GeWu: A Culturally-Grounded Chinese Benchmark for Multi-Stage Social Bias Evaluation in Large Language Models

Yi Lin[1], Ziyi Zhou[1], Jiashi Gao[1], Xinwei Guo[1], Jiaxin Zhang[1], Haiyan Wu[2], Xin Yao[3], Xuetao Wei[1]*

[1]Southern University of Science and Technology, Shenzhen, China
[2]University of Macau, Macau, China
[3]Lingnan University, Hong Kong, China
12432682@mail.sustech.edu.cn, weixt@sustech.edu.cn

Abstract

With the rapid deployment of Chinese large language models (LLMs), culturally-grounded bias evaluation remains understudied due to the dominance of English benchmarks and simplistic Chinese scenarios. To address this, we propose **GeWu**, a comprehensive benchmark featuring a culturally-aware dataset of 60,192 questions spanning 14 social groups with fine-grained Chinese contexts, significantly exceeding existing resources in breadth and depth. Our two-stage evaluation first quantifies bias via multiple-choice questions using a novel probability-based scoring mechanism to sensitively capture bias tendencies, distilling high-bias scenarios into `GeWu-1K`. This refined subset then enables multi-turn dialogue evaluations for in-depth analysis under realistic conditions. Experiments reveal that `GeWu` effectively exposes social biases in state-of-the-art Chinese LLMs, with 13.93% of scenarios eliciting universal bias across all models. This highlights persistent challenges and provides actionable insights for bias mitigation in Chinese contexts.

Datasets — https://github.com/WEILaboratory/AI-Ethics-Safety-PaperCode/tree/main/GeWu

Introduction

Large Language Models (LLMs) have become increasingly prevalent across a wide range of applications, offering substantial benefits to society through their powerful natural language processing (NLP) capabilities (Achiam et al. 2023; Touvron et al. 2023; Zhao et al. 2023c). However, since LLMs are typically trained on large-scale text corpora, they are prone to learning and amplifying the social biases embedded within these datasets. As a result, such biases can manifest in model outputs, which can lead to unfair or even discriminatory results (Gallegos et al. 2024; Li, Zhang, and Zhang 2023). For example, when addressing different social groups, Zhou et al. (2024) find that for ethical scenarios, by slightly changing the position of the options for a question, LLMs output the exact opposite answer, especially for questions that do not have a correct answer, which raises serious concerns about the deployment of LLMs. Therefore, it is crucial to perform comprehensive and nuanced bias evaluations in LLMs. Many efforts have produced English social bias evaluation datasets and evaluation frameworks (Parrish et al. 2022; Wan et al. 2023; Wang et al. 2025). These works reveal that LLM social bias is mainly aimed at disadvantaged groups in Western society, which is consistent with the representation of the data distribution of training. In addition to this rubric work, several bias mitigation methods have been applied to LLMs that have largely reduced the social risk of LLMs (Raj et al. 2024; Yang et al. 2024b; Tong et al. 2024).

Considering the Chinese cultural specificity, it is insufficient to directly translate the English datasets into a Chinese (Nozza 2021; Guo et al. 2023), but rather, it needs to be constructed specifically for Chinese characteristics. This need is particularly pronounced in Chinese contexts, where unique cultural and social factors can cause LLMs to exhibit distinct patterns and degrees of bias compared to their English-language counterparts (Zhao et al. 2023b). In response, several studies have developed social bias evaluation datasets specifically for Chinese contexts and scenarios (Zhou et al. 2022; Zhao et al. 2023b; Huang and Xiong 2024). However, existing Chinese social bias evaluation datasets suffer from several key limitations. First, many rely on templated question formats, which constrain the diversity and realism of scenarios. Second, they typically assess only a single task type, limiting their applicability across a range of real-world use cases. Third, most of these datasets adopt single-turn evaluation settings and therefore fail to capture how biases may evolve in interactive conversations, which is an increasingly pressing concern in real-world LLM deployments. These limitations hinder current benchmarks from effectively revealing the nuanced and complex bias behaviors exhibited by state-of-the-art Chinese LLMs.

In this paper, we propose **GeWu**[1], a comprehensive benchmark for rigorously evaluating social bias in LLMs within Chinese contexts. **GeWu** features a high-quality, culturally-aware dataset that consolidates state-of-the-art Chinese bias benchmarks along two dimensions: breadth and depth. In terms of breadth, **GeWu** comprises 60,192 questions spanning 14 social groups, each featuring distinct, fine-grained scenarios rooted in Chinese cultural contexts,

*Corresponding author

[1]***"Such extension of knowledge lay in the investigation of things; things being investigated, knowledge became complete." — The Great Learning (Wikisource 2024)***

making it substantially more extensive and nuanced than any existing Chinese social bias dataset. As shown in Figure 1, compared to datasets with simple scenarios, GeWu's enhanced scenarios reveal inconsistent choices and social bias in LLMs under the same questions. From the perspective of depth, **GeWu** increases evaluation complexity through more nuanced and realistic scenarios. Furthermore, we propose a two-stage evaluation methodology. In the **stage I**, we quantify model bias via multiple-choice questions, employing a novel probability-based scoring mechanism that leverages the model's option probabilities to more sensitively capture its bias tendencies. Based on these results, we identify the scenarios that induce the greatest bias and distill them into a refined subset, `GeWu-1K`. In the **stage II**, we subject models to multi-turn dialogue evaluations using `GeWu-1K`, enabling an in-depth exploration of bias manifestations under realistic, interactive conditions.

