3**	91.3	76.6	84.8	85.3	88.1
PeBC(Pert)	**98.0**(+1.0)	**84.3**(+0.0)	91.3(+0.0)	**78.7**(+2.1)	84.8(+0.0)	85.9(+0.6)	88.5(+0.4)
PeBC(Pert+Ana)	97.0(+0.0)	82.8(-1.5)	92.4(+1.1)	**78.7**(+2.1)	84.8(+0.0)	85.3(+0.0)	88.1(+0.0)
PeBC(Pert+Ana+Rules)	97.0(+0.0)	83.6(-0.7)	**93.5**(+2.2)	**78.7**(+2.1)	**91.3**(+6.5)	**86.8**(+1.5)	**89.3**(+1.2)

Table 3: Accuracy of GPT-3.5-based evaluators on baselines and PeBC. The highest average accuracy is marked by **bold**. *Baseline results are reported from the original paper (Zeng et al., 2023), with consistent model version and settings.

Method	NAT	ADV					Avg
	Acc.(↑)	NEI(↑)	GPTI(↑)	GPTO(↑)	MANU(↑)	Avg(↑)	Acc.(↑)
Vanilla*	29.3	43.6	37.0	17.9	27.7	79.0	40.0
Vanilla+rule	35.9	38.3	43.5	38.1	38.3	82.5	48.8
Metrics+Reference*	31.5	46.8	34.8	39.6	37.6	79.0	47.5
PeBC(Pert)	44.6(+8.7)	59.6(+12.8)	47.8(+4.3)	40.3(+0.7)	45.5(+7.2)	82.0(-0.5)	54.2(+5.4)
PeBC(Pert+Ana)	46.7(+10.8)	**61.7**(+14.9)	**54.3**(+10.8)	41.8(+2.2)	48.0(+9.7)	82.0(-0.5)	56.1(+7.3)
PeBC(Pert+Ana+Rules)	**53.3**(+17.4)	59.6(+12.8)	47.8(+4.3)	**44.8**(+5.2)	**49.8**(+11.5)	**85.0**(+2.5)	**58.2**(+9.4)

often rely on fast, intuitive, and heuristic-driven decision-making, leading to biases like verbosity and positional bias. PeBC enables the model to systematically identify and correct these biases, resulting in more accurate and reliable evaluations.

Overall, PeBC shows superior performance across both strong and weaker evaluators, underscoring its robustness and generalizability. We speculate that the introduction of perturbation information and bias-aware evaluations effectively mitigates the bias of weaker judges, leading to significant performance enhancements.

Ablation Study

Alignment with Human Evaluation: To assess the alignment of PeBC with human evaluation standards, we analyzed the kappa correlation coefficient (Kap) results(use GPT-4o as the evaluator). The experimental result, as illustrated in Figure 3, indicates a strong agreement between PeBC and human judgments. Specifically, on the ADVERSARIAL dataset, (PeBC(Pert+Ana+Rules)) achieved a kappa value of 0.738, outperforming the SOTA method by 0.093, demonstrating stronger alignment with human evaluators. This suggests that PeBC not only improves evaluation accuracy but also aligns closely with human standards, particularly in complex and adversarial scenarios.

