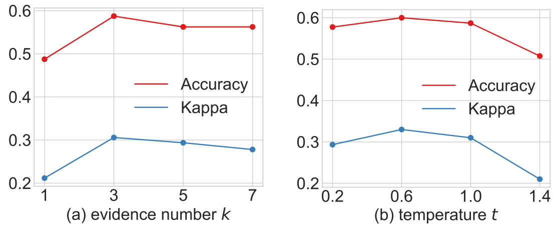

Figure 4: Variation of accuracy and kappa coefficient with a different number of evidence k and sampling temperature t when ChatGPT is used as the evaluator.

model's performance increases and then tends to be constant or decreases slightly as k becomes larger. Despite the slight decrease, the enhancement of the model effect by the MCE strategy is still significant, illustrating the stability of the MEC strategy. Consequently, we found that a k value of 3 yields an optimal performance. With this value, the model achieves a notable level of performance while keeping the API cost relatively low.

We further investigate the impact of sampling temperature t on evaluation performance. Figure 4(b) illustrates that both low temperature (i.e., 0.2) and high temperature (i.e., 1.4) result in sub-optimal evaluation alignment. We believe that low temperature eliminates the randomness of sampling, weakening the effect of MEC, while high temperature compromises the quality of generation results, leading to poor performance. Hence, it is crucial to select an appropriate temperature (e.g., 0.6 or 1.0 in our experiments) for the LLM evaluators.

5.2 Effectiveness of the BPDE

Our HITLC strategy utilizes BPDE score to select examples for human annotations. In order to analyze the efficiency of BPDE score, we compare BPDE with two typical baselines, *Random* and *Vanilla Diversity Entropy*, where Random denotes randomly select examples for human annotations, and Vanilla Diversity Entropy is calculated by using only the evaluation results of one position without swapping the position of two responses. To ensure fairness, the total number of evaluation results is 6 for both BPDE and Vanilla Diversity Entropy. As shown in Figure 5: **1)** Two Diversity Entropy methods outperform Random, showing the effectiveness of selecting examples based on the diversity entropy; **2)** BPDE outperforms Vanilla DE, which shows LLMs are sensitive to position exchange, and the results of BPC can significantly improve the performance of HITLC compared to relying solely on the results of MEC.

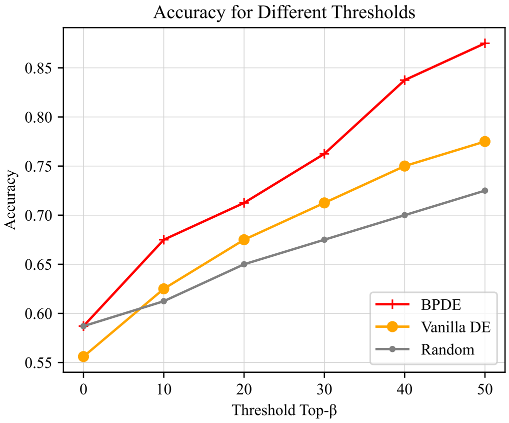

Figure 5: The accuracy of various methods changes with different human assistant thresholds (Top-β) when ChatGPT is used as the evaluator.

TEMPLATES	METHODS	ACC.	KAP.	C.R
SCORING	VANILLA	44.4%	0.06	82.5%
SCORING	MEC	53.2%	0.24	35.0%
SCORING	MEC + BPC	58.7%	0.31	N/A
COMPARING	VANILLA	50.2%	0.18	50.0%
COMPARING	MEC	54.8%	0.27	42.5%
COMPARING	MEC + BPC	60.3%	0.35	N/A

Table 5: Effectiveness of our proposed two automatic calibrated methods on two different evaluation templates with ChatGPT as the evaluator. ACC., KAP. and C.R are short for Accuracy, Kappa correlation coefficient, and Conflict Rate, respectively. N/A means the Conflict Rate is not valid for BPC methods.

5.3 Generalization on the Pairwise Comparison Evaluation Template

To provide a more comprehensive validation of our proposed calibration methods, in addition to the previous SCORING evaluation template that rates each response, we extend our analysis to incorporate the COMPARING evaluation template. This template facilitates a direct comparison between two responses, eschewing explicit scores in its assessment. Specifically, we prompt LLMs to produce results labeled as "Assistant 1", "Assistant 2", or "Same", indicating whether the response from Assistant 1 is better, worse, or equal to that of Assistant 2. As is shown in Table 5: **1)** Our proposed methods are applicable to both of these templates, leading to enhanced accuracy and a heightened correlation coefficient for ChatGPT; **2)** The significant

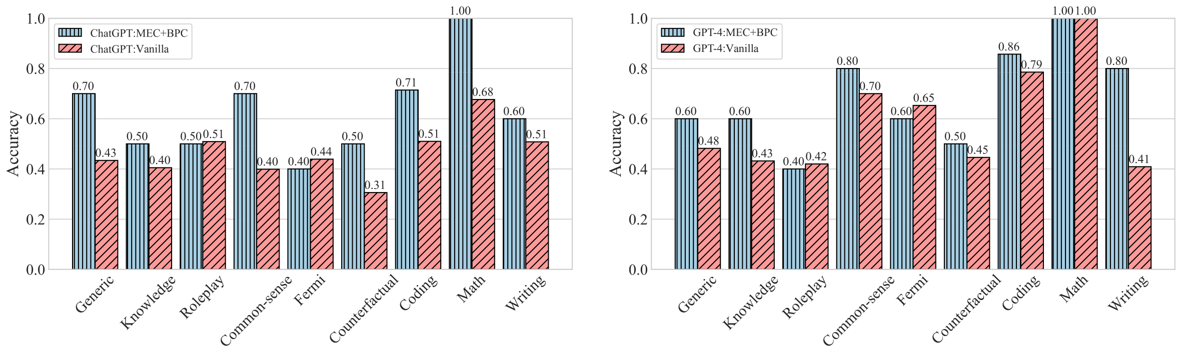

Figure 6: Fine-grained analysis of evaluation quality. Our MEC and BPC improve the evaluation performance of ChatGPT and GPT-4 in nearly all categories. Especially on the complex task categories such as common sense, coding, and math for ChatGPT.

performance gap (nearly 6% accuracy) between the VANILLA method of two templates, coupled with the high conflict rate, highlights the sensitivity and unreliability of LLMs. However, our methods effectively narrow this performance gap and reduce conflict, showcasing how calibration enhances LLM robustness.