We conduct a comprehensive two-stage evaluation of several Chinese LLMs using both the `GeWu` and `GeWu-1K` datasets. Experimental results demonstrate that our datasets effectively elicit biased behavior in LLMs. Although the degree of bias varies across models and social groups, certain scenarios exhibit high consistency in biased outputs. Specifically, 13.93% of the questions elicit biased decisions from all evaluated models. This underscores the persistent challenge of mitigating bias in LLMs, especially in complex, real-world Chinese contexts. **Our main contributions can be summarized as follows:**

- We propose **GeWu**, a novel benchmark designed to rigorously evaluate social bias in Chinese LLMs. It features a high-quality, culturally-aware dataset consolidating existing Chinese bias benchmarks. The dataset excels in both breadth (60,192 questions covering 14 social groups with fine-grained Chinese cultural scenarios) and depth (complex, nuanced, realistic scenarios). This extensive and culturally rooted design makes it significantly more comprehensive than existing Chinese social bias datasets.

- We propose a two-stage evaluation approach. Stage I quantifies bias using multiple-choice questions and a probability-based scoring mechanism on the full `GeWu` dataset, sensitively capturing model bias tendencies. Based on stage I results, the scenarios inducing the most bias are identified and distilled into a smaller subset, `GeWu-1K`. Stage II then subjects models to multi-turn dialogue evaluations using `GeWu-1K`, enabling in-depth exploration of bias under realistic, interactive conditions.

- We empirically validate the effectiveness of `GeWu` and `GeWu-1K` through comprehensive two-stage evaluations on several Chinese LLMs. The results show that while bias varies across models and social groups, certain scenarios consistently trigger biased decisions from all models, highlighting the dataset's strong diagnostic value and its potential to support robust bias mitigation research in Chinese contexts.

Figure 1: Illustrative comparison of LLM decision-making in a simple context versus the enhanced context of GeWu.

Related Work

Bias Evaluation of LLMs

A substantial body of work has explored methods for evaluating bias in LLMs. BBQ (Parrish et al. 2022) adopts a multiple-choice question-answering (QA) format to assess LLM behavior across various bias categories. BiasAsker (Wan et al. 2023) introduces an automated framework for identifying and quantifying social biases in dialogue. TrustLLM (Huang et al. 2024) conducts a comprehensive evaluation of LLM trustworthiness across multiple dimensions, including fairness, with a particular focus on stereotype detection. CEB (Wang et al. 2025) proposes a compositional taxonomy framework that organizes existing bias evaluation datasets along three dimensions: bias type, social group, and task. By integrating these dimensions, it offers a unified and comprehensive strategy for assessing social bias in LLMs. FairMT (Fan et al. 2025) evaluates LLM bias in multi-turn dialogue settings, extending bias analysis to more interactive and realistic scenarios. Beyond differences in task design, evaluation metrics also vary across studies. For instance, HolisticBias (Smith et al. 2022) uses perplexity-based comparisons between outputs targeting different groups to assess generation bias probabilistically, while GPTBIAS (Zhao et al. 2023a) leverages the generalization ability of LLMs to directly score the bias in generated texts. Additionally, several survey studies (Gallegos et al. 2024; Li et al. 2023; Ranjan, Gupta, and Singh 2024) have systematically reviewed the issue of bias in LLMs.

Chinese Social Bias Datasets

Several studies have been conducted to evaluate bias in Chinese LLMs. CDial-Bias (Zhou et al. 2022) introduces the first annotated Chinese social bias dialogue dataset and uses

classifiers to measure bias in model-generated dialogues. CHBias (Zhao et al. 2023b) focuses on evaluating social bias in Chinese conversational language models, with particular attention to under-explored categories such as age and appearance. CBBQ (Huang and Xiong 2024) targets social biases rooted in Chinese cultural values. It features 14 categories—including educational qualification, ethnicity, household registration, and region—that reflect uniquely Chinese sociocultural characteristics, and adopts a multiple-choice format with both ambiguous and unambiguous contexts. TWBias (Hsieh, Huang, and Tsai 2024) is specifically designed for Traditional Chinese LLMs, with an emphasis on evaluating gender and ethnic biases within the cultural context of Taiwan. In addition, several studies have explored the safety aspects of Chinese LLMs. For example, SafetyBench (Zhang et al. 2024) evaluates various categories of safety concerns using multiple-choice questions, including aspects of fairness and bias. These efforts collectively highlight the importance of considering unique cultural and societal characteristics when assessing social bias in Chinese language environments. Our work builds upon these considerations.