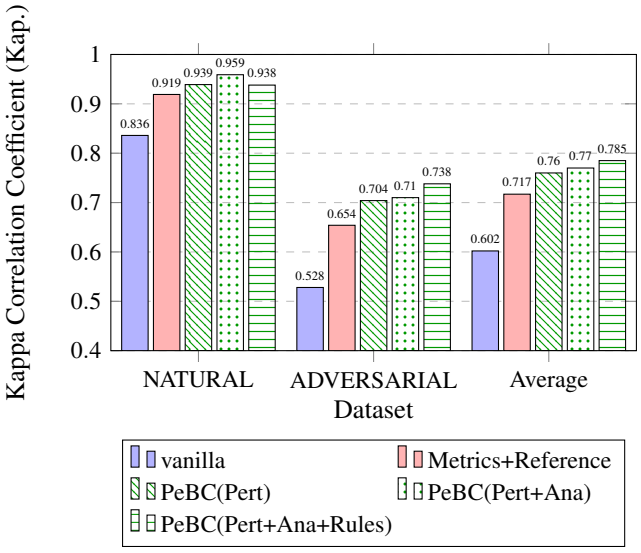

Figure 3: GPT-4o results on NATURAL and ADVERSARIAL datasets using the Kappa Correlation Coefficient (Kap.), which reflects consistency with human evaluation.

Effectiveness in Bias Reduction: To evaluate the effectiveness of PeBC in reducing biases, we randomly selecte 50 cases from the LLMBar-Adversarial dataset and have three professional annotators assess the evaluation results(use GPT-4o as

the evaluator). The annotators evaluate the presence of content bias, position bias, and verbosity bias in the results generated by different methods.

As presented in Table 4, PeBC achieved the lowest error rates across all bias categories. Specifically, we noted that the instance-specific perturbation information generated by PeBC significantly mitigated the impact of positional and verbosity biases and overall result highlights the robustness of PeBC in addressing common evaluation biases and aligning with objective standards.

Table 4: Error analysis of bias reduction across different methods on 50 cases. The values represent the average number of bias instances, along with the standard deviation.

Method	Content Bias(↓)	Position Bias(↓)	Verbosity Bias(↓)
Vanilla	9±0.82	12.33±1.37	11.67±1.37
Metrics+Reference	5.33±1.16	11.33±1.37	6±0.82
PeBC	**5±1**	**8±2**	**4±0.82**

Case Study

Continuing from the previous error analysis, we observed clear reduction in biases after applying PeBC. In this case, we use different methods to evaluate two outputs for the instruction: *"What is the typical wattage of a bulb in a lightbox?"*

- *Output 1*: The wattage of a bulb in a lightbox is typically measured in watts, which refers to the amount of power consumed by the bulb and determines its brightness level. The wattage can vary depending on the size, purpose, and specific design of the lightbox.
- *Output 2*: Its wattage typically ranges from 5 watts to 100 watts.

The outputs differ in content: one provides a more precise numerical range, while the other includes a lot of extra but unnecessary information. Other methods tend to choose the **detailed answer** as the gold standard. However, by leveraging instance-specific perturbation information, PeBC accurately assesses and prioritizes the more concise and direct response, effectively mitigating the bias towards unnecessary verbosity and ensuring that the evaluation remains both relevant and free from content redundancy.

Related works

LLM-as-a-Judge and Evaluation Biases

The use of large language models (LLMs) as evaluators, or "LLM-as-a-judge," has gained significant traction due to their efficiency and cost-effectiveness in automating tasks such as text evaluation, ranking, and scoring (Xu et al., 2023; Peng et al., 2023). However, recent studies reveal that LLMs exhibit systematic biases during evaluation, including positional bias, verbosity bias, and contextual bias (Zheng et al., 2023; Saito et al., 2023). These biases often lead to inconsistencies with human judgment, raising concerns about the reliability of LLM-based evaluations. For instance, positional bias—where LLMs favor responses based on their order—parallels the anchoring effect in human cognitive biases (Lieder, Griffiths, M. Huys, & Goodman, 2018), while verbosity bias—where LLMs prefer longer or more detailed responses—resembles the availability heuristic (Ehrlinger, Readinger, & Kim, 2016). To address these issues, recent efforts have proposed rule-based prompting (Zeng et al., 2023) and reference-guided metrics (P. Wang et al., 2023; Chan et al., 2023), though these approaches often lack a systematic framework for bias identification and correction. Inspired by the parallels between LLM biases and human cognitive biases, we propose PeBC, a perturbation-based calibration framework that leverages perturbation injection to help LLMs recognize and correct biases.

