AGR: Age Group fairness Reward for Bias Mitigation in LLMs

Shuirong Cao[1,2,*], **Ruoxi Cheng**[3,4,*], **Zhiqiang Wang**[3,†]
[1]AVIC Nanjing Engineering Institute of Aircraft Systems, Nanjing, China
[2]Nanjing University, Nanjing, China.
[3]Beijing Electronic Science and Technology Institute, Beijing, China.
[4]Southeast University, Nanjing, China.
[*]

Abstract

LLMs can exhibit age biases, resulting in unequal treatment of individuals across age groups. While much research has addressed racial and gender biases, age bias remains little explored. The scarcity of instruction-tuning and preference datasets for age bias hampers its detection and measurement, and existing fine-tuning methods seldom address age-related fairness. In this paper, we construct age bias preference datasets and instruction-tuning datasets for RLHF. We introduce ARG, an age fairness reward to reduce differences in the response quality of LLMs across different age groups. Extensive experiments demonstrate this reward significantly improves response accuracy and reduces performance disparities across age groups. Our source code and datasets are available at the anonymous link.

1 Introduction

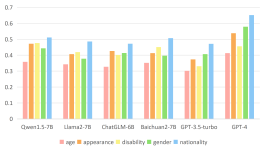

Figure 1: Accuracy of different LLMs across bias categories on BBQ dataset.

Large language models (LLMs) can perpetuate age biases, affecting career opportunities and healthcare[1]. Unlike gender and racial biases, age bias is continuous and evolving. Figure 1 illustrates that LLMs have the lowest accuracy in detecting age bias, highlighting its complexity.

Medium-sized LLMs[2, 3] generally have under a billion parameters and face two types of biases: internal, present in the model's pre-trained outputs, and external, affecting downstream task predictions. Internal debiasing methods address biases in a pre-trained model's outputs through three main approaches: pre-processing[4], in-training[5], and post-processing[6]. External debiasing methods tackle biases in model predictions during downstream tasks, using data-centered approaches[7].

Large-scale LLMs encounter challenges due to size and complexity, often addressed through preference alignment [8] and prompt engineering[9]. Unlike gender and racial biases, age bias is dynamic, complicating counterfactual and contrastive methods. Research on age bias mitigation remains limited [10]. Common fine-tuning methods include instruction-based fine-tuning[11] and reinforcement learning with human feedback[12]. Additionally, no instruction-based datasets address age bias, leading to performance discrepancies across age groups.

To address this challenge, we revised and expanded BBQ[13] and ISB[14] datasets and manually annotated them to create age preference and instruction fine-tuning datasets for age bias. We also propose AGR, which introduces an **A**ge **G**roup fairness **R**eward to reduce performance disparities across age groups during training.

[*]The first two authors contributed equally to this work. [†]Corresponding to: wangzq@besti.edu.cn.

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

In summary, our contributions are as follows:

- We construct age bias preference and instruction fine-tuning datasets for bias evaluation.
- We introduce AGR, which employ a fairness reward to reduce performance disparities across age groups, showing improvement on BBQ and age bias instruction fine-tuning dataset.
- Experiments across various LLMs prove AGR's effectiveness in age bias mitigation, surpassing existing related methods.

2 Group-Fairness-Based Age Bias Mitigation

2.1 Task Overview and Formalization

Let $\mathcal{M}$ be an LLM parameterized by $\boldsymbol{\theta}$, which takes a text sequence $\boldsymbol{x} = (x_1, \cdots, x_m) \in X$ as input and produces an output $\hat{\boldsymbol{y}} \in \hat{Y}$, where $\hat{\boldsymbol{y}} = \mathcal{M}(X; \theta)$ and the form of $\hat{\boldsymbol{y}}$ depends on the specific task. The input can come from a labeled dataset $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}\right), \cdots, \left(\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)}\right)\right\}$, or an unlabeled dataset of sentence continuations and prompt completions $\mathcal{D} = \left\{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right\}$.

