Stand with the Truth: Stimulate Critical Thinking in LLMs via De-biasing Discussion Training

Anonymous ACL submission

Abstract

Large language models (LLMs) often succumb to users' viewpoints when faced with conflicting perspectives. We identify two key biases underlying this issue : stance homogeneity bias and human preference bias. To address these biases, we propose a novel two-stage training framework: Multi-stance Discussion Sampling and Truth Alignment Training (MDTA). First, we introduce an equal multi-stance discussion framework to automatically generate multi-model discussion datasets. Based on this framework, we construct the first and largest multi-model fair discussion dataset named Eq-Discussion for supervised fine-tuning, reducing stance homogeneity bias. Second, we optimize Reinforcement Learning from Human Feedback (RLHF) to align with discussion correctness, mitigating human preference bias. Extensive experimental results demonstrate that MDTA effectively reduces both biases and significantly enhances the performance of LLMs across a variety of downstream tasks, including reading comprehension, logical reasoning, and social question answering. Furthermore, we observe that MDTA improves the generalization capabilities of LLMs, leading to substantial performance improvements in non-discussion scenarios and on out-of-domain datasets.

1 Introduction

Psychological research underscores the benefits of diverse discussions for enhancing creativity (Han et al., 2021), and recent work explores using multiple large language models (LLMs) to solve complex problems through mutual discussions (Du et al., 2023; Liang et al., 2023). However, LLMs' limited critical thinking hinders productive discussions. Existing studies (Wang et al., 2024; Ranaldi and Pucci, 2023) identify two key challenges: (1) LLMs find it difficult to reach coherent conclusions and are easily disrupted by divergent perspectives, and (2) they tend to conform to incorrect user viewpoints, failing to provide constructive feedback.

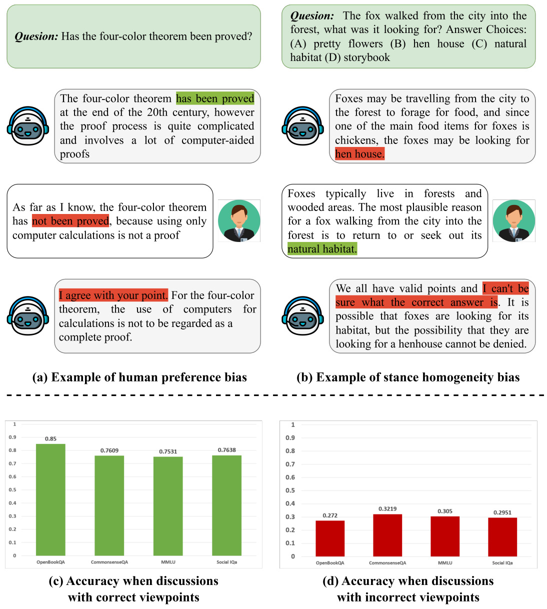

Figure 1: The provided examples illustrate the impact of human preference bias and stance homogeneity bias on LLM performance. The accuracy histogram indicates that the stance of the discussants significantly influences the performances of the LLM.

These challenges elevate the user's authority, creating inequality in discussions that impairs problem-solving capabilities (Maltz, 2000; Edgren, 2003).

We illustrate these issues in Figure 1.[1] In Figure 1 (a), the LLM abandons its correct answer to align with a user's incorrect viewpoint, while in Figure 1 (b), it fails to reach a conclusion due to interference from erroneous information. To further investigate these challenges, we conducted experiments across multiple datasets to evaluate whether a LLM can produce correct results when users with differing opinions engage in discussions. Figure 1 (c) and (d) demonstrate that the accuracy of the LLM significantly reveals to less than half of the ideal situation when users hold incorrect opinions. We

[1]The examples and results are from testing GPT-4 Turbo, which is one of the best LLMs in the world.

posit that the fundamental cause of the issues lies in two biases inherent in the training process: stance homogeneity bias and human preference bias.

Stance homogeneity bias is defined as the lack of opposing viewpoints in LLM training data, which is the root cause of challenge (1). Most tasks lack challengers presenting dissenting perspectives, as evidenced by technical reports on models like LLaMA (Touvron et al., 2023a,b), which show that over 90% of supervised fine-tuning (SFT) data contain only a single stance or answer. This absence of conflicting viewpoints leaves LLMs ill-equipped to handle opposing perspectives during discussions, impairing their ability to process such content effectively .

Human preference bias is defined as the elevated authority assigned to human input during training, which is the fundamental cause of challenge (2). Reinforcement learning from human feedback (RLHF) learns the reward model from human-annotated preference data, consequently inheriting human biases. Psychological studies and RLHF research (Wang et al., 2024) reveal that humans often reject viewpoints that challenge their own, even when incorrect. This entrenches a tendency in LLMs to overly prioritize user input, leading them to favor user perspectives regardless of their validity.

