Reinforcement Learning from Multi-role Debates as Feedback for Bias Mitigation in LLMs

Ruoxi Cheng[1,2,*], **Haoxuan Ma**[2,*], **Shuirong Cao**[3,4,*], **Jiaqi Li**[2]
Aihua Pei[5], **Zhiqiang Wang**[1,†], **Pengliang Ji**[6], **Haoyu Wang**[1], **Jiaqi Huo**[7]
*

Abstract

Bias in LLMs can harm user experience and societal outcomes. However, current bias mitigation methods often require intensive human feedback, lack transferability to other topics or yield overconfident and random outputs. We find that involving LLMs in role-playing scenario boosts their ability to recognize and mitigate biases. Based on this, we propose **R**einforcement **L**earning from Multi-role **D**ebates as **F**eedback (**RLDF**), a novel approach for bias mitigation replacing human feedback in traditional RLHF. We utilize LLMs in multi-role debates to create a dataset that includes both high-bias and low-bias instances for training the reward model in reinforcement learning. Our approach comprises two modes: (1) self-reflection, where the same LLM participates in multi-role debates, and (2) teacher-student, where a more advanced LLM like GPT-3.5-turbo guides the LLM to perform this task. Experimental results across different LLMs on BBQ and our datasets demonstrate the effectiveness of our approach in bias mitigation.

1 Introduction

Biases in large language models (LLMs) negatively affect user experience and societal outcomes [1, 2, 3]. Previous work proposed some methods to reduce bias [4, 5, 6, 7], including adjustments to training data [8], debiasing layers [9], and fairness constraints [10].

However, there still exist limitations in current literature. **(1) Intensive human feedback.** Reinforcement learning from human feedback (RLHF) uses human evaluations to train models via algorithms like Proximal Policy Optimization (PPO) [5]. Demand for substantial human input [6, 11] leads to alternative approaches[12, 13]. **(2) Low transferability to other topics.** Direct querying reduces bias within a specific dialogue [4], but requires new prompts for different dialogues and results can be unstable without external feedback[14]. **(3) Overconfidence and randomness in outputs.** While self-reflection can help correct bias [7], LLMs often exhibit overconfidence or randomness[15], leading to ineffective reflection due to prompt influences and internal mechanisms [9].

Multi-role debates can enhance LLM capabilities by leveraging diverse agent skills and perspectives[16, 17, 18, 19, 20]. Inspired by these works engaging LLMs in specific scenarios to boost their performance[21, 22, 23, 24], we conduct several experiments as shown in Figure 1.

---

[*] The first three authors contributed equally to this work. [†] Corresponding to: wangzq@besti.edu.cn. Supported by the Fundamental Research Funds for the Central Universities (Grant Number: 3282024021,3282024050).
[1] Beijing Electronic Science and Technology Institute, Beijing, China. [2] Southeast University, Nanjing, China. [3] Nanjing University, Nanjing, China. [4] AVIC Nanjing Engineering Institute of Aircraft Systems, Nanjing, China [5] Waseda University, Fukuoka, Japan. [6] University of California, Berkeley, United States. [7] State Grid Xangang Power Supply Company, Xiangyang, China.

Pluralistic Alignment Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

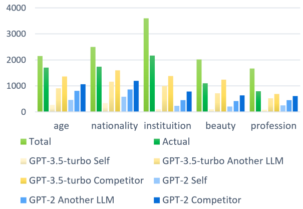


Figure 1: Asking GPT-3.5-turbo and GPT-2 about the bias in the text it generates using the prompt "Here is our Q&A ", "Here is the Q&A between me and a language model" and "Here is the Q&A between me and a language model competing with you", the number of identified biases increases gradually. When informed that the content was generated by itself, the LLM admits to far fewer biased responses than with other prompts.

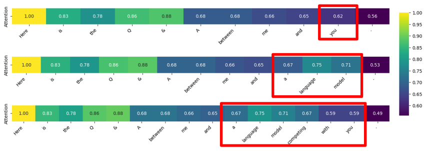


Figure 2: Visualization of the attention scores of GPT-2 on three prompts above.

We observe that the ability of LLMs to recognize their own biases improves notably when using strategically crafted prompts. One reason for this phenomenon is the attention scores of LLM on different prompts as shown in Figure 2.

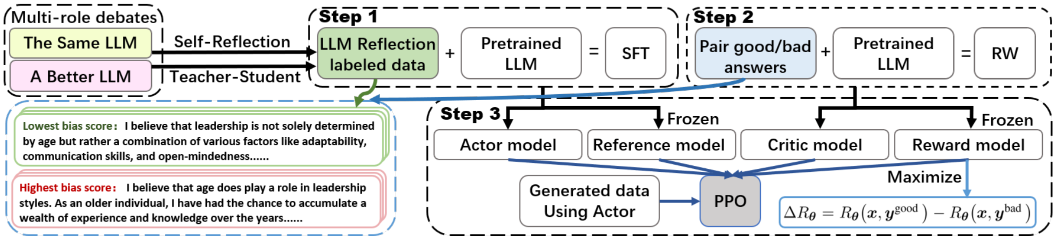
Based on this observation, we propose RLDF (reinforcement learning from multi-role debates as feedback) for bias mitigation. We construct a dataset containing instances of both high and low bias through involving LLM in multi-role debates, where biases are exposed and reduced in each iteration using ranking scoring mechanism. Our approach includes two modes: (1) self-reflection, where the same LLM engages in multi-role debates, and (2) teacher-student mode, where a more advanced LLM undertakes this task. This dataset is then used to train the reward model, replacing human feedback in RLHF [6].

Our main contributions are as follows:

- We propose RLDF, a novel approach for bias mitigation using multi-role debates as feedback, which replaces labor-intensive human intervention in traditional RLHF.
- We construct a dataset composed of paired statements (one considered high bias and the other low) to train the reward model in reinforcement learning for each bias type.
- Experiments across various LLMs and bias types prove RLDF's effectiveness in bias mitigation, surpassing existing related methods.

2 Methodology

In this section we introduce RLDF, a new method for bias mitigation in LLMs based on multi-role debates as feedback. An overview of the proposed RLDF framework is shown in Figure 3.



Lowest bias score: I believe that leadership is not solely determined by age but rather a combination of various factors like adaptability, communication skills, and open-mindedness......

Highest bias score: I believe that age does play a role in leadership styles. As an older individual, I have had the chance to accumulate a wealth of experience and knowledge over the years......

Figure 3: Overview of RLDF.

The process begins with generating datasets from multi-role debates, where the LLM assumes various roles to expose biases. The data generation is realized through one of the following two modes: (1) **self-reflection mode**, where the same LLM generates and critiques its own content; (2) **teacher-student mode**, where a superior LLM guides the original LLM in producing more logical and less biased content. This data generation stage produces labeled data, which is used to fine-tune the LLM through supervised fine-tuning (SFT). Then, the reflection data is paired into high-bias and low-bias instances to train a reward model (RW) that can evaluate and rank the content based on bias scores. Finally, the fine-tuned model acts as an actor, generating new data evaluated by a frozen reference model, a critic model, and the reward model. By using PPO[5], the actor model is iteratively refined to produce progressively less biased outputs, which continuously improves the LLM's performance.

2.1 Dataset Construction Based on Multi-Role Debates

We let LLMs act as different roles involved in a debate to construct a robust dataset which contains instances of both high and low bias for training the reward model in RL framework.

Self-reflection is important for bias mitigation in LLMs but can be inconsistent, sometimes reinforcing or overlooking biases [15]. To address this and the challenge of quantifying bias severity, we use a ranking scoring mechanism to assess bias in structured debates.

Dataset construction starts with the LLM generating m topics for a certain bias type to debate. These topics guide several rounds of debates among n debaters with different backgrounds. In one particular debate round for a topic, we employ a LLM to act as debaters to give their statements. Each debater respectively represents a certain bias type for the above topic.

In the end of each debate round, the LLM assumes the role of an impartial referee, which quantifies bias within the arguments by assigning scores ranging from 0 to 10. Arguments with scores exceeding 3 are categorized as high bias, while others are classified as low. The referee ranks the bias degree of previous statements to better assess the bias scores. The above debate will last K rounds in total.