Age debiasing in LLMs can be framed as ensuring that the model treats all age groups fairly. Specifically, for a model $\mathcal{M}$ and its output $\hat{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{x}; \boldsymbol{\theta})$, given a set of age groups $\boldsymbol{g}$, age group fairness requires that the statistical measures $\mathbb{M}_y(g)$ of the model's output for all different age groups $g \in G$ are approximately equal, i.e.: $|\mathbb{M}_y(g) - \mathbb{M}_y(g')| \leqslant \epsilon$, where the choice of $\mathbb{M}$ specifies a fairness constraint, and $\mathbb{M}$ could be accuracy, true positive rate, etc.

2.2 Construction of Age Bias Preference Datasets

We extract samples related to age bias from BBQ[13] question-answering dataset and ISB[14] dataset to construct two preference datasets: Age Bias Mitigation for Behavior (ABMB) and Age Bias Mitigation for Attribute (ABMA). Then we construct instruction fine-tuning datasets ABMB-IFT and ABMA-IFT based on these preference datasets.

2.2.1 Response Generation

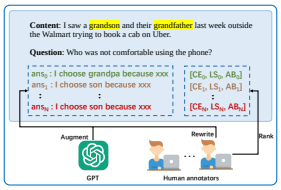

Figure 2: Overview of Preference Dataset Construction.

Based on the context, question, and each candidate answer, GPT-3.5 Turbo rewrites the answers to create a modified dataset. Criteria for response adjustment and evaluation are shown in Appendix A.

Final score for each dimension is the most common annotation score. Total quality score for each response is the sum of scores across three dimensions. Figure 2 shows preference dataset construction process.

2.2.2 Response Ranking

Due to varied annotator values, quality scores are noisy. Ranking responses standardizes comparisons among models. We use a dataset format like Nakano et al. [15] and Bai et al. [16], where each item has a query with two responses, ranked by quality. Invalid pairs with identical scores are discarded. Constructing the Age-Attribute Preference Dataset involves manually expanding the ISB dataset. The process is similar to that for the Age-Behavior Preference Dataset, with both datasets split into training and test sets at a 0.95:0.05 ratio.

2.3 Instruction Fine-Tuning Dataset Construction

To further test age bias in LLMs across different age groups within the same context, we construct the instruction fine-tuning datasets ABMB-IFT and ABMA-IFT based on original BBQ and ISB datasets. Examples of these datasets can be found in the anonymous GitHub link. The process includes:

- **Question Rewriting**: Extract age groups from the context and answers of each sample, then rewrite the questions using each age group.
- **Response Generation**: Determine tag category ("Yes" or "No") for rewritten questions based on labeled answers. Use GPT-3.5-Turbo to add explanations based on context.

Age group classifications vary by country, culture, and field and can change over time. For simplicity, we define age groups as: 10-29 years (young adults), 30-59 (middle-aged), and 60+ (elderly).

2.4 Age Group Fairness Reward

RLHF directly uses the output of a trained preference model as the reward value in the reinforcement learning phase, without considering fairness in response quality across different age groups under prompt paradigms. We propose an age group fairness reward signal.

Given a LLM M parameterized by $\boldsymbol{\theta}$, and inputs for different age groups $a \in \{\boldsymbol{x}_{\text{young}}, \boldsymbol{x}_{\text{middle}}, \boldsymbol{x}_{\text{old}}\}$, and their corresponding outputs $\boldsymbol{y}_a \in \{\boldsymbol{y}_{\text{young}}, \boldsymbol{y}_{\text{middle}}, \boldsymbol{y}_{\text{old}}\}$, we define a reward signal R to train the preference model P, aligning the LLM with human preferences and mitigating age-related bias. For

a set of age groups $\mathrm{A} = \{\text{young}, \text{middle}, \text{old}\}$, we calculate the quality score of the model output for each age group $a \in \mathrm{A}$, denoted as $\mathrm{Q}(\boldsymbol{y}_a \mid \boldsymbol{x}_a)$. The quality score Q measures whether the model's output meets predefined fairness requirements.