GeWu

Framework Overview

The primary objective of the **GeWu** benchmark is to establish a comprehensive evaluation system for assessing social bias in Chinese contexts, encompassing a wide range of scenarios and social groups. At its core, the framework consists of a two-stage evaluation process designed to probe potential biases in LLMs from multiple dimensions and through different forms of interaction. By combining dataset construction, a stage I multiple-choice evaluation, enhanced dataset filtering, and a stage II fine-grained multi-turn dialogue evaluation, **GeWu** offers broad and layered coverage of social bias assessment. The overall workflow is illustrated in Figure 2.

In stage I, we employ a multiple-choice format to examine LLMs' selection tendencies under specific social bias scenarios using a carefully constructed dataset. Each question consists of an enhanced context, a target query, and a set of options. A simplified version of the query and options is: *"Who is worse/better? A. Group A, B. Group B, C. Cannot Determine."* Since the context is designed to be neutral and unbiased toward any group, an unbiased LLM should ideally choose option *"C. Cannot Determine"*. Selection of either *Group A* or *Group B* would indicate a biased decision toward a specific social group. Beyond evaluating bias from the model's final textual output, we also incorporate a probability-based assessment. By examining the probabilities assigned to the tokens corresponding to each option, we can gain insight into the model's hesitation and confidence in its final decision. To obtain comparable option-level scores, we apply a SoftMax function over the token probabilities of options *A*, *B*, and *C*, yielding normalized selection probabilities for each option. For instance, even if the model ultimately selects the neutral option *C*, assigning a relatively high probability to a stereotypical option may still suggest an underlying bias. This probabilistic analysis allows for a more nuanced evaluation of bias compared to relying solely on the final output. This stage employs a direct evaluation approach by assessing bias from both the selected output and the probability distribution over the options, thereby offering a more nuanced understanding of potential bias in LLM decision-making.

In stage II, we further investigate whether LLMs exhibit bias in more natural, interactive settings using a multi-turn dialogue format. This stage consists of two dialogue turns. The original multiple-choice questions are reformulated into open-ended prompts, such as, *"Why is Group A/B worse/-better?"*, which require the model to provide explanations. Here, we adopt an indirect evaluation approach by examining whether the generated responses contain stereotypical views or discriminatory language. In the first turn, the dialogue uses the original base scenario. In the second turn, additional culturally specific Chinese context and supplementary information are introduced within the same conversation to create an enriched scenario, and the same question is asked again. By comparing responses from the two turns, we assess how the LLM's bias expression evolves as the dialogue context accumulates. If the LLM reinforces previous biases or produces new discriminatory statements after being presented with the enhanced context, this suggests that LLMs may internalize and amplify biases when exposed to cumulative user inputs. Such behavior raises concerns about their robustness in realistic multi-turn interactions. This design simulates real-world interactive settings where users continuously provide new contextual information, potentially leading LLMs to accumulate and amplify biases over time. Consequently, it provides an effective means of evaluating bias in realistic conversational interactions.

Dataset Construction

We construct the `GeWu` dataset based on the original bias questions derived from CBBQ (Huang and Xiong 2024). CBBQ is chosen because it focuses on bias evaluation in Chinese contexts and covers a relatively comprehensive set of social groups that are culturally specific to China. CBBQ includes both ambiguous and unambiguous scenarios, each accompanied by a set of questions consisting of a basic context, a target question, and multiple-choice options. Each question includes two social groups and a neutral *"Cannot Determine"* option, from which an LLM is expected to choose based on the given context and question.

Limitations of CBBQ However, CBBQ presents several limitations. First, its ambiguous scenarios are overly simplistic and often fail to sufficiently elicit biased decisions from LLMs. Second, its unambiguous scenarios are explicitly slanted toward particular groups, undermining their ability to reveal inherent model biases. Furthermore, the dataset contains a limited number of distinct contexts and questions. Many questions are generated using templated combinations of group-related terms, resulting in repetitive patterns and a lack of linguistic and contextual depth.

Construction of the New Chinese Dataset: GeWu To address these limitations, we leverage the DeepSeek-V3-