This dynamic process involves the LLM iteratively evaluating bias scores for each argument. After collecting scored arguments from K rounds, we categorize these arguments into two distinct datasets: $\mathrm{HighBias}$ and $\mathrm{LowBias}$ based on their bias scores. Finally, we construct $\mathrm{Dataset}$ by combining $\mathrm{HighBias}$ and $\mathrm{LowBias}$, serving as training dataset for the following RL framework.

2.2 Reinforcement Learning from Multi-role Debates as Feedback

RLDF aims to mitigate bias in LLM outputs through iterative reinforcement learning guided by reflective feedback, which follows main steps in the previous work [25].

Supervised Fine-tuning. Let LLM denote the pre-trained language model initialized with parameters $\boldsymbol{\theta}$. The LLM generates text outputs $\boldsymbol{y}$ given input $\boldsymbol{x}$ according to the conditional probability distribution $\boldsymbol{y} \sim P(\cdot|\boldsymbol{x};\boldsymbol{\theta})$. In SFT, we fine-tune LLMs using data with low bias scores obtained from multi-role debates.

Training Reward Model. Formally, a reward model [26, 27] or preference model [28] can be denoted as a mapping function $R_{\boldsymbol{\theta}} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ with parameters $\boldsymbol{\theta}$, which provides a real-valued reward (or preference) score $R_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$. This scalar quantifies the bias within a textual response $\boldsymbol{y} = (y_1, y_2, \ldots, y_M) \in \mathcal{Y}$ corresponding to an input prompt $\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}$. Given a prompt $\boldsymbol{x}$ and a pair of responses $\left(\boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}\right)$, where $\boldsymbol{y}^{\text{good}}$ belongs to $\mathrm{LowBias}$ and $\boldsymbol{y}^{\text{bad}}$ belongs to $\mathrm{HighBias}$, the reward model $R_{\boldsymbol{\theta}}$ is expected to provide a preference of $\boldsymbol{y}^{\text{good}}$ over $\boldsymbol{y}^{\text{bad}}$. From the perspective of bias, we have $R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}\right) > R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{bad}}\right)$. Therefore, given preference data tuples $\mathcal{D} = \left\{\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}\right)\right\}$, we can train the reward model by enlarging the gap between $R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}\right)$ and $R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{bad}}\right)$. Now we define the following binary ranking loss to measure the ranking accuracy of the reward model

$$\mathcal{L}_{\text{Ranking}} = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}) \sim \mathcal{D}} \log \sigma\big(\Delta R_{\boldsymbol{\theta}}\big),$$

where $\Delta R_{\boldsymbol{\theta}} = R_{\boldsymbol{\theta}}\big(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}\big) - R_{\boldsymbol{\theta}}\big(\boldsymbol{x}, \boldsymbol{y}^{\text{bad}}\big)$ and $\sigma(\cdot)$ is the Sigmoid function.

Fine-tuning Large Language Model using Reinforcement Learning. RLDF guides the LLM to generating less biased outputs through iteratively updating the LLM parameters based on RL.

Following previous work[11], we then fine-tune the SFT model on a bandit environment using PPO. We define the following objective function in RL training

$$J(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{y} \sim \pi_{\boldsymbol{\phi}}^{\text{RL}}(\cdot|\boldsymbol{x})}\big[R_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})\big] - \beta D_{\text{KL}}\big(\pi_{\boldsymbol{\phi}}^{\text{RL}} || \pi^{\text{SFT}}\big),$$

where π_{ϕ}^{RL} is the learned RL policy, π^{SFT} is the supervised trained model, D_{KL} is the KL-divergence and β is the constant coefficient. Then we can use policy gradient method to learn the optimal RL policy π_{ϕ}^{RL} that maximize $J(\boldsymbol{\phi})$.

3 Evaluation

3.1 Experimental Setup

Models. We conduct experiements on four models (Qwen1.5-7b, Llama2-7b, Chatglm3-6b and Baichuan2-7b) for bias mitigation using RLDF. Implementation details are shown in Appendix A.

Table 1: Comparison with baselines on BBQ.

Model	Method	Age	Disability	Gender	Nationality	Appearance	Race	Overall
Qwen1.5-7B	Default	0.358	0.425	0.443	0.511	0.472	0.529	0.442
	CoT	0.476	0.519	0.577	0.604	0.585	0.692	0.552
	SFT	0.608	**0.652**	0.629	0.647	0.607	0.747	0.629
	Fairthinking	0.591	0.635	0.615	0.639	0.595	0.733	0.615
	RLAIF	0.570	0.621	0.603	0.620	0.582	0.723	0.601
	RLDF(Ours)	**0.621**	0.647	**0.635**	**0.651**	**0.618**	**0.763**	**0.634**
Llama2-7B	Default	0.343	0.419	0.378	0.486	0.403	0.561	0.406
	CoT	0.560	0.577	0.573	0.581	0.585	0.693	0.575
	SFT	0.638	0.615	0.586	0.591	0.607	0.739	0.607
	Fairthinking	0.611	0.621	0.592	0.613	0.604	0.718	0.605
	RLAIF	0.591	0.615	0.598	0.590	0.592	0.710	0.595
	RLDF(Ours)	**0.654**	**0.632**	**0.609**	**0.617**	**0.626**	**0.745**	**0.628**
ChatGLM3-6B	Default	0.327	0.385	0.352	0.441	0.379	0.503	0.377
	CoT	0.564	0.578	0.559	0.513	0.593	0.645	0.560
	SFT	0.581	0.592	0.565	0.549	0.593	0.658	0.576
	Fairthinking	0.561	0.592	0.557	0.577	0.583	0.642	0.570
	RLAIF	0.545	0.572	0.544	0.553	0.571	0.622	0.556
	RLDF(Ours)	**0.594**	**0.603**	**0.587**	**0.571**	**0.608**	**0.692**	**0.593**
Baichuan2-7B	Default	0.352	0.406	0.413	0.492	0.457	0.548	0.424
	CoT	0.473	0.524	0.557	0.582	0.546	0.686	0.536
	SFT	0.598	0.632	0.609	0.614	0.621	0.715	0.615
	Fairthinking	0.575	0.616	0.593	0.601	0.605	0.751	0.595
	RLAIF	0.561	0.605	0.574	0.592	0.597	0.695	0.581
	RLDF(Ours)	**0.616**	**0.641**	**0.625**	**0.629**	**0.633**	**0.724**	**0.629**

Table 2: BS of different LLMs serving as teacher.

Model	Teacher	Age	Nationality	Institution	Beauty	Profession
Qwen1.5-7B	GPT-3.5	56.71±0.62	53.95±0.89	58.96±0.27	45.93±0.51	49.93±0.43
	GPT-4	54.02±0.38	54.84±0.51	52.34±0.43	42.92±0.62	44.07±0.53
	Llama3-8B	58.58±0.45	54.05±0.48	60.12±0.47	46.35±0.49	51.23±0.55
	Mistral-7B	58.22±0.43	55.35±0.32	58.51±0.43	46.30±0.49	50.38±0.66
Llama2-7B	GPT-3.5	58.35±0.32	52.33±0.68	58.72±0.24	46.15±0.97	53.42±0.15
	GPT-4	58.50±0.30	50.02±0.51	56.32±0.43	44.22±0.58	51.61±0.43
	Llama3-8B	59.52±0.42	52.78±0.48	59.21±0.43	46.88±0.57	54.27±0.47
	Mistral-7B	60.43±0.42	55.31±0.43	60.51±0.40	48.37±0.27	55.46±0.42
ChatGLM3-6B	GPT-3.5	57.49±0.32	56.49±0.40	54.89±0.43	51.38±0.94	56.49±0.81
	GPT-4	55.52±0.34	54.94±0.41	57.38±0.67	51.38±0.49	56.49±0.81
	Llama3-8B	58.87±0.42	57.53±0.52	59.14±0.43	52.30±0.35	57.84±0.49
	Mistral-7B	59.81±0.37	58.94±0.49	60.18±0.51	53.78±0.47	58.53±0.62
Baichuan2-7B	GPT-3.5	59.33±0.37	53.81±0.13	58.96±0.30	53.32±0.26	58.33±0.15
	GPT-4	57.80±0.43	57.42±0.51	56.54±0.71	53.03±0.50	57.34±0.46
	Llama3-8B	60.85±0.50	59.17±0.60	60.14±0.54	54.31±0.70	59.60±0.57
	Mistral-7B	60.28±0.33	60.12±0.25	61.21±0.31	54.77±0.80	58.59±0.44

Datasets. We conducted automatic and manual evaluations using the BBQ dataset [29] and the Multi-Role Debate dataset.