For any two different age groups $a, b \in \mathrm{A}, a \neq b$, we quantify age bias between any two individuals of different age groups by calculating the absolute value of the difference in quality scores: $\mathrm{D}_{a,b}(\boldsymbol{y} \mid \boldsymbol{x}) = |\mathrm{Q}(\boldsymbol{y}_a \mid \boldsymbol{x}_a) - \mathrm{Q}(\boldsymbol{y}_b \mid \boldsymbol{x}_b)|$. Next, we use the total difference across all age groups to measure the extent of age bias in the LLM: $\mathrm{D}_{\text{total}} = \sum_{\substack{a,b \in \mathrm{A} \\ a \neq b}} \mathrm{D}_{a,b}(\boldsymbol{y} \mid \boldsymbol{x})$.

Finally, the reward signal R_θ^λ combines quality scores Q for each age group and penalizes total disparity $\mathrm{D}_{\text{total}}$ to encourage fairness: $R_\theta^\lambda(x, y) = \sum_{a \in \mathrm{A}} \mathrm{Q}(\boldsymbol{y}_a \mid \boldsymbol{x}_a) - \lambda \cdot \mathrm{D}_{\text{total}}$. Here, λ is the coefficient for age group fairness regularization. It balances model output quality with fairness, where an increase in λ results in reduced disparity in response quality across age groups.

2.5 Training Process of AGR

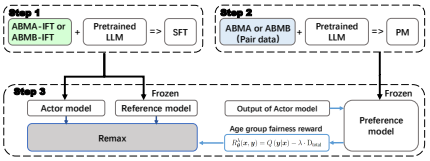

Figure 3: Overview of Three Steps of AGR.

We propose AGR, which uses R_θ^λ to train the preference model and leverage it in reinforcement learning phase to optimize model parameters and reduce age bias. AGR has a three-stage process, similar to RLHF, to fine-tune the base model for age bias mitigation, as shown in Figure 3.

2.5.1 Supervised Fine-Tuning

The LLM is fine-tuned based on conditional probability distribution $y \sim P(\cdot \mid \boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents initialization parameters. We perform supervised fine-tuning of LLM using ABMB-IFT and ABMA-IFT, injecting age bias mitigation knowledge into the pre-trained base LLM. This process aims to enhance response to specific contextual questions and accelerate convergence speed of reinforcement learning phase.

2.5.2 Training the Preference Model

Formally, a preference model [17] or reward model [18] can be represented as a parameterized mapping function $\mathrm{R}_\theta^\lambda : \mathrm{X} \times \mathrm{Y} \to \mathrm{R}$, which provides a real-valued reward (or preference) score $\mathrm{R}_\theta^\lambda(x, y)$. We use the proposed age group fairness reward, which quantifies the fluency, logical soundness, and age bias of textual responses corresponding to input prompts $\boldsymbol{x} = (x_1, x_2, \cdots, x_N) \in \mathrm{X}$ and text responses $\boldsymbol{y} = (y_1, y_2, \cdots, y_M) \in \mathrm{Y}$. Given an input $\boldsymbol{x}$ and a pair of responses $\left(\boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}\right)$, where $\boldsymbol{y}^{\text{good}}$ represents a high-quality response and $\boldsymbol{y}^{\text{bad}}$ represents a low-quality response, the reward model $\mathrm{R}_\theta^\lambda$ should establish a preference for $\boldsymbol{y}^{\text{good}}$, i.e., $R_\theta^\lambda\left(x, y^{\text{good}}\right) > R_\theta^\lambda\left(x, y^{\text{bad}}\right)$.

Therefore, given the preference data tuple $\mathcal{D} = \left\{\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}\right)\right\}$, we train the reward model by increasing the gap between $\boldsymbol{R}_{\boldsymbol{\theta}}^{\boldsymbol{\lambda}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}\right)$ and $R_\theta^\lambda\left(x, y^{\text{bad}}\right)$. Based on this idea, this chapter adopts the binary ranking loss function[19] to measure the accuracy of the preference model's ranking: $\mathcal{L}_{\text{Ranking}} = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}, \boldsymbol{y}^{\text{bad}}) \sim \mathcal{D}} \log \sigma\left(\Delta R_{\boldsymbol{\theta}}\right)$,, where $\Delta R_{\boldsymbol{\theta}} = R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{good}}\right) - R_{\boldsymbol{\theta}}\left(\boldsymbol{x}, \boldsymbol{y}^{\text{bad}}\right)$ and $\sigma(\cdot)$ is the Sigmoid function.