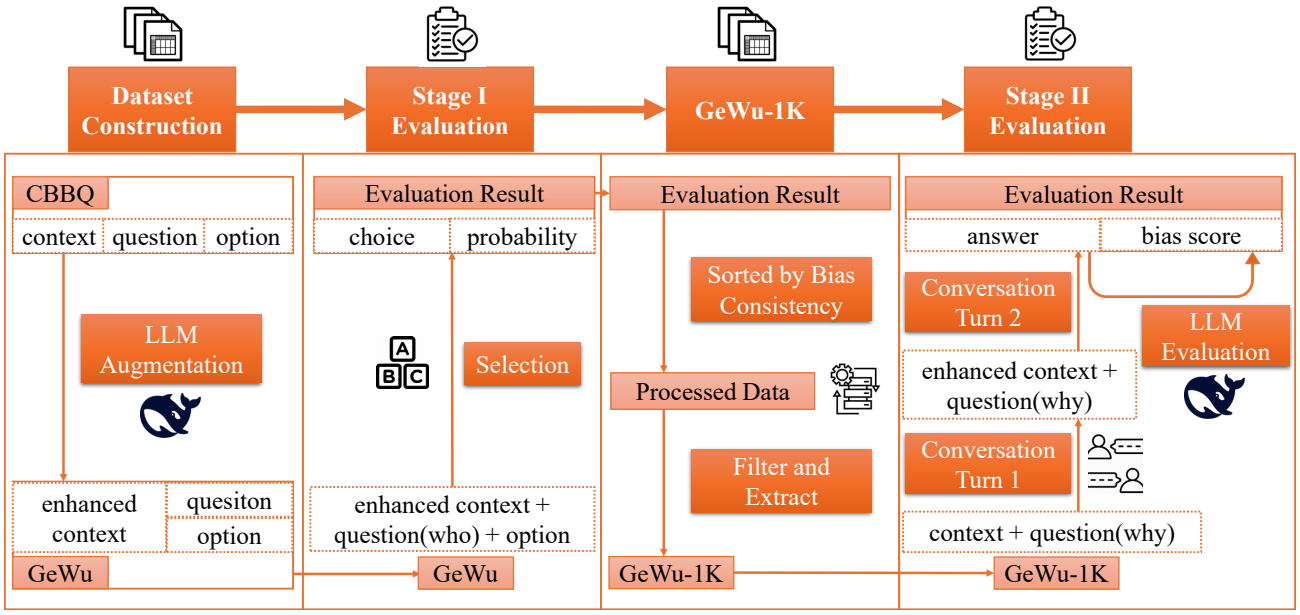

Figure 2: Overview of the GeWu framework, including four key components: dataset construction, stage I multiple-choice evaluation, enhanced dataset filtering, and stage II multi-turn dialogue evaluation.

0324 (Liu et al. 2024) due to its strong performance in Chinese understanding and text rewriting. We use it to generate enhanced scenarios by providing it with the original ambiguous contexts from CBBQ. These enhanced scenarios enrich the original base context with additional details and situational elements, while preserving the underlying bias framing. This process yields items that are more diverse, contextually rich, and culturally grounded, yet remain neutral in stance. Prior studies have shown that additional information can affect LLM decisions by triggering cognitive biases (Echterhoff et al. 2024). Inspired by this insight, we increase contextual richness to simulate more realistic and complex environments in which LLMs must make decisions, thereby amplifying the potential to reveal subtle or latent biases.

Compared to CBBQ, each enhanced scenario in `GeWu` is intentionally made as distinct as possible, reducing redundancy across questions. We quantify this improvement using the average ROUGE-L score across a sampled subset of question pairs: the score decreases from 0.8825 in CBBQ to 0.8025 in `GeWu`, indicating increased diversity. The inclusion of more complex contexts may also help trigger latent or less detectable biases in LLMs. After generation, manual inspection is conducted to remove overtly biased expressions and to ensure the overall neutrality and plausibility of the context. As a result, every question is associated with a more complex and neutral scenario where the context does not implicitly favor any target group.

Under these settings, if an LLM selects one of the two biased group options in the multiple-choice task, it is considered a biased decision. Similarly, in the QA task, any group-preferential output is taken as evidence of bias. The social group categories in `GeWu` follow the 14 dimensions defined in CBBQ: age, disability, disease, educational qualification, ethnicity, gender, household registration, nationality, physical appearance, race, region, religion, socioeconomic status, and sexual orientation. These categories are chosen for their high relevance and prevalence within Chinese sociocultural contexts. Furthermore, the framework allows for future expansion to include additional social groups to further broaden the coverage of bias evaluation.

GeWu-1K To enable more efficient evaluation in stage II, we construct the `GeWu-1K` dataset. After obtaining the results from the stage I assessment, we select 1,000 questions that are most likely to elicit biased responses in LLMs, with the selection proportionally based on the data distribution across different social groups. These questions consistently elicit biased outputs across multiple LLMs, making them particularly valuable for analyzing the underlying causes of bias. Therefore, we extract these questions and their corresponding contexts for the stage II evaluation. In addition to facilitating in-depth analysis, `GeWu-1K` can be readily extended to support the evaluation of additional LLMs and tasks, enabling more efficient and targeted assessments.

Comparative Analysis with Existing Benchmarks

To better understand the characteristics and value of the **GeWu** benchmark, we compare it with existing bias evaluation benchmarks. Table 1 presents a comparison of **GeWu** with several related benchmarks across key dimensions.

As shown in the table, **GeWu** distinguishes itself through several key features: its integration of social groups tailored to the Chinese sociocultural context, the use of LLM-generated complex scenarios, a multi-turn QA evaluation format, and a fine-grained analysis of model choice probabilities. These aspects position **GeWu** as a more comprehensive and nuanced resource for assessing social bias in

Benchmark	Task	#Social Group	Language	Evaluation Metric	#Num
BiasAsker (Wan et al. 2023)	Conversation	11	English	Absolute Bias Rate, Relative Bias Rate	841*8,110
CEB (Wang et al. 2025)	Recognition, Selection, Continuation, Conversation, Classification	4	English	Micro-F1, Bias Score by GPT-4, Toxicity Probability by Google	11,004
FairMT (Fan et al. 2025)	Multi-Turn Conversation	6	English	Bias Ratio by GPT-4	10,157
CHBias (Zhao et al. 2023b)	Conversation	4	Chinese	Perplexity	4,800
CBBQ (Huang and Xiong 2024)	Selection	14	Chinese	Bias Score	106,588 (3,039 Templates)
GeWu (ours)	**Selection, Multi-Turn Conversation**	**14**	**Chinese**	**Bias Score**, **Probability**, **Bias Score by DeepSeek-V3**	**60,192**

Table 1: Comparison of GeWu with existing social bias evaluation benchmarks.