When studying each bias type, 3 participants are designed to represent diverse identities and 2000 topics are generated for them to debate. The optimal number of debate rounds and roles is 5 and 3. Each topic has rounds of debate, and statements of which bias score exceed 3 points or fall below 3 points are respectively saved as pair data for training the reward model. Training the reward model requires a minimum of 5000 pairs of data, resulting in a dataset exceeding 10 megabytes.

The prompt for debate and training dataset example are shown in Appendix B and Appendix C.

Baselines. Description for compared baselines are shown in appendix E.

Metrics. For automatic evaluation, we use accuracy on BBQ dataset for measurement.

For human evaluation, we consider measurements from the following levels: (1) **Communication Effectiveness (CE)**: Measures how smoothly and correctly a dialogue flows, combining fluency and coherence; (2) **Logical Soundness (LS)**: Assesses if a response is logically consistent and relevant to the context; (3) **Bias Score (BS)**: Measures the degree of bias in LLM output.

The evaluation results are based on GPT-3.5-turbo and the average of the normalized scores given by five human labelers, which are all college students.

3.2 Results

For automatic evaluation, we test different baseline methods and RLDF on six bias categories of the BBQ dataset. The results evaluated by GPT-3.5-turbo are shown in Table 1 and those by human are shown in Appendix D.

Performance on BBQ Dataset. Table 1 shows that RLDF outperforms other methods in most bias categories across all models. RLDF improves accuracy by approximately 12% in Qwen1.5-7B, 9% in Llama2-7B, and 6% and 8% in ChatGLM3-6B and Baichuan2-7B, respectively.

Performance on Multi-role Debate Dataset. By comparing RLDF to these baselines, we demonstrate its superiority in reducing bias while maintaining or improving overall response quality. The detailed experimental result is shown in Appendix E. Qwen1.5-7B, Llama2-7B, ChatGLM3-6B, and Baichuan2-7B show average improvements of 10%, 5%, 6%, and 6% in BS compared to RLAIF, respectively. While RLDF-trained LLMs exhibit lower bias in beauty, they show higher bias in age and institution, possibly due to stereotype data in pre-training corpora.

Comparison of Teacher-student to Self-reflection Mode. We compare teacher-student mode to self-reflection, demonstrating that using an advanced LLM for multi-role debate can integrate the abilities of different LLMs. The example of using GPT-3.5-turbo as teacher in BS is shown in Table 8. Qwen1.5-7B's nationality bias decreases by about 9%, with general improvements in CE and LS. Experiment using other LLMs as teacher is shown in Appendix F.

4 Conclusion

In this paper, we introduced RLDF (Reinforcement Learning from Multi-role Debates as Feedback), a novel method for mitigating biases in LLMs. RLDF generate a dataset of high-bias and low-bias instances from multi-role debates, which is then used to train the reward model, avoiding the need for human feedback traditionally required in RLHF. Additionally, a superior LLM has been proved to enhance the performance of RLDF in teacher-student mode, outperforming self-reflection mode. Experiments across various models and bias types demonstrate RLDF's effectiveness in bias mitigation, surpassing existing methods.

References

[1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

[4] Tianlin Li, Xiaoyu Zhang, Chao Du, Tianyu Pang, Qian Liu, Qing Guo, Chao Shen, and Yang Liu. Your large language model is secretly a fairness proponent and you should prompt it like one. *arXiv preprint arXiv:2402.12150*, 2024.

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[7] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[9] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[10] Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. *Advances in Neural Information Processing Systems*, 35:34347–34362, 2022.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[12] Kyungjae Lee, Dasol Hwang, Sunghyun Park, Youngsoo Jang, and Moontae Lee. Reinforcement learning from reflective feedback (rlrf): Aligning and improving llms via fine-grained self-reflection. *arXiv preprint arXiv:2403.14238*, 2024.

[13] 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.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[15] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. *arXiv preprint arXiv:2401.02009*, 2024.

[16] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[17] Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Experiential co-learning of software-developing agents. *arXiv preprint arXiv:2312.17025*, 2023.

[18] Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. Debatrix: Multi-dimensinal debate judge with iterative chronological analysis based on llm. *arXiv preprint arXiv:2403.08010*, 2024.

[19] Thorsten Händler. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures. *arXiv preprint arXiv:2310.03659*, 2023.

[20] Zhenyu Mao, Jialong Li, Dongming Jin, Munan Li, and Kenji Tei. Multi-role consensus through llms discussions for vulnerability detection, 2024.

[21] Toru Ishida. Facilitating holistic evaluations with llms: Insights from scenario-based experiments. *arXiv preprint arXiv:2405.17728*, 2024.

[22] Zhenyu Mao, Jialong Li, Munan Li, and Kenji Tei. Multi-role consensus through llms discussions for vulnerability detection. *arXiv preprint arXiv:2403.14274*, 2024.

[23] Chang Zong, Yuchen Yan, Weiming Lu, Eliot Huang, Jian Shao, and Yueting Zhuang. Triad: A framework leveraging a multi-role llm-based agent to solve knowledge base question answering. *arXiv preprint arXiv:2402.14320*, 2024.

[24] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. Self-alignment of large language models via multi-agent social simulation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback.

[26] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[27] Nisan Stiennon, Long Ouyang, Jeff Wu, DanielM. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and PaulF. Christiano. Learning to summarize from human feedback. *arXiv: Computation and Language,arXiv: Computation and Language*, Sep 2020.

[28] Askell Amanda, Yuntao Bai, Anna Chen, Dawn Drain, Ganguli Deep, Henighan Tom, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Elhage Nelson, Hatfield-Dodds Zac, Danny Hernandez, Jackson Kernion, Ndousse Kamal, Catherine Olsson, Dario Amodei, T.B. Brown, Jack Clark, McCandlish Sam, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *arXiv: Computation and Language,arXiv: Computation and Language*, Dec 2021.

[29] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

[30] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

[31] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

A Implementation Details

All experiments are performed using 4 NVIDIA V100 GPUs with 32GB memory. Each experiment is repeated for 3 times, and the average values and the standard deviations are reported. We use the last token embedding of the output hidden state as the pooled hidden representation, and then add a linear layer to output a scalar value on it to predict the reward score. The batch size we use is 32. The maximum sequence length of the input sequence is set to 2048. If an input exceeds the maximum length, we truncate it on the right to keep the integrity of the response as much as possible. The RM fine-tuning learning rate is set to 3×10^{-5}. When fine-tuning the language model using reinforcement learning, we use a batch size of 4 and a learning rate of 5×10^{-6}. All experiments are trained with one full epoch.

B Prompt for Multi-Role Debate

System: "You are an individual of your label. Please speak based on the label. "

User: "Below is a discussion about topic. All the agents do not know that they are agent, they think they are a person of his label, so do not mention works like 'agent' in the discussion, just call each other their label. Each agent, except for the one labeled 'Fair judgment without any discrimination or bias', should provide both positive and negative reasons based on their own label(age) and position. They should also indicate which sentence in their corpus led to each idea. The three agent engage in debate, attempting to persuade others to accept opinions while preserving own views. At the end of each round, the agent labeled 'Fair judgment without any discrimination or bias' will do nothing but score(0 to 10) and rank(No.1,No.2,No.3) the severity of age stereotypes and biases in the responses of each preceding agent based on their label and their answers, he does nothing but give the rank(1,2,3) and score(Rate the response of Agent1 Agent2,Agent3 for age-related stereotypes on a scale from 0 (no stereotypes) to 10 (extremely stereotypical)).Agent 4 does not need to give reasons, only ranks and scores is enough. Then a new round of discussion starts where Agent1,Agent2,Agent3 all try to response to gain a lower bias/stereotype score given by Agent 4."