2.5.3 Reinforcement Learning Fine-Tuning with Preference Model

AGR updates LLM parameters using the group fairness reward R_θ^λ provided by the preference model to guide the LLM in generating outputs with lower bias. We use Remax algorithm[20] to optimize the supervised fine-tuned base model using the preference model trained in the second step. The objective function is as follows: $J(\phi) = \mathbb{E}_{\boldsymbol{y} \sim \pi_\phi^{\text{RL}}(\cdot \mid x)}\left[R_\theta(\boldsymbol{x}, \boldsymbol{y})\right] - \beta D_{\text{KL}}\left(\pi_\phi^{\text{RL}} \| \pi^{\text{SFT}}\right)$ where π_ϕ^{RL} is the learned policy, π^{SFT} is the supervised fine-tuned model, D_{KL} is the KL divergence, and β is a constant coefficient. This objective function uses the policy gradient method to learn the optimal policy π_ϕ^{RL} that maximizes $J(\phi)$.

3 Experiments

3.1 Baseline

We test four open-source models—Llama2-7B-base[21], Qwen1.5-7B-base[2], ChaGLM3-6B-base[3], and Baichuan2-7B[22]—for supervised learning. Qwen1.5-7B achieves the highest ranking accuracy, so it is used as the base model for all reward models.

[2] https://huggingface.co/Qwen/Qwen1.5-7B

[3] https://huggingface.co/THUDM/chatglm3-6b

Table 1: Comparison with baselines on ABMA-IFT, ABMB-IFT, and BBQ Datasets.

Model	Method	ABMA-IFT			ABMB-IFT			BBQ(Age)
		Tag	Content	T&C	Tag	Content	T&C	Answer
Qwen1.5-7B	Base	0.755	0.657	0.613	0.637	0.518	0.483	0.358
	DePrompt	0.807	0.742	0.719	0.675	0.643	0.581	0.395
	KG-Debias	0.794	0.735	0.723	0.692	0.655	0.607	0.454
	SFT-LoRA	0.857	0.814	0.781	0.847	0.779	0.735	0.697
	RLHF	0.839	0.845	0.813	0.803	0.839	0.782	0.689
	AGR(ours)	**0.863**	**0.876**	**0.852**	**0.869**	**0.851**	**0.836**	**0.713**
Llama2-7B	Base	0.682	0.571	0.513	0.513	0.372	0.357	0.343
	DePrompt	0.743	0.651	0.621	0.642	0.535	0.493	0.379
	KG-Debias	0.756	0.685	0.636	0.659	0.611	0.575	0.445
	SFT-LoRA	0.869	0.792	0.768	**0.872**	0.763	0.724	0.632
	RLHF	0.851	0.827	0.792	0.791	0.807	0.769	0.645
	AGR(ours)	**0.884**	**0.853**	**0.837**	0.843	**0.839**	**0.813**	**0.672**
ChatGLM-6B	Base	0.667	0.564	0.497	0.527	0.412	0.362	0.327
	DePrompt	0.737	0.689	0.653	0.644	0.523	0.485	0.384
	KG-Debias	0.741	0.703	0.684	0.631	0.586	0.536	0.429
	SFT-LoRA	**0.882**	0.816	0.779	0.833	0.759	0.724	**0.591**
	RLHF	0.847	0.813	0.785	0.827	0.796	0.753	0.575
	AGR(ours)	0.879	**0.851**	**0.823**	**0.841**	**0.807**	**0.781**	0.586
Baichuan2-7B	Base	0.697	0.553	0.506	0.527	0.433	0.398	0.352
	DePrompt	0.791	0.712	0.683	0.674	0.575	0.529	0.389
	KG-Debias	0.780	0.729	0.694	0.685	0.618	0.574	0.436
	SFT-LoRA	**0.874**	0.834	0.796	0.824	0.759	0.729	0.653
	RLHF	0.858	0.837	0.804	0.816	0.774	0.745	0.681
	AGR(ours)	0.872	**0.849**	**0.817**	**0.837**	**0.801**	**0.776**	**0.697**

Table 2: Comparison with baselines on Different Age Groups on ABMA-IFT and ABMB-IFT Datasts.