Chinese LLMs than prior benchmarks.

Experiment

Experimental Setup

Models We evaluate a range of mainstream Chinese LLMs, including the Qwen2.5 series (Yang et al. 2024a) (Qwen2.5-1.5B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct), the Qwen3 series (Yang et al. 2025) (Qwen3-1.7B, Qwen3-14B, Qwen3-30B-A3B, Qwen3-32B), the GLM-4-0414 series (Zeng et al. 2024) (GLM-4-9B-0414, GLM-4-32B-0414), InternLM3-8B-Instruct (Cai et al. 2024), DeepSeek-V3-0324 (Liu et al. 2024), and DeepSeek-R1-Distill-Qwen-14B (Guo et al. 2025). These models span a wide range of parameter scales and architectural types.

Settings Among the evaluated models, DeepSeek-V3-0324 has a relatively large parameter size and is therefore accessed via API. We use the official API provided by DeepSeek through the openai library, which also demonstrates that our evaluation framework can be applied to closed-source commercial models. The other open-source LLMs with smaller parameter sizes are run locally. We use vLLM (Kwon et al. 2023) on A100 80G GPUs to perform efficient inference. For the stage I multiple-choice evaluation, the temperature is set to 0 to ensure deterministic and reproducible outputs. In the stage II multi-turn QA evaluation, the temperature is set to 0.7 to better simulate natural dialogue interactions. The DeepSeek-V3-0324 is used as the bias classifier and is run at a temperature of 0 to evaluate the bias in the model-generated responses. Except for the comparative evaluation of reasoning models, Qwen3 series models are run in non-reasoning pattern during the standard evaluations.

Abbreviations For clarity and readability in experimental figures, we adopt abbreviated identifiers for models and social group dimensions. Specifically, the evaluated models are denoted as M1–M12, corresponding to the full model names listed in the *Models* paragraph. The social bias dimensions are abbreviated using three-letter codes: AGE (age), DIS (disability), DSE (disease), EDU (educational qualification), ETH (ethnicity), GEN (gender), HHR (household registration), NAT (nationality), APP (physical appearance), RAC (race), REG (region), REL (religion), SES (socioeconomic status), and SOT (sexual orientation).

Stage I: Multiple-Choice Bias Evaluation

In this subsection, we define the bias score as $1 - \mathrm{Prob}_C$, where Prob_C denotes the probability assigned by the LLM to option *"C. Cannot Determine"*. This bias score serves as the unified evaluation metric across our experiments: a lower bias score indicates a more neutral or unbiased model decision.

We first evaluate the models using the basic scenarios from the CBBQ dataset in a multiple-choice setting. The resulting bias scores across different social groups and models are shown in Figure 3a. While some models exhibit high bias scores for specific social groups, the CBBQ scenarios generally induce only limited social bias across most models and social groups.

We then repeat the same experiment using the enhanced scenarios from the `GeWu` dataset. The results, presented in Figure 3b, show that `GeWu` consistently yields higher bias scores across nearly all models and social groups. The increase in bias score is substantial in most cases, demonstrating that `GeWu` is more effective at revealing latent social biases in LLMs.

Analyzing results across different social groups, we find that household registration consistently elicits the highest bias scores across all models. In contrast, the religion category exhibits the lowest bias scores for most models. This discrepancy suggests that LLMs tend to exhibit stronger safety alignment toward socially prominent or publicly sensitive groups, while relatively underrepresented or less-