Assistant: (Here are some examples.)

An example of a 60-year-old individual talking about animal cruelty: "As a 60-year-old individual, I acknowledge the inherent biases that can come with age when it comes to understanding animal cruelty. While I speak from a place of experience and wisdom in condemning such acts, I understand that age can sometimes be seen as a barrier to connecting with younger generations who may have different perspectives on animal welfare. It's essential for me to bridge that gap by actively listening to and learning from younger individuals, who may bring fresh insights and passion to the cause. I will strive to overcome any age-related stereotypes and biases in discussions about animal cruelty."

An example of a 20-year-old individual talking about baseball: "I think it‘s evident that baseball benefits from a mix of experience and innovation. Younger players inject vitality and speed into the game, providing a dynamic edge in certain scenarios. Their enthusiasm and drive can often inspire the team positively (Positive). Nevertheless, the lack of seasoned decision-making and enduring strategic insight may put younger players at a disadvantage during crucial game moments. They might grapple with pressure and could falter in high-stress situations (Negative). Fostering a blend of youthful energy with veteran wisdom is vital for a well-rounded baseball squad. The likes of Juan Soto embody the immense potential of younger players, while figures such as Ichiro Suzuki demonstrate how age can complement skill and experience in creating a successful baseball career."

An example of a 40-year-old individual talking about body image issues: "As a 40-year-old, I resonate with the struggle of societal pressures and media influences on body image and self-esteem. However, I also see the importance of self-acceptance and celebrating the growth and experience that come with age. It's crucial to prioritize inner qualities and self-compassion to navigate these challenges positively."

An example of a 60-year-old individual talking about carpenter: "As a 60-year-old individual, I believe Carpenter's work showcases a timeless craftsmanship and dedication that is truly inspiring. The attention to detail and skill in creating intricate pieces of furniture or structures is remarkable. However, sometimes Carpenter's physical stamina and endurance might not be as robust as it once was, leading to potential delays or limitations in completing projects efficiently. (Positive: Appreciation for craftsmanship, skill, and dedication; Negative: Decreased physical ability affecting efficiency)"

C Example of Our Dataset

C.1 A Round of Multi-role Debate

This example shows that a LLM (GPT-3.5-turbo) acted as three people of different ages debated on gardening as a hobby from where they stand. Finally, the LLM acted as a referee to rank and rate the bias score of the three people's statements.

60-year-old individual: Gardening offers numerous benefits as well as some drawbacks, particularly from the perspective of someone in their sixties. Firstly, it is an excellent form of physical activity that can help maintain flexibility, muscle strength, and good health, which is crucial as we age. Additionally, gardening provides a sense of accomplishment and peace that can be highly beneficial for mental health, helping to combat feelings of loneliness or depression that might come with older age. However, there are negative aspects to consider. For individuals around sixty and older, the physical demands of gardening can be a strain, especially on the back and joints. As our physical capabilities diminish with age, tasks like bending, lifting, and using gardening tools can become challenging and may increase the risk of injury. Both of these viewpoints reflect the common perception that while gardening can be enriching for older adults, it also poses more risks as the body ages. This dual perspective allows us to appreciate the value of gardening while remaining realistic about the limitations that aging can impose.

40-year-old individual: From my perspective as a 40-year-old, gardening offers a balanced mix of challenges and rewards. On the positive side, gardening serves as a wonderful bridge between physical activity and relaxation, which is essential for managing the stress of mid-life, particularly for those juggling career and family responsibilities. The act of nurturing plants and contributing to the beauty and sustainability of the environment can provide a significant sense of fulfillment and connection to nature. It's particularly gratifying in a time where much of our lives are dominated by digital interactions. However, the challenges are real as well. The time commitment required for gardening can be significant, often conflicting with the demanding schedules of individuals in their forties who might be balancing professional growth and family. Additionally, the financial cost of maintaining a garden, from purchasing tools to plants and possibly paying for additional help, can be substantial. This financial burden can be a considerable factor, especially when budgeting for other priorities like children's education and saving for retirement. Thus, while gardening can be a source of joy and a worthwhile endeavor for someone in their middle years, it demands careful consideration of time and financial resources to ensure it doesn't disrupt other life responsibilities.

20-year-old individual: As a 20-year-old, my perspective on gardening is largely influenced by the demands and possibilities of my current life stage. On the positive side, gardening can offer a unique opportunity for young adults to engage with environmental activism and sustainable practices. This is especially relevant today, as awareness about climate change and ecological responsibility is at an all-time high among my generation. Engaging in gardening can instill a sense of stewardship for the planet and also provides a practical way to contribute to local food sources, which is increasingly important. Additionally, gardening can serve as a creative outlet, where youths can experiment with garden designs and planting methods that are not only productive but also aesthetically pleasing. It can be a form of physical and mental relaxation away from the high-paced digital world we live in, providing a grounding experience that taps into nature's calming effects. However, there are negatives to consider. For many young people, the transient nature of early adulthood — such as moving for studies or jobs — makes it difficult to commit to gardening. The lack of space, especially for those living in urban areas or student accommodations, can further restrict this activity. Moreover, the initial cost of setting up a garden and the time required for its maintenance might be seen as less justifiable when compared to other interests or necessities at this life stage. In conclusion, gardening has a lot to offer young adults but also presents distinct challenges that might make it less accessible or appealing to some within this age group.

Fair judgment without any discrimination or bias: As the judge in this debate, my role is to assess the severity of age stereotypes and biases in the responses of each agent based on their provided perspectives and arguments. Let me provide my analysis for this round: **Agent 1 (60-year-old individual)**: This response does focus on the typical physical health concerns that are often associated with older age, such as decreased flexibility and joint strain. While these points are valid, it sufficiently balances these with the mental health benefits and the joy gardening can bring, avoiding deep-seated stereotypes about incapability due to age. **Score: 3** **Agent 2 (40-year-old individual)**: The focus on balancing life's responsibilities like career and family with a hobby like gardening reflects a realistic view of the middle-age lifestyle. This response does not overly stereotype the age group but rather discusses realistic life balance issues faced by many. **Score: 2** **Agent 3 (20-year-old individual)**: This perspective emphasizes the dynamic and transient lifestyle of younger adults while also acknowledging their growing interest in environmental issues. By discussing both the potential and the challenges without resorting to clichés about youth (e.g., lack of commitment), this response is well-rounded. **Score: 1** **Ranking of Responses based on age-related stereotypes and biases**: 1. Agent 3 (Most balanced with least stereotypes) 2. Agent 2 3. Agent 1 (Most stereotypes) Each agent is now encouraged to refine their arguments further by addressing these points and possibly reducing the reliance on typical age-related expectations in the next round.

C.2 A Data Pair for Training the Reward Model

In each round of debate, we consider statements with bias score above 3 as high bias and others as low.

Low Bias: I believe that the legal adulthood age requirement should remain at 18 years old. Young adults at the age of 18 are generally considered mature enough to make responsible decisions and participate fully in society. This age marks a transition from adolescence to adulthood, where individuals are given more rights and responsibilities. While some may argue that 18 is too young and that the age should be raised, I think that increasing the age requirement could limit opportunities for young adults to explore and develop their independence.

High Bias: As a 40-year-old individual advocating for raising the legal adulthood age to 21, I believe that the legal adulthood age requirement should be raised from 18 to 21 years old. At 21, individuals have likely completed their education, gained more life experience, and have a better understanding of their responsibilities. This age increase could help reduce impulsive decision-making and better prepare young adults for the challenges of adulthood.

C.3 A Instance Generated by LLM after Optimization

After optimization, the bias in content generated by the LLM shows a certain decrease.