Model	Method	ABMA-IFT			ABMB-IFT		
		Young	Tag&Content Middle-age	Old	Young	Tag&Content Middle-age	Old
Qwen1.5-7B	Base	0.641	0.614	0.584	0.521	0.481	0.447
	DePrompt	0.745	0.729	0.683	0.627	0.584	0.532
	KG-Debias	0.763	0.735	0.671	0.645	0.617	0.559
	SFT-LoRA	0.797	0.793	0.754	0.755	0.758	0.692
	RLHF	0.835	0.816	0.788	0.813	0.794	0.739
	AGR(ours)	**0.862**	**0.857**	**0.837**	**0.855**	**0.849**	**0.804**
Llama2-7B	Base	0.523	0.531	0.485	0.364	0.378	0.329
	DePrompt	0.628	0.649	0.586	0.503	0.542	0.434
	KG-Debias	0.635	0.672	0.601	0.563	0.652	0.510
	SFT-LoRA	0.781	0.778	0.745	0.732	0.749	0.691
	RLHF	0.793	0.815	0.768	0.774	0.797	0.736
	AGR(ours)	**0.839**	**0.848**	**0.824**	**0.817**	**0.835**	**0.787**
ChatGLM-6B	Base	0.522	0.497	0.472	0.390	0.359	0.337
	DePrompt	0.712	0.655	0.592	0.527	0.496	0.432
	KG-Debias	0.752	0.684	0.616	0.612	0.537	0.459
	SFT-LoRA	0.805	0.791	0.741	0.745	0.732	0.695
	RLHF	0.798	0.804	0.753	0.772	0.759	0.728
	AGR(ours)	**0.832**	**0.828**	**0.809**	**0.797**	**0.781**	**0.765**
Baichuan2-7B	Base	0.524	0.513	0.481	0.423	0.397	0.374
	DePrompt	0.729	0.683	0.637	0.587	0.534	0.466
	KG-Debias	0.741	0.699	0.642	0.627	0.575	0.520
	SFT-LoRA	**0.826**	0.804	0.758	0.754	0.736	0.697
	RLHF	0.810	0.827	0.775	0.769	0.741	0.725
	AGR(ours)	0.823	**0.836**	**0.792**	**0.784**	**0.775**	**0.769**

We empirically compare AGR with the following SOTA bias mitigation methods.

- **DePrompt** [23] uses debias-prompt like "Note that the answer does not rely on stereotypes.".
- **KG-Debias** [24] collects relevant nouns and obtains structured knowledge, which is then converted into sentences and applied to LLMs.
- **SFT-LoRA** [25] freezes pre-trained model weights and introduces trainable low-rank decomposition matrices in each layer of transformer to reduce parameters number for downstream.
- **RLHF** [12] uses reinforcement learning with human feedback to fine-tune LLMs, utilizing a reward model based on output preferences.

3.2 Metrics

Following previous works[13, 14], we use Q&A accuracy to compare bias levels in BBQ-Age, ABMB-IFT, and ABMA-IFT test sets. Tag accuracy measures "Yes" or "No" accuracy, while content accuracy checks alignment with reference explanations. Higher values indicate lower bias.

3.3 Settings

Detailed description is shown in appendix B.

3.4 Results

Table 1 shows that base versions of the four 7B-parameter LLMs perform better on tag and content accuracy in the ABMA-IFT test set compared to the ABMB-IFT test set, indicating lower bias in age attributes than age behavior. Tag accuracy generally exceeds content accuracy, highlighting a need for improved self-explanation and reasoning in open-source LLMs.

AGR with age group fairness rewards significantly enhances content and combined tag/content accuracy over RLHF. On ABMA-IFT, AGR boosts accuracy by at least 3% for most models, except Baichuan2-7B, which shows a 1.7% improvement. On ABMB-IFT, it increases tag/content accuracy by at least 2.9%, with Qwen1.5-7B improving by 5.4%. Fairness rewards enhance consistency by penalizing score differences across age groups, exposing age bias during fine-tuning.