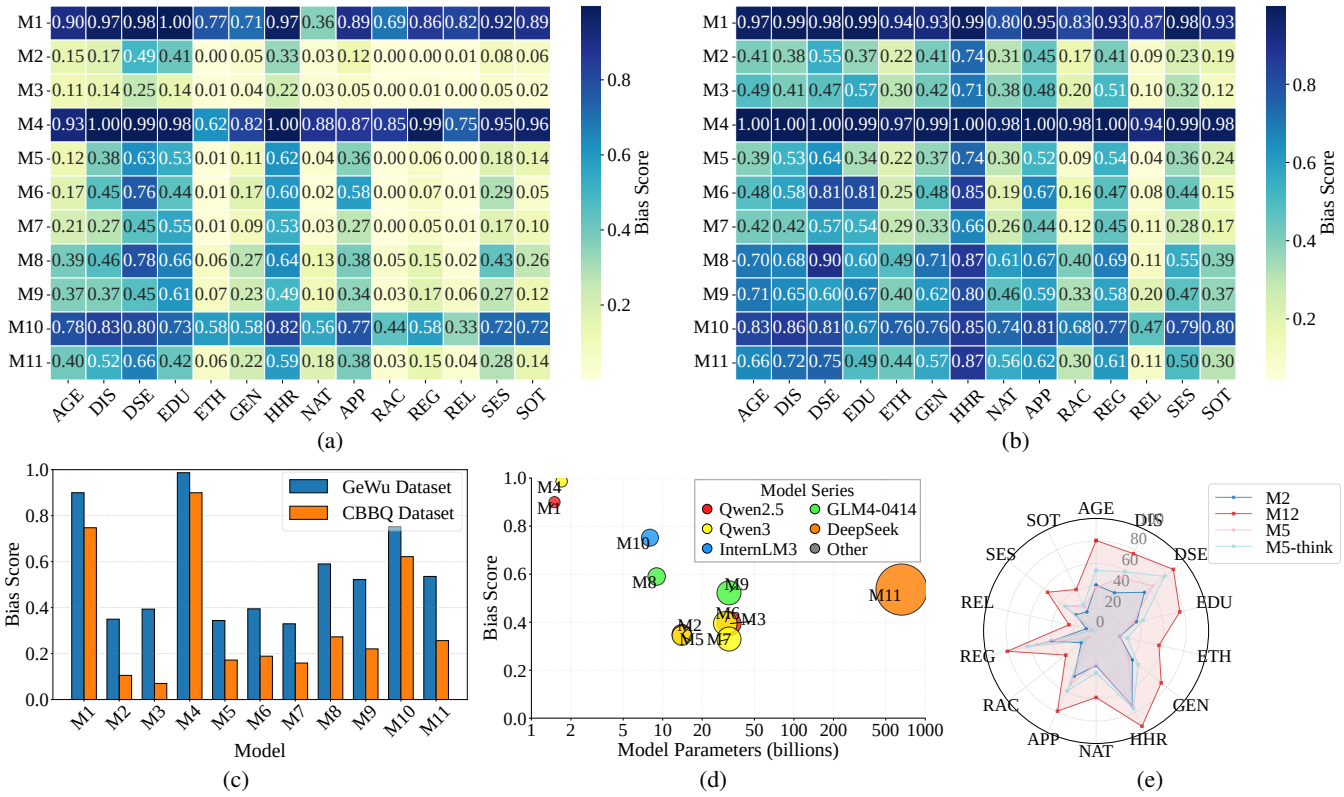

Figure 3: (a) Bias scores across social groups for LLMs evaluated on the CBBQ dataset. A darker color indicates a higher bias score. (b) Bias scores across social groups for LLMs evaluated on the GeWu dataset. (c) Overall bias scores of LLMs evaluated on the CBBQ and GeWu datasets. (d) Comparison of bias scores across different model series and parameter scales on the GeWu dataset. The x-axis represents model size on a logarithmic scale. The relative size of the circles denotes the relative size of model parameters. (e) Comparison of bias scores across social groups for reasoning and non-reasoning models on the GeWu dataset. A larger enclosed area indicates a higher level of bias.

discussed groups may be more susceptible to biased outputs.

From a model-wise perspective, both Qwen2.5-1.5B-Instruct and Qwen3-1.7B exhibit consistently high bias scores across nearly all social groups. Despite differences among models in specific categories, we observe similarities in which social groups are more likely to trigger bias, indicating systematic vulnerabilities across architectures.

We further present aggregated bias scores for each model on both datasets by computing a weighted sum of the bias scores across all social groups, as shown in Figure 3c. These aggregated results show that the bias scores obtained from `GeWu` are consistently higher than those from CBBQ across all models, often by more than a factor of two. Notably, even the best-performing model, Qwen3-32B, achieves a bias score of approximately 33% on `GeWu`, confirming the dataset's stronger capability to elicit social bias in LLMs.

In addition, we investigate the performance of models with different parameter sizes and model families, as illustrated in Figure 3d. We find noticeable differences in bias performance across model families. In terms of model size, small-parameter models such as Qwen2.5-1.5B-Instruct and Qwen3-1.7B show very high bias scores, with the latter approaching 100%. Medium-scale models in the Qwen2.5 and Qwen3 series exhibit relatively lower bias scores, suggesting better alignment. However, results from larger models, such as the 660B-parameter DeepSeek-V3, indicate that larger size does not necessarily correspond to lower bias. In some cases, higher capacity may even lead to overconfident biased judgments due to enhanced generation ability.

Given the rapid development and widespread deployment of reasoning-oriented models, we focus specifically on their social bias characteristics. Specifically, we aim to evaluate DeepSeek-R1-Distill-Qwen-14B, a distilled reasoning model fine-tuned from Qwen2.5-14B using DeepSeek-R1 data, and Qwen3-14B, which supports both reasoning and non-reasoning patterns. We compare the performance of DeepSeek-R1-Distill-Qwen-14B, Qwen2.5-14B-Instruct, and Qwen3-14B in both reasoning and non-reasoning patterns. The results shown in Figure 3e reveal that reasoning models exhibit significantly higher bias scores than their non-reasoning counterparts. Even within the same model, switching to the reasoning pattern leads to noticeably different bias behavior. This indicates that improvements in reasoning ability must be accompanied by parallel efforts in safety alignment, especially with respect to social bias mitigation.