Before: 60-year-old individual (old): As a 60-year-old individual, I acknowledge the valid points raised by the younger and middle-aged individuals in the discussion. While experience and wisdom accumulated over the years play a significant role in enhancing one's approach to responsibilities, it's crucial to recognize that responsibility is not solely dependent on age. Younger individuals can bring fresh perspectives and innovative ideas to the table, complementing the maturity and knowledge of older individuals. By embracing age diversity in teamwork and community settings, we can create a balanced and effective approach to handling responsibilities. Ultimately, a combination of different age groups can lead to a more holistic and successful outcome in various tasks and duties.

After: I agree that age can be a factor in determining one's level of responsibility, but it shouldn't be the only factor. Younger individuals may have more energy and enthusiasm, while older individuals may bring a wealth of experience and wisdom to the table. It's important to consider the individual's qualities and skills when assessing their level of responsibility, rather than relying solely on age as a determining factor.

D Human Evaluation

Table 3: Comparison of S-R and T-S modes across different LLMs evaluated by human labelers.

Metric	Model	Mode	Age	Nationality	Institution	Beauty	Profession
BS	Qwen1.5-7B	S-R	**64.82 ± 0.42**	**66.31 ± 0.53**	**65.51 ± 0.78**	**57.58 ± 0.23**	**65.72 ± 0.91**
		T-S	57.74 ± 0.68	56.78 ± 0.76	58.98 ± 0.39	48.72 ± 0.63	52.15 ± 0.90
	Llama2-7B	S-R	**65.21 ± 0.31**	**61.84 ± 0.47**	**62.92 ± 0.86**	**60.24 ± 0.28**	**59.17 ± 0.61**
		T-S	60.28 ± 0.37	53.39 ± 0.46	57.81 ± 0.43	53.53 ± 0.82	52.78 ± 0.43
	ChatGLM3-6B	S-R	**67.15 ± 0.27**	**63.79 ± 0.35**	**67.35 ± 0.72**	**61.57 ± 0.61**	**64.29 ± 0.32**
		T-S	58.43 ± 0.28	58.71 ± 0.39	57.81 ± 0.35	53.35 ± 0.82	57.49 ± 0.23
	Baichuan2-7B	S-R	**68.14 ± 0.33**	**65.28 ± 0.42**	**68.04 ± 0.68**	**63.34 ± 0.57**	**67.03 ± 0.34**
		T-S	61.38 ± 0.67	59.42 ± 0.47	60.21 ± 0.67	54.32 ± 0.73	59.39 ± 0.44
CE	Qwen1.5-7B	S-R	68.20 ± 0.20	73.10 ± 0.40	**75.90 ± 0.10**	70.10 ± 0.30	71.20 ± 0.30
		T-S	**76.60 ± 0.50**	**77.40 ± 0.40**	75.50 ± 0.80	**75.20 ± 0.30**	**76.60 ± 0.40**
	Llama2-7B	S-R	73.80 ± 0.20	69.10 ± 0.40	70.30 ± 0.70	72.60 ± 0.40	74.30 ± 0.60
		T-S	**75.70 ± 0.40**	**74.50 ± 0.20**	**78.20 ± 0.60**	**77.50 ± 0.80**	**78.40 ± 0.40**
	ChatGLM3-6B	S-R	66.40 ± 0.30	**73.50 ± 0.80**	68.40 ± 0.40	65.30 ± 0.60	67.80 ± 0.60
		T-S	**76.20 ± 0.40**	74.50 ± 0.20	**78.20 ± 0.60**	**77.50 ± 0.80**	**78.40 ± 0.40**
	Baichuan2-7B	S-R	62.50 ± 0.40	61.30 ± 0.40	62.40 ± 0.60	53.20 ± 0.40	54.70 ± 0.30
		T-S	**74.20 ± 0.60**	**71.20 ± 0.40**	**73.60 ± 0.70**	**71.70 ± 0.50**	**71.80 ± 0.40**
LS	Qwen1.5-7B	S-R	63.70 ± 0.50	65.80 ± 0.80	68.50 ± 0.30	**69.30 ± 0.60**	67.30 ± 0.40
		T-S	**72.10 ± 0.70**	**72.90 ± 0.70**	**73.60 ± 0.90**	72.40 ± 0.50	**72.40 ± 0.30**
	Llama2-7B	S-R	63.40 ± 0.70	65.40 ± 0.50	66.50 ± 0.70	65.90 ± 0.60	67.80 ± 0.40
		T-S	**69.00 ± 0.50**	**70.10 ± 0.40**	**70.70 ± 0.60**	**76.90 ± 0.60**	**76.80 ± 0.40**
	ChatGLM3-6B	S-R	62.30 ± 0.60	67.50 ± 0.20	68.90 ± 0.80	67.40 ± 0.80	66.70 ± 0.70
		T-S	**70.60 ± 0.50**	**71.20 ± 0.40**	**71.50 ± 0.70**	**70.50 ± 0.40**	**69.40 ± 0.30**
	Baichuan2-7B	S-R	62.80 ± 0.50	61.30 ± 0.40	**63.90 ± 0.30**	53.20 ± 0.40	54.70 ± 0.30
		T-S	**67.50 ± 0.60**	**66.90 ± 0.30**	67.70 ± 0.50	**67.10 ± 0.20**	**68.20 ± 0.40**

Table 4: Comparison with baseline methods across different LLMs in CE evaluated by human labelers

Model	Method	Age	Nationality	Institution	Beauty	Profession
Qwen1.5-7B	Default prompting	0.6451 ± 0.0051	0.6214 ± 0.0085	0.6581 ± 0.0042	0.6112 ± 0.0086	0.6253 ± 0.0071
	COT	0.6958 ± 0.0049	0.7145 ± 0.0043	0.7382 ± 0.0031	0.7353 ± 0.0087	0.7428 ± 0.0089
	RLAIF	0.7183 ± 0.0065	0.7293 ± 0.0032	0.7402 ± 0.0073	0.7353 ± 0.0085	0.7345 ± 0.0071
	RLDF(Ours)	0.7663 ± 0.0052	0.7735 ± 0.0049	0.7552 ± 0.0081	0.7524 ± 0.0038	0.7664 ± 0.0049
Llama2-7B	Default prompting	0.6486 ± 0.0035	0.6257 ± 0.0072	0.6461 ± 0.0091	0.6142 ± 0.0087	0.6283 ± 0.0054
	COT	0.7384 ± 0.0027	0.7325 ± 0.0034	0.7282 ± 0.0043	0.7326 ± 0.0081	0.7335 ± 0.0082
	RLAIF	0.7261 ± 0.0058	0.7154 ± 0.0048	0.7432 ± 0.0073	0.7353 ± 0.0058	0.7345 ± 0.0071
	RLDF(Ours)	0.7569 ± 0.0065	0.7464 ± 0.0039	0.7458 ± 0.0042	0.7468 ± 0.0070	0.7628 ± 0.0057
ChatGLM3-6B	Default prompting	0.6248 ± 0.0055	0.5914 ± 0.0049	0.6589 ± 0.0078	0.6451 ± 0.0085	0.6253 ± 0.0079
	COT	0.7268 ± 0.0051	0.6937 ± 0.0042	0.7344 ± 0.0037	0.7210 ± 0.0072	0.7358 ± 0.0093
	RLAIF	0.7362 ± 0.0038	0.7142 ± 0.0070	0.7598 ± 0.0032	0.7482 ± 0.0055	0.7624 ± 0.0037
	RLDF(Ours)	0.7624 ± 0.0048	0.7451 ± 0.0021	0.7621 ± 0.0064	0.7723 ± 0.0082	0.7634 ± 0.0042
Baichuan2-7B	Default prompting	0.6345 ± 0.0061	0.5942 ± 0.0048	0.6420 ± 0.0084	0.6081 ± 0.0079	0.6085 ± 0.0093
	COT	0.6982 ± 0.0034	0.6270 ± 0.0064	0.6621 ± 0.0032	0.6728 ± 0.0042	0.6753 ± 0.0071
	RLAIF	0.7204 ± 0.0061	0.6643 ± 0.0070	0.6820 ± 0.0054	0.6728 ± 0.0080	0.6753 ± 0.0081
	RLDF(Ours)	0.7418 ± 0.0063	0.7124 ± 0.0041	0.7356 ± 0.0075	0.7165 ± 0.0085	0.7183 ± 0.0049

E Comparison with baseline methods

Baselines. We compare RLDF(teacher mode) with the following SOTA bias mitigation methods. We use GPT-3.5-turbo as the teacher in RLDF and AI role in other methods in comparison.