5.4 Fine-Grained Analysis of Evaluation Quality

In order to further analyze the evaluation capabilities of the model, we perform a fine-grained analysis of the questions by dividing them into 9 categories following Zheng et al. (2023). We calculate the performance of different evaluators within these categories. As shown in Figure 6, we find that: **1)** In certain complex tasks such as common-sense, coding and math, GPT-4 performs significantly better than ChatGPT, highlighting the strength of GPT-4 as a more fair evaluator in these scenarios; **2)** Our proposed MEC+BPC strategy demonstrates noticeable improvement in evaluating ChatGPT's performance on complex tasks, allowing us to obtain satisfactory evaluation results with a low API cost.

6 Related Work

6.1 Evaluation of Large Language Models

LLMs have demonstrated powerful general generation capabilities, becoming universal assistants (OpenAI, 2022, 2023; Song et al., 2023b). With the rapid advancement of LLMs, it becomes crucial to evaluate their ability to follow human instructions. Traditional evaluation methods assess the ability by calculating a metric, like BLEU, ROUGE, BERTScore, or BARTScore, to compare the generated response with a reference response. However, these metrics do not adequately measure the alignment of the generated response with human intent (He et al., 2023). While human evaluation is treated as the most accurate measurement of model performance, it is costly and time-consuming to operate at scales. Considering the potent capabilities of LLMs, researchers have started utilizing LLMs to evaluate the proficiency of generative models in adhering to human instructions (Zheng et al., 2023; Lu et al., 2023; Li et al., 2023). In these works, Vicuna's evaluation paradigm (Zheng et al., 2023) is widely adopted, where it provides a question and two responses from two models, and uses GPT-4 to determine which response has better quality.

6.2 Bias of Deep Neural Networks

Deep Neural Networks have been proven to easily learn biases from the data, which significantly impacts their reliability. Specifically, bias has also been investigated in natural language inference (Gururangan et al., 2018; McCoy, Pavlick, and Linzen, 2019; Belinkov et al., 2019; Liu et al., 2020a,b), question answering (Min et al., 2019), ROC story cloze (Cai, Tu, and Gimpel, 2017; Schwartz et al., 2017), lexical inference (Levy et al., 2015), visual question answering (Goyal et al., 2017), information extraction (Wang et al., 2021, 2022; Song et al., 2023a; Xia et al., 2023) and so on. LLMs are pre-trained using a vast amount of data from the internet, making it highly likely for them to learn biases present in those materials. Although the LLMs are already widely adopted as a proxy of human evaluators, the reliability of this paradigm is not well explored. In this paper, we critically examine the LLMs-as-evaluator paradigm and uncover a significant positional bias. Furthermore, we propose three simple yet effective methods to calibrate the positional bias to achieve reliable and fair evaluation results.

7 Conclusion

In this paper, we reveal a systematic positional bias in evaluation with advanced ChatGPT/GPT-4 models: by manipulating the order of candidate responses during evaluation, the quality ranking results can be significantly influenced. To this end, we introduce three effective strategies, namely Multiple Evidence Calibration (MEC), Balanced Position Calibration (BPC), and Human-in-the-Loop Calibration (HITLC). MEC requires the LLM evaluator to first provide multiple evaluation evidence to support their subsequent ratings and BPC aggregates the results from various orders to determine the final score. Based on the results of MEC and BPC, HITLC further calculates a balanced position diversity entropy to select examples for human annotations. These strategies successfully reduce the evaluation bias and improve alignment with human judgments. We provide our code and human annotations to support future studies and enhance the evaluation of generative models.

References

Belinkov, Y.; Poliak, A.; Shieber, S.; Van Durme, B.; and Rush, A. 2019. Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).*

Bowman, S. R. 2023. Eight things to know about large language models. *arXiv preprint arXiv:2304.00612.*

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Cai, Z.; Chang, B.; and Han, W. 2023. Human-in-the-Loop through Chain-of-Thought. *arXiv preprint arXiv:2306.07932.*

Cai, Z.; Tu, L.; and Gimpel, K. 2017. Pay Attention to the Ending:Strong Neural Baselines for the ROC Story Cloze Task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).*

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N. M.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B. C.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; García, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Díaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K. S.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2022. PaLM: Scaling Language Modeling with Pathways. *ArXiv*, abs/2204.02311.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A Survey for In-context Learning. *arXiv preprint arXiv:2301.00234.*

Dubois, Y.; Li, X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387.*

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; Li, H.; and Qiao, Y. J. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *ArXiv*, abs/2304.15010.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics.*

He, T.; Zhang, J.; Wang, T.; Kumar, S.; Cho, K.; Glass, J.; and Tsvetkov, Y. 2023. On the Blind Spots of Model-Based Evaluation Metrics for Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12067–12097. Toronto, Canada: Association for Computational Linguistics.