in accuracy compared to its original performance, whereas GLM-4 demonstrated a more substantial increase of 7%.

5 DISCUSSION

Explicit and implicit influence of bias. We identified two distinct types of biases: explicit and implicit. Explicit biases are those where the LLM clearly states its preference for certain attributes in its decision-making process. Implicit biases are influences that affect judgments without being directly acknowledged in their reasoning. Our case studies illustrate these biases in Appendix E. The Authority bias exemplifies an explicit bias, where the LLM openly favored answers containing citations, even when these were fake. This demonstrates a clear preference for responses that appear scholarly, regardless of their actual validity. Conversely, the refinement-aware bias represents an implicit bias. Here, the LLM consistently scored refined answers higher, despite providing similar justifications for different instances and never explicitly mentioning refinement as a factor in its decision-making process. The findings indicate that LLMs are influenced by various factors. The disparity between their internal processing and expressed reasoning underscores the importance of conducting more research into the nature of LLM bias. It is essential to comprehend these biases to enhance the trustworthiness and reliability of LLM-as-a-Judge.

Table 6: Bias recognition performance across different bias types. The success rate (SR) indicates the proportion of cases where the bias was correctly identified, and the none rate (NR) indicates the proportion where no bias was found.

Bias Type	GPT-4-Turbo		Claude-3.5	
	SR↑	NR↓	SR↑	NR↓
Authority	0.84	0.14	0.84	0.00
Bandwagon-effect	1.00	0.00	0.92	0.00
Compassion-fade	0.48	0.34	0.96	0.00
Distraction	1.00	0.00	1.00	0.00
Diversity	0.46	0.02	0.96	0.00
Fallacy-oversight	0.52	0.04	0.46	0.00
Sentiment	0.96	0.04	0.72	0.00
Verbosity	0.90	0.10	1.00	0.00

Suggestions for application. In discussing potential strategies to mitigate biases in LLM-as-a-Judge, we propose the following recommendations aimed at enhancing the fairness of models while mitigating bias interference:

- ▷ **Carefully construct prompts and implement advanced reasoning strategies.** We recommend creating prompts that include specific protective phrases to guard against various types of biases, such as instructing the model to disregard the identity information of the person being evaluated. Additionally, implementing advanced reasoning strategies similar to CoT can guide the model through a step-by-step decision-making process.
- ▷ **Establish prompt injection safeguards.** We recommend instituting protective measures against prompt injection related to the bias types discussed in this paper. These safeguards can prevent models from being influenced by biased information embedded in prompts. By implementing such protective measures, we can enhance the fairness of LLM-as-a-Judge, ensuring that the judging process is not compromised by external attempts to introduce bias.
- ▷ **Implement bias detection mechanisms.** Based on our experimental findings, we suggest implementing a simple, prompt-based bias detection mechanism similar to the one we developed in Figure 32. This approach can proactively identify potential biases in judging templates before the actual judging process begins. As presented in Table 6, our results demonstrate that while the effectiveness varies across different bias types, this method shows promise in uncovering a majority of biases.

6 CONCLUSION

This paper presents CALM, an automated evaluation framework for assessing potential bias when LLMs are employed as judges in various application scenarios. CALM provides a comprehensive examination of 12 types of biases and utilizes an automated bias injection and qualification method, resulting in an objective and scalable evaluation approach. Our experiments show that while models may reliably serve as judges for specific tasks, there remains significant room for improvement in the broader use of LLMs as judges. CALM could be used to evaluate future, more advanced LLM-based judge solutions, ensuring they meet higher standards of bias mitigation.

Ethical Consideration

It is significant to emphasize that some of the question sets and bias-related responses may contain NSFW content. While we have manually reviewed and curated this data to ensure its appropriateness for research purposes, we urge readers and potential users of our findings to exercise caution and discretion. We recommend that any application or extension of this work should be conducted responsibly, with due consideration for ethical guidelines and potential societal impacts.

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

AI@Mistral. Mistral 7b: The best 7b model to date, apache 2.0, 2023. URL https://mistral.ai/news/announcing-mistral-7b/.

AI@Mistral. Cheaper, better, faster, stronger, 2024. URL https://mistral.ai/news/mixtral-8x22b/.

Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. URL https://arxiv.org/abs/2302.04023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024a. URL https://arxiv.org/abs/2402.04788.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*, 2024b.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement biases, 2024c. URL https://arxiv.org/abs/2402.10669.

Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*, 2024d.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. Scholarchemqa: Unveiling the power of language models in chemical research question answering, 2024e. URL https://arxiv.org/abs/2407.16931.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Jon Durbin. Truthy-dpo-v0.1. https://huggingface.co/datasets/jondurbin/truthy-dpo-v0.1, 2023. Accessed: 2024-07-15.

Jon Durbin. Py-dpo-v0.1. https://huggingface.co/datasets/jondurbin/py-dpo-v0.1, 2024. Accessed: 2024-07-15.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*, November 2023. ISSN 1396-0466. doi: 10.5210/fm.v28i11.13346. URL http://dx.doi.org/10.5210/fm.v28i11.13346.

Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang Zhang, and Lichao Sun. The best of both worlds: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Taicheng Guo, kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 59662–59688. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/bbb330189ce02be00cf7346167028ab1-Paper-Datasets_and_Benchmarks.pdf.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL https://arxiv.org/abs/2402.01680.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Hui Huang, Yingqi Qu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. On the limitations of fine-tuned judge models for llm evaluation, 2024a. URL https://arxiv.org/abs/2403.02839.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, et al. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*, 2023a.

Yue Huang, Qihui Zhang, Lichao Sun, et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023b.

Yue Huang, Jingyu Tang, Dongping Chen, Bingda Tang, Yao Wan, Lichao Sun, and Xiangliang Zhang. Obscureprompt: Jailbreaking large language models via obscure input. *arXiv preprint arXiv:2406.13662*, 2024b.

Intel. Orca-dpo-pairs. https://huggingface.co/datasets/Intel/orca_dpo_pairs, 2023. Accessed: 2024-07-15.