From Intuitive to Analytical Evaluation

Recent advancements in prompting techniques, such as chain-of-thought (COT) (Wei et al., 2022) and self-reflection (Madaan et al., 2024), have enabled LLMs to transition from intuitive, heuristic-based decision-making to more analytical reasoning. These highlight the potential for LLMs to exhibit behaviors analogous to slow, deliberate thinking. Building on this, (Chan et al., 2023) use multi-agent debate to enhance evaluation robustness, while (P. Wang et al., 2023) proposes "Human-in-the-Loop Calibration" to mitigate position bias. However, these methods often address biases in isolation and lack a unified framework for handling multiple biases effectively. In contrast, our work is the first to systematically address LLM evaluation biases from a cognitive bias perspective, drawing parallels between LLM biases (e.g., positional, verbosity, and contextual biases) and human cognitive biases (e.g., anchoring effect and availability heuristic). By introducing PeBC, we propose a comprehensive framework that leverages LLMs' self-refinement capabilities to shift evaluations from intuitive to analytical reasoning, ensuring greater objectivity and consistency in LLM-as-a-judge evaluations.

Conclusion

In this work, we addressed the issue of systematic biases in LLM-as-a-judge scenarios, where LLMs often exhibit various biases such as contextual, verbosity, and positional biases. Inspired by the well-established Dual Process Theory, we introduced the Perturbation-Based Calibration (PeBC) framework to shift LLM evaluations from intuitive, system 1-like reasoning to more analytical, system 2-driven thinking. By injecting perturbation information, conducting bias analysis, and calibrating assessment rules, PeBC enables LLMs to engage in reflective decision-making, significantly reducing biases and improving alignment with human judgment. Our experiments on the LLMBar benchmarks demonstrated the effectiveness of PeBC in enhancing evaluation accuracy and robustness, particularly in adversarial contexts. This work makes a valuable contribution to the development of fairer and more reliable automated evaluation systems, paving the way for more objective LLM-as-a-judge applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 72432007 and the National Key R&D Program of China under Grant No. 2022YFC3302900.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., ... Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Chen, L., Li, B., Zheng, L., Wang, H., Meng, Z., Shi, R., ... others (2024). What factors influence llms' judgments? a case study on question answering. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 17473–17485).

Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., ... Hashimoto, T. B. (2024). Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, *36*.

Ehrlinger, J., Readinger, W. O., & Kim, B. (2016). Decision-making and cognitive biases. *Encyclopedia of mental health*, *12*(3), 83–87.

Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*(1), 255–278.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C. (2024). Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 13806–13834).

He, T., Zhang, J., Wang, T., Kumar, S., Cho, K., Glass, J., & Tsvetkov, Y. (2022). On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*.

Huang, H., Qu, Y., Zhou, H., Liu, J., Yang, M., Xu, B., & Zhao, T. (2024). *On the limitations of fine-tuned judge models for llm evaluation.* Retrieved from https://arxiv.org/abs/2403.02839

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213.

Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C., & Liu, Y. (2023). Split and merge: Aligning position biases in large language model based evaluators. *arXiv preprint arXiv:2310.01432*.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, *25*, 322–349.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Liu, A., Bai, H., Lu, Z., Kong, X., Wang, X., Shan, J., ... Wen, L. (2024, August). Direct large language model alignment through self-rewarding contrastive prompt distillation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9688–9712). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.523

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulic, I., Korhonen, A., & Collier, N. (2024). Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... others (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., ... Lewis, P. (2024). Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., ... Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., ... others (2023). Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv*

preprint arXiv:2306.05087.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Xu, C., Guo, D., Duan, N., & McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Yang, K., Klein, D., Celikyilmaz, A., Peng, N., & Tian, Y. (2024). RLCD: Reinforcement learning from contrastive distillation for LM alignment. In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=v3XXtxWKi6

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., & Chen, D. (2023). Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … others (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*, 46595–46623.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., … others (2024). Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, *36*.