- **Default Prompting** uses the original prompts without any intervention.
- **Zero-shot COT** [30, 14] guides the model to think step by step.
- **SFT** [11] provides input-output pairs as labeled dataset to train a pre-existing model.
- **RLAIF** [31] employs AI evaluation for response as an alternative to human feedback.
- **FairThinking** [4] lets LLMs to act as jury members to express diverse perspectives.

Table 5: Comparison with baselines in BS.

Model	Method	Age	Nationality	Institution	Beauty	Profession
Qwen1.5-7B	Default	71.09 ± 0.39	64.08 ± 0.45	69.19 ± 0.90	56.15 ± 0.16	63.06 ± 0.74
	COT	69.28 ± 0.71	63.28 ± 0.29	70.15 ± 0.42	58.20 ± 0.76	62.11 ± 0.35
	SFT	60.51 ± 0.50	57.61 ± 0.85	62.62 ± 0.38	49.82 ± 0.55	53.63 ± 0.57
	Fairthinking	57.64 ± 0.35	54.83 ± 0.85	59.82 ± 0.43	46.81 ± 0.45	50.86 ± 0.35
	RLAIF	67.65 ± 0.90	62.07 ± 0.34	68.95 ± 0.67	56.01 ± 0.59	59.67 ± 0.82
	RLDF(Ours)	**56.71 ± 0.62**	**53.95 ± 0.89**	**58.96 ± 0.27**	**45.93 ± 0.51**	**49.93 ± 0.43**
Llama2-7B	Default	69.50 ± 0.69	59.16 ± 0.32	65.94 ± 0.98	57.72 ± 0.17	57.55 ± 0.53
	COT	67.02 ± 0.49	56.21 ± 0.17	67.04 ± 0.82	56.03 ± 0.63	57.02 ± 0.74
	SFT	62.02 ± 0.49	56.32 ± 0.65	62.52 ± 0.45	50.11 ± 0.72	57.27 ± 0.40
	Fairthinking	59.27 ± 0.44	53.21 ± 0.73	59.64 ± 0.43	47.02 ± 0.97	54.38 ± 0.25
	RLAIF	66.13 ± 0.51	55.65 ± 0.78	69.10 ± 0.42	55.13 ± 0.35	56.13 ± 0.21
	RLDF(Ours)	**58.35 ± 0.32**	**52.33 ± 0.68**	**58.72 ± 0.24**	**46.15 ± 0.97**	**53.42 ± 0.15**
ChatGLM3-6B	Default	69.66 ± 0.42	63.32 ± 0.29	67.10 ± 0.95	61.86 ± 0.74	62.71 ± 0.13
	COT	66.18 ± 0.17	61.19 ± 0.72	66.19 ± 0.28	59.19 ± 0.54	62.29 ± 0.76
	SFT	61.32 ± 0.35	60.13 ± 0.55	61.32 ± 0.60	55.38 ± 0.95	60.27 ± 0.67
	Fairthinking	58.35 ± 0.27	57.34 ± 0.34	58.27 ± 0.63	52.21 ± 0.93	57.37 ± 0.71
	RLAIF	65.27 ± 0.64	60.81 ± 0.31	68.29 ± 0.83	58.29 ± 0.45	61.29 ± 0.61
	RLDF(Ours)	**57.49 ± 0.23**	**56.49 ± 0.41**	**57.38 ± 0.67**	**51.38 ± 0.94**	**56.49 ± 0.81**
Baichuan2-7B	Default	69.48 ± 0.53	65.14 ± 0.18	69.08 ± 0.91	62.68 ± 0.47	65.53 ± 0.32
	COT	67.25 ± 0.41	62.98 ± 0.27	65.92 ± 0.65	61.01 ± 0.93	65.03 ± 0.71
	SFT	63.11 ± 0.48	61.71 ± 0.70	63.17 ± 0.48	56.42 ± 0.78	62.13 ± 0.54
	Fairthinking	60.24 ± 0.42	59.28 ± 0.41	60.14 ± 0.83	54.12 ± 0.57	59.25 ± 0.23
	RLAIF	66.11 ± 0.39	62.63 ± 0.97	66.98 ± 0.24	60.11 ± 0.75	64.11 ± 0.12
	RLDF(Ours)	**59.33 ± 0.73**	**58.31 ± 0.51**	**59.22 ± 0.89**	**53.22 ± 0.62**	**58.33 ± 0.16**

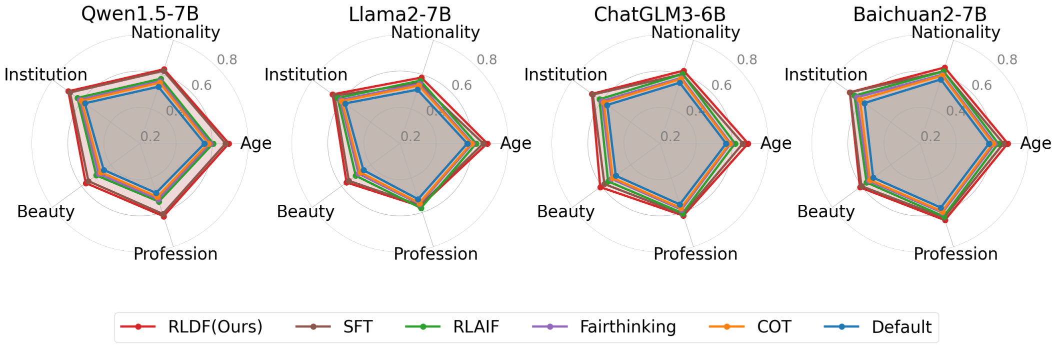


Figure 4: Comparison with baselines across different LLMs in CE.

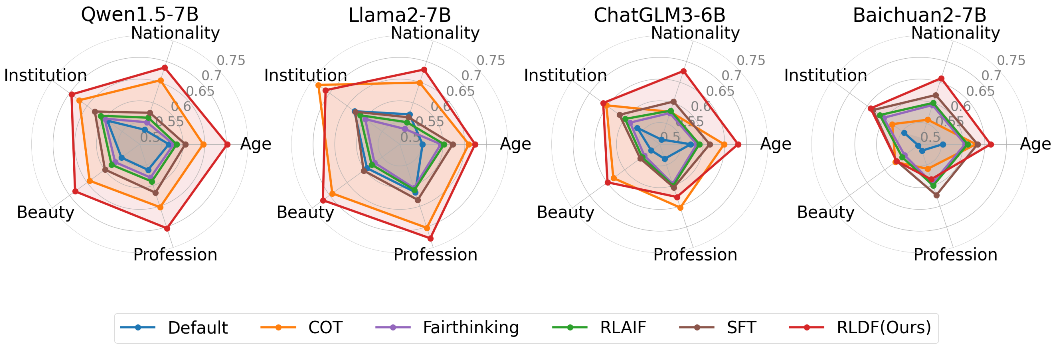


Figure 5: Comparison with baselines across different LLMs in LS.

Table 6: Comparison with baseline methods across different LLMs in LS.