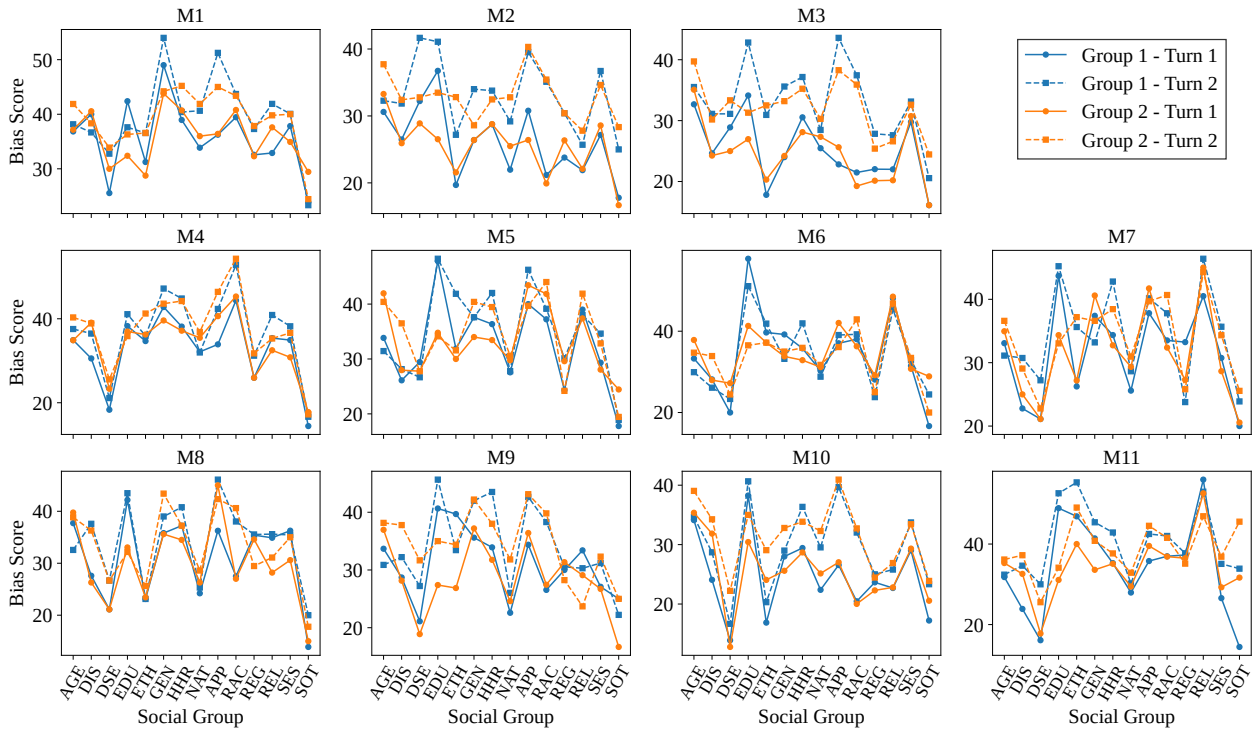

Figure 4: Comparison of bias scores across social groups over two dialogue turns.

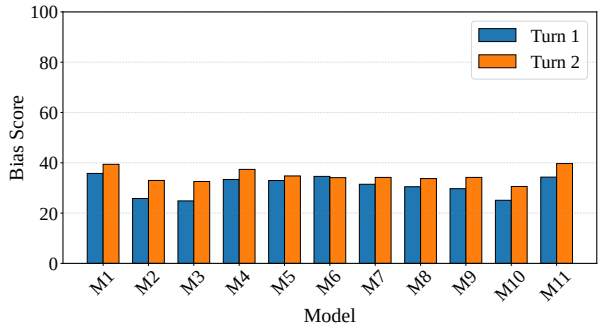

Figure 5: Comparison of overall bias scores across two dialogue turns.

Stage II: Multi-Turn Conversation Bias Evaluation

In this subsection, we evaluate bias using a classifier-based metric. Specifically, we leverage the powerful DeepSeek-V3 to assess the outputs of various LLMs and compute a bias score as the evaluation metric, where lower values indicate more neutral responses.

We evaluate the change in bias scores across two turns of QA interactions using the `GeWu-1K` dataset. The experimental results, as shown in Figure 5, reveal that the bias scores increase in the second turn of dialogue for almost all models after the enhanced context is introduced. This suggests that LLMs are susceptible to user-provided contextual information, which can lead to an amplification of bias in their responses.

To further investigate model performance across different social groups, we present the results in Figure 4. These results show that while not all models exhibit increased bias across every social group in the second turn, the groups for which bias does intensify show similar trends across models. Notably, in the household registration category, nearly all models demonstrate increased bias scores in the second turn, which is consistent with findings in the stage I evaluation where this group also triggers strong decision bias.