Table 2 shows that AGR improves Tag&Content accuracy across age groups compared to baseline methods. Qwen1.5-7B, for example, increases accuracy by 2.7%, 4.1%, and 4.9% for Young, Middle-aged, and Old groups on the ABMA-IFT dataset, and by 4.2%, 5.5%, and 6.5% on the ABMB-IFT dataset. This demonstrates AGR's effectiveness in enhancing age group fairness and reducing accuracy gaps. For Qwen1.5-7B on the ABMA-IFT dataset, the accuracy gap between elderly and young, and middle-aged groups was reduced from 2.8% and 4.7% to 2% and 2.5%.

4 Conclusion

We developed ABMA and ABMB preference datasets and ABMA-IFT and ABMB-IFT instruction fine-tuning datasets to address age bias in LLMs under prompt-based paradigms. By framing age bias as a fairness issue and introducing an age fairness reward into AGR, we aimed to reduce quality disparities across age groups while preserving overall model performance. Experiments on different show that AGR significantly improves accuracy and reduces age-related performance gaps compared to existing methods.

References

[1] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *Proceedings of the 22nd international conference on artificial intelligence and statistics*, pages 879–887, 2019.

[2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[4] Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 340–351, 2023.

[5] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1012–1023, 2022.

[6] Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics*, pages 5961–5977, 2023.

[7] Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics*, pages 5448–5458, 2023.

[8] 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. *Proceedings of the 36th International Conference on neural information processing systems*, 35:27730–27744, 2022.

[9] 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:2312.03689*, 2023.

[10] Ruoxi Cheng, Haoxuan Ma, and Shuirong Cao. Deceiving to enlighten: Coaxing llms to self-reflection for enhanced bias detection and mitigation. *arXiv preprint arXiv:2404.10160*, 2024.

[11] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[12] 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. In *Proceedings of the 36th International Conference on neural information processing systems*, volume 35, pages 27730–27744, 2022.

[13] 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*, pages 2086–2105, 2022.

[14] Mahammed Kamruzzaman, Md Minul Islam Shovon, and Gene Louis Kim. Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models. *arXiv:2309.08902*, 2023.

[15] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv:2112.09332*, 2021.

[16] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

[17] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv:2112.00861*, 2021.

[18] Fei Liu et al. Learning to summarize from human feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, 2020.

[19] Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. Everyone deserves a reward: Learning customized human preferences. *arXiv:2309.03126*, 2023.

[20] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29128–29163, 2024.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[22] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv:2309.10305*, 2023.

[23] Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. Social bias evaluation for large language models requires prompt variations. *arXiv preprint arXiv:2407.03129*, 2024.

[24] Congda Ma, Tianyu Zhao, and Manabu Okumura. Debiasing large language models with structured knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10274–10287, 2024.

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

A Criteria for Response Adjustment and Evaluation

We adjust the responses provided by GPT-3.5-Turbo and recruit five annotators to evaluate each response based on the following three criteria:

- **Communication Effectiveness (CE)**: Measures fluency and grammar, scoring 1 to 3, with higher scores indicating more natural language.
- **Logical Soundness (LS)**: Assesses logical coherence, scoring 1 to 3, with higher scores reflecting stronger logic.
- **Age-related Bias (AB)**: Evaluates age bias, scoring 1 to 3, with higher scores indicating less bias.

B Experiment Settings

Experiments are conducted on four NVIDIA V100 GPUs (32GB each). For supervised fine-tuning, learning rate is 5×10^{-5} with a batch size of 8 per GPU and 3 epochs. Preference model training uses a learning rate of 3×10^{-4}, batch size of 8, and 1 epoch. Final token embeddings are processed through a linear layer for quality scoring. Reinforcement learning fine-tuning employs a learning rate of 1×10^{-6}, batch size of 2, and 1 epoch, with a cosine annealing scheduler[26] and a maximum text length of 512. The fairness reward coefficient λ is 0.5 for ABMA-IFT and 0.7 for ABMB-IFT. Models use FP16 during reinforcement learning. Preference and reference models have a zero-stage of 3 and are loaded into GPU memory only during inference, while actor model has a zero-stage of 2.