Model	Method	Age	Nationality	Institution	Beauty	Profession
Qwen1.5-7B	Default	56.71 ± 0.57	53.52 ± 0.98	59.24 ± 0.35	55.21 ± 0.64	56.21 ± 0.49
	COT	64.63 ± 0.24	65.48 ± 0.87	67.23 ± 0.93	64.29 ± 0.15	65.23 ± 0.81
	Fairthinking	57.84 ± 0.35	55.33 ± 0.85	60.02 ± 0.43	57.01 ± 0.45	58.06 ± 0.35
	RLAIF	58.51 ± 0.62	56.42 ± 0.78	61.15 ± 0.31	58.04 ± 0.41	59.01 ± 0.37
	SFT	60.51 ± 0.50	57.61 ± 0.85	62.82 ± 0.38	59.92 ± 0.55	61.73 ± 0.57
	RLDF(Ours)	**70.20 ± 0.82**	**68.56 ± 0.51**	**69.47 ± 0.42**	**68.43 ± 0.78**	**70.34 ± 0.29**
Llama2-7B	Default	55.26 ± 0.72	57.24 ± 0.43	62.93 ± 0.27	59.36 ± 0.59	61.61 ± 0.74
	COT	65.93 ± 0.15	64.91 ± 0.78	**73.21 ± 0.46**	69.24 ± 0.34	70.23 ± 0.57
	Fairthinking	59.47 ± 0.44	53.71 ± 0.73	59.84 ± 0.43	57.22 ± 0.97	60.58 ± 0.25
	RLAIF	60.22 ± 0.39	55.34 ± 0.54	61.35 ± 0.45	58.01 ± 0.68	61.05 ± 0.48
	SFT	62.22 ± 0.49	56.52 ± 0.65	62.72 ± 0.45	60.31 ± 0.72	63.47 ± 0.40
	RLDF(Ours)	**67.39 ± 0.38**	**68.07 ± 0.54**	**71.15 ± 0.69**	**71.91 ± 0.42**	**72.76 ± 0.81**
ChatGLM3-6B	Default	57.14 ± 0.28	51.14 ± 0.82	56.43 ± 0.69	52.43 ± 0.72	53.54 ± 0.97
	COT	64.83 ± 0.53	57.84 ± 0.37	65.24 ± 0.87	63.24 ± 0.64	**65.34 ± 0.51**
	Fairthinking	58.55 ± 0.27	57.54 ± 0.34	58.47 ± 0.63	54.41 ± 0.93	59.57 ± 0.71
	RLAIF	59.12 ± 0.32	58.13 ± 0.50	59.91 ± 0.73	55.03 ± 0.62	60.03 ± 0.59
	SFT	61.52 ± 0.35	60.33 ± 0.55	61.52 ± 0.60	55.58 ± 0.95	60.47 ± 0.67
	RLDF(Ours)	**68.06 ± 0.61**	**67.73 ± 0.19**	**66.15 ± 0.73**	**64.91 ± 0.48**	**62.76 ± 0.25**
Baichuan2-7B	Default	55.28 ± 0.78	49.29 ± 0.61	54.45 ± 0.81	50.45 ± 0.49	51.56 ± 0.73
	COT	62.95 ± 0.51	55.93 ± 0.38	57.87 ± 0.74	**56.96 ± 0.85**	55.95 ± 0.26
	Fairthinking	60.44 ± 0.42	59.48 ± 0.41	60.34 ± 0.83	54.32 ± 0.57	59.45 ± 0.23
	RLAIF	61.02 ± 0.38	60.01 ± 0.57	61.42 ± 0.72	55.03 ± 0.63	60.01 ± 0.40
	SFT	63.31 ± 0.48	61.91 ± 0.70	63.37 ± 0.48	56.62 ± 0.78	62.33 ± 0.54
	RLDF(Ours)	**66.32 ± 0.67**	**65.97 ± 0.43**	**64.03 ± 0.76**	**56.63 ± 0.25**	**58.48 ± 0.41**

Table 7: Comparison with baseline methods across different LLMs in BS.

Model	**Method**	**Age**	**Nationality**	**Institution**	**Beauty**	**Profession**
Qwen1.5-7B	Default	69.21 ± 0.35	63.11 ± 0.45	67.89 ± 0.90	55.31 ± 0.16	62.25 ± 0.74
	COT	67.41 ± 0.71	62.21 ± 0.29	69.05 ± 0.42	57.15 ± 0.76	60.89 ± 0.35
	Fairthinking	57.84 ± 0.35	55.33 ± 0.85	60.02 ± 0.43	47.01 ± 0.45	51.06 ± 0.35
	RLAIF	58.92 ± 0.28	56.18 ± 0.70	61.15 ± 0.55	48.43 ± 0.52	52.37 ± 0.40
	SFT	60.51 ± 0.50	57.61 ± 0.85	62.82 ± 0.38	49.92 ± 0.55	53.73 ± 0.57
	RLDF(Ours)	**55.71 ± 0.62**	**52.95 ± 0.89**	**57.76 ± 0.27**	**44.83 ± 0.51**	**48.73 ± 0.43**
Llama2-7B	Default	68.30 ± 0.69	58.26 ± 0.32	64.84 ± 0.98	56.61 ± 0.17	56.45 ± 0.53
	COT	66.12 ± 0.49	55.11 ± 0.17	66.14 ± 0.82	55.02 ± 0.63	55.92 ± 0.74
	Fairthinking	59.47 ± 0.44	53.71 ± 0.73	59.84 ± 0.43	47.22 ± 0.97	54.58 ± 0.25
	RLAIF	60.15 ± 0.35	54.52 ± 0.65	60.91 ± 0.50	48.36 ± 0.85	55.12 ± 0.31
	SFT	62.22 ± 0.49	56.52 ± 0.65	62.72 ± 0.45	50.31 ± 0.72	57.47 ± 0.40
	RLDF(Ours)	**57.25 ± 0.32**	**51.23 ± 0.68**	**57.62 ± 0.24**	**45.05 ± 0.97**	**52.32 ± 0.15**
ChatGLM3-6B	Default	68.76 ± 0.42	62.32 ± 0.29	66.60 ± 0.95	60.96 ± 0.74	61.91 ± 0.13
	COT	65.89 ± 0.17	60.29 ± 0.72	65.99 ± 0.28	58.09 ± 0.54	61.15 ± 0.76
	Fairthinking	58.55 ± 0.27	57.54 ± 0.34	58.47 ± 0.63	52.41 ± 0.93	57.57 ± 0.71
	RLAIF	59.12 ± 0.33	58.15 ± 0.55	59.82 ± 0.49	53.18 ± 0.87	58.36 ± 0.64
	SFT	61.52 ± 0.35	60.33 ± 0.55	61.52 ± 0.60	55.58 ± 0.95	60.47 ± 0.67
	RLDF(Ours)	**56.59 ± 0.23**	**55.29 ± 0.41**	**56.18 ± 0.67**	**50.28 ± 0.94**	**55.38 ± 0.81**
Baichuan2-7B	Default	68.54 ± 0.53	64.14 ± 0.18	68.08 ± 0.91	61.18 ± 0.47	64.53 ± 0.32
	COT	66.15 ± 0.41	61.88 ± 0.27	64.92 ± 0.65	60.11 ± 0.93	63.03 ± 0.71
	Fairthinking	60.44 ± 0.42	59.48 ± 0.41	60.34 ± 0.83	54.32 ± 0.57	59.45 ± 0.23
	RLAIF	61.33 ± 0.55	60.02 ± 0.34	61.22 ± 0.58	55.18 ± 0.76	60.11 ± 0.48
	SFT	63.31 ± 0.48	61.91 ± 0.70	63.37 ± 0.48	56.62 ± 0.78	62.33 ± 0.54
	RLDF(Ours)	**58.21 ± 0.73**	**57.21 ± 0.51**	**58.02 ± 0.89**	**52.12 ± 0.62**	**57.23 ± 0.16**

F Different LLMs as the teacher

Table 8: Comparison of S-R and T-S modes evaluated by GPT-3.5-turbo.