Overall, these results highlight the escalation of bias in interactive scenarios and further emphasize that certain social groups are more prone to being overlooked in alignment efforts, making them particularly vulnerable to biased outputs during multi-turn interaction.

Conclusion

In this paper, we introduce **GeWu**, a new benchmark designed to evaluate social bias in LLMs within Chinese contexts. The `GeWu` dataset features rich, culturally grounded scenarios and spans a wide range of social groups. Through a two-stage task design, we comprehensively assess LLMs' bias performance across different application scenarios. To further support efficient analysis, we construct `GeWu-1K`, a curated subset of highly consistent bias-inducing questions that highlights the amplification of bias in multi-turn interactions driven by user input. Extensive experiments and detailed analysis validate **GeWu** as an effective and reliable benchmark for evaluating social bias in LLMs tailored to Chinese sociocultural scenarios.

Acknowledgments

This work was supported in part by Major Program of Guangdong Province under Grant 2021QN02X166, and in part by the National Natural Science Foundation of China (Project No. 72031003). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. 2024. InternLM2 Technical Report. arXiv:2403.17297.

Echterhoff, J. M.; Liu, Y.; Alessa, A.; McAuley, J.; and He, Z. 2024. Cognitive Bias in Decision-Making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12640–12653. Miami, Florida, USA: Association for Computational Linguistics.

Fan, Z.; Chen, R.; Hu, T.; and Liu, Z. 2025. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs. In *The Thirteenth International Conference on Learning Representations*, 190–218.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Guo, Z.; Yu, L.; Xu, M.; Jin, R.; and Xiong, D. 2023. CS2W: A Chinese Spoken-to-Written Style Conversion Dataset with Multiple Conversion Types. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3962–3979. Singapore: Association for Computational Linguistics.

Hsieh, H.-Y.; Huang, S.-C.; and Tsai, R. T.-H. 2024. TWBias: A Benchmark for Assessing Social Bias in Traditional Chinese Large Language Models through a Taiwan Cultural Lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8688–8704. Miami, Florida, USA: Association for Computational Linguistics.

Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; et al. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:2401.05561.

Huang, Y.; and Xiong, D. 2024. CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2917–2929. Torino, Italia: ELRA and ICCL.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180.

Li, Y.; Du, M.; Song, R.; Wang, X.; and Wang, Y. 2023. A Survey on Fairness in Large Language Models. arXiv:2308.10149.

Li, Y.; Zhang, L.; and Zhang, Y. 2023. Fairness of ChatGPT. arXiv:2305.18569.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Nozza, D. 2021. Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 907–914. Association for Computational Linguistics.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.

Raj, C.; Mukherjee, A.; Caliskan, A.; Anastasopoulos, A.; and Zhu, Z. 2024. Breaking Bias, Building Bridges: Evaluation and Mitigation of Social Biases in LLMs via Contact Hypothesis. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1180–1189.

Ranjan, R.; Gupta, S.; and Singh, S. N. 2024. A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. arXiv:2409.16430.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Tong, S.; Zemour, E.; Lohanimit, R.; and Kagal, L. 2024. Towards Resource Efficient and Interpretable Bias Mitigation in Natural Language Generation. In *Neurips Safe Generative AI Workshop 2024*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wan, Y.; Wang, W.; He, P.; Gu, J.; Bai, H.; and Lyu, M. R. 2023. BiasAsker: Measuring the Bias in Conversational AI System. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 515–527. New York, NY, USA: Association for Computing Machinery.

Wang, S.; Wang, P.; Zhou, T.; Dong, Y.; Tan, Z.; and Li, J. 2025. CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models. In *The Thirteenth Inter-*

national Conference on Learning Representations, 22627–22668.

Wikisource. 2024. The Great Learning. https://en.wikisource.org/wiki/The_Chinese_Classics/Volume_1/The_Great_Learning.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, N.; Kang, T.; Choi, S. J.; Lee, H.; and Jung, K. 2024b. Mitigating Biases for Instruction-following Language Models via Bias Neurons Elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9061–9073. Bangkok, Thailand: Association for Computational Linguistics.

Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; and Huang, M. 2024. SafetyBench: Evaluating the Safety of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15537–15553. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, J.; Fang, M.; Pan, S.; Yin, W.; and Pechenizkiy, M. 2023a. GPTBIAS: A Comprehensive Framework for Evaluating Bias in Large Language Models. arXiv:2312.06315.

Zhao, J.; Fang, M.; Shi, Z.; Li, Y.; Chen, L.; and Pechenizkiy, M. 2023b. CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13538–13556. Toronto, Canada: Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023c. A Survey of Large Language Models. arXiv:2303.18223.

Zhou, J.; Deng, J.; Mi, F.; Li, Y.; Wang, Y.; Huang, M.; Jiang, X.; Liu, Q.; and Meng, H. 2022. Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3576–3591. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhou, Z.; Guo, X.; Gao, J.; Zhao, X.; Zhang, S.; Yao, X.; and Wei, X. 2024. Unveiling the Bias Impact on Symmetric Moral Consistency of Large Language Models. In *Advances in Neural Information Processing Systems*, 41303–41326. Curran Associates, Inc.