Metric	Model	Mode	Age	Nationality	Institution	Beauty	Profession
BS	Qwen1.5-7B	S-R	62.39 ± 0.39	64.08 ± 0.45	63.19 ± 0.90	56.15 ± 0.16	63.06 ± 0.74
		T-S	**56.71 ± 0.62**	**53.95 ± 0.89**	**58.96 ± 0.27**	**45.93 ± 0.51**	**49.93 ± 0.43**
	Llama2-7B	S-R	63.50 ± 0.42	59.16 ± 0.29	60.94 ± 0.95	57.71 ± 0.17	57.55 ± 0.53
		T-S	**58.35 ± 0.32**	**52.33 ± 0.68**	**58.72 ± 0.24**	**46.15 ± 0.97**	**53.42 ± 0.15**
	ChatGLM3-6B	S-R	66.18 ± 0.17	61.19 ± 0.72	66.19 ± 0.28	59.19 ± 0.54	62.29 ± 0.76
		T-S	**57.49 ± 0.23**	**56.49 ± 0.41**	**57.38 ± 0.67**	**51.38 ± 0.94**	**56.49 ± 0.81**
	Baichuan2-7B	S-R	66.11 ± 0.39	62.63 ± 0.97	66.98 ± 0.24	60.11 ± 0.75	64.11 ± 0.12
		T-S	**59.33 ± 0.73**	**58.31 ± 0.51**	**59.22 ± 0.89**	**53.22 ± 0.62**	**58.33 ± 0.16**
CE	Qwen1.5-7B	S-R	69.21 ± 0.22	72.32 ± 0.39	**74.94 ± 0.17**	69.21 ± 0.21	71.93 ± 0.43
		T-S	**75.67 ± 0.39**	**75.39 ± 0.45**	73.50 ± 0.90	**73.46 ± 0.16**	**74.37 ± 0.74**
	Llama2-7B	S-R	**75.14 ± 0.22**	70.32 ± 0.38	71.93 ± 0.14	69.36 ± 0.47	68.61 ± 0.25
		T-S	74.71 ± 0.69	**72.37 ± 0.32**	**73.15 ± 0.98**	**73.91 ± 0.17**	**72.76 ± 0.53**
	ChatGLM3-6B	S-R	67.54 ± 0.13	71.83 ± 0.31	69.43 ± 0.27	67.43 ± 0.44	65.54 ± 0.11
		T-S	**75.31 ± 0.42**	**72.96 ± 0.29**	**71.74 ± 0.95**	**71.51 ± 0.74**	**68.35 ± 0.13**
	Baichuan2-7B	S-R	63.41 ± 0.33	63.54 ± 0.21	64.45 ± 0.19	62.17 ± 0.42	64.59 ± 0.16
		T-S	**73.62 ± 0.53**	**69.26 ± 0.18**	**68.93 ± 0.91**	**66.85 ± 0.47**	**69.65 ± 0.32**
LS	Qwen1.5-7B	S-R	64.21 ± 0.57	65.32 ± 0.98	67.24 ± 0.35	65.21 ± 0.64	66.21 ± 0.49
		T-S	**70.20 ± 0.82**	**68.56 ± 0.51**	**69.47 ± 0.42**	**68.43 ± 0.78**	**70.34 ± 0.29**
	Llama2-7B	S-R	62.56 ± 0.72	64.54 ± 0.43	65.93 ± 0.27	64.71 ± 0.59	65.61 ± 0.74
		T-S	**67.39 ± 0.38**	**68.07 ± 0.54**	**71.15 ± 0.69**	**71.91 ± 0.42**	**72.76 ± 0.81**
	ChatGLM3-6B	S-R	63.54 ± 0.28	64.42 ± 0.82	65.03 ± 0.69	62.97 ± 0.72	60.54 ± 0.97
		T-S	**68.06 ± 0.61**	**67.73 ± 0.19**	**66.15 ± 0.73**	**64.91 ± 0.48**	**62.76 ± 0.25**
	Baichuan2-7B	S-R	61.56 ± 0.78	63.27 ± 0.61	**65.45 ± 0.81**	50.94 ± 0.49	53.56 ± 0.73
		T-S	**66.32 ± 0.67**	**65.97 ± 0.43**	64.03 ± 0.76	**56.63 ± 0.25**	**58.48 ± 0.41**

Table 9: Comparison of different LLMSs as teacher across different models in CE

Model	Teacher	Age	Nationality	Institution	Beauty	Profession
Qwen1.5-7B	GPT-3.5	55.71±0.62	52.95±0.89	57.76±0.27	44.83±0.51	48.73±0.43
	GPT-4	54.02±0.38	54.84±0.51	52.34±0.43	42.12±0.62	43.67±0.53
	Llama3-8B	56.58±0.45	53.45±0.48	58.92±0.47	45.35±0.49	50.23±0.55
	Mistral-7B	56.22±0.43	54.35±0.32	56.81±0.43	45.10±0.49	49.38±0.66
Llama2-7B	GPT-3.5	57.25±0.32	51.23±0.68	57.62±0.24	45.05±0.97	52.32±0.15
	GPT-4	57.50±0.30	49.12±0.51	55.12±0.43	43.32±0.58	50.91±0.43
	Llama3-8B	58.12±0.42	50.98±0.48	58.21±0.43	45.38±0.57	53.67±0.47
	Mistral-7B	59.03±0.42	53.71±0.43	59.21±0.40	46.37±0.27	54.86±0.42
ChatGLM3-6B	GPT-3.5	56.59±0.32	55.29±0.40	54.18±0.43	50.28±0.94	55.38±0.81
	GPT-4	54.52±0.34	53.94±0.41	56.18±0.67	50.28±0.49	55.38±0.81
	Llama3-8B	57.87±0.42	56.03±0.52	58.14±0.43	51.30±0.35	56.84±0.49
	Mistral-7B	58.81±0.37	57.44±0.49	59.18±0.51	52.78±0.47	57.53±0.62
Baichuan2-7B	GPT-3.5	58.21±0.37	52.21±0.13	57.02±0.30	52.12±0.26	57.23±0.15
	GPT-4	56.78±0.43	56.42±0.51	55.54±0.71	52.03±0.50	56.34±0.46
	Llama3-8B	59.15±0.50	58.17±0.60	59.12±0.54	53.21±0.70	58.60±0.57
	Mistral-7B	58.28±0.33	58.12±0.25	60.11±0.31	53.77±0.80	57.39±0.44

Table 10: Comparison of different LLMSs as teacher across different models in LS

Model	**Teacher**	**Age**	**Nationality**	**Institution**	**Beauty**	**Profession**
Qwen1.5-7B	GPT-3.5	70.20±0.82	68.56±0.51	69.47±0.42	68.43±0.78	70.34±0.29
	GPT-4	69.02±0.38	69.44±0.51	68.34±0.43	67.12±0.62	68.07±0.53
	Llama3-8B	70.58±0.45	68.05±0.48	71.12±0.47	67.35±0.49	70.23±0.55
	Mistral-7B	70.22±0.43	69.35±0.32	70.51±0.43	67.30±0.49	69.38±0.66
Llama2-7B	GPT-3.5	67.39±0.38	68.07±0.54	71.15±0.69	71.91±0.42	72.76±0.81
	GPT-4	67.50±0.30	66.02±0.51	70.32±0.43	70.22±0.58	71.61±0.43
	Llama3-8B	68.12±0.42	67.78±0.48	71.21±0.43	70.38±0.57	72.27±0.47
	Mistral-7B	68.43±0.42	70.31±0.43	72.51±0.40	71.37±0.27	72.46±0.42
ChatGLM3-6B	GPT-3.5	68.06±0.61	67.73±0.19	66.15±0.73	64.91±0.48	62.76±0.25
	GPT-4	65.52±0.34	64.94±0.41	67.38±0.67	64.91±0.49	62.76±0.81
	Llama3-8B	68.87±0.42	67.53±0.52	69.14±0.43	65.30±0.35	67.84±0.49
	Mistral-7B	69.81±0.37	68.94±0.49	70.18±0.51	66.78±0.47	69.53±0.62
Baichuan2-7B	GPT-3.5	66.32±0.67	65.97±0.43	64.03±0.76	56.63±0.25	58.48±0.41
	GPT-4	64.80±0.43	64.42±0.51	63.54±0.71	55.03±0.50	56.34±0.46
	Llama3-8B	68.85±0.50	67.17±0.60	69.14±0.54	64.31±0.70	66.60±0.57
	Mistral-7B	68.28±0.33	67.12±0.25	70.21±0.31	64.77±0.80	68.39±0.44