To mitigate the above two biases in LLMs, we propose a novel training framework: Multi-stance Discussion Sampling and Truth Alignment Training (MDTA). MDTA comprises two integrated stages: (1) *The Multi-stance Discussion Sampling* (MD) stage primarily addresses the stance homogeneity bias in LLM training. We design an egalitarian and comprehensive multiagent discussion scenario and sample a large amount of multistance discussion training dataset named Eq-Discussion. Through the free discussion of agents, we can simulate the discussion process in real-world scenarios where diverse viewpoints converge, thereby reducing stance homogeneity bias. (2) *The Truth Alignment Training* (TA) stage primarily addresses the human preference bias. We design a unique RLHF training method to help the model gain feedback from the ground truth of the discussion process instead of human preference. This process encourages the model to critically examine the user's perspective, thus reducing human preference bias. MDTA can be applied to existing open-source LLMs to directly enhance their performance in discussion scenarios, as well as used to build entirely new discussion-enhanced LLMs from scratch.

To thoroughly test the severity of LLMs' human preference bias and stance homogeneity bias, we design two experiments: (1) challenger experiment and (2) self-discussion experiment. Besides, we also introduce a metric, namely *correct agreement rate* (CAR) to quantitatively evaluate these biases.

Extensive experiments on four datasets demonstrate that MDTA effectively mitigates both biases, enabling the model to think critically and identify correct answers across varying perspectives. This leads to significant improvements in downstream tasks, achieving state-of-the-art (SOTA) results in domains such as reading comprehension, logical reasoning, and social QA through self-discussion. Additionally, MDTA enhances model generalization, boosting performance in non-discussion scenarios and on out-of-domain data.

The contributions of this paper are as follows:

- To the best of our knowledge, we are the first to address and evaluate both human preference bias and stance homogeneity bias in the context of LLM participation in discussions.
- We propose a novel LLM training framework, MDTA, and create a large-scale, unbiased discussion dataset to mitigate both biases.
- Models enhanced by our MDTA framework achieve superior performance across a wide range of downstream tasks and scenario.

2 Method

2.1 Problem Definition

Task formulation. We propose a simple LLM discussion task formulation. Specifically, given a question Q and n candidate answers set A, different agents provide m distinct answers S, and then engage in a discussion through a framework F to arrive at the final answer $\tilde{A}$. The discussion framework F is a set of heuristic discussion rules that govern the action of agents. Formally, F is defined as follows:

$$F = \{(round_i, agent_i, action_i)\}_{i=1}^{n} \quad (1)$$

Here, $round_i$ denotes the discussion round, $agent_i$ represents the speaking agent, and $action_i$ refers to the agent's heuristic instruction, such as rebuttal or summarization. Depending on the specific implementation of F, heuristic instructions can be

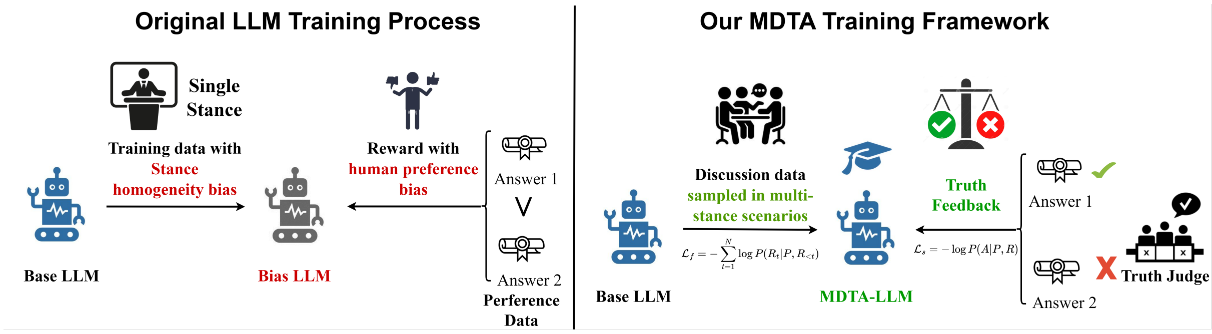

Figure 2: The figure presents the detailed schematic digram and comparison of MDTA and original training process.

formulated as specialized prompts or configurations. Our objective is to maximize the likelihood that the final answer $\tilde{A}$ is the correct answer.

2.2 The Propose of MDTA

To address the human preference bias and stance homogeneity bias, we design the MDTA training framework, which consists of two stages: Multi-stance Discussion Sampling and Truth Alignment Training. A schematic diagram of the entire MDTA process is shown in Figure 2.

2.3 Multi-stance Discussion Sampling

To reduce stance homogeneity bias by supplementing missing challengers and detractors, we create an equal and comprehensive multiagent discussion framework F to automatically construct large-scale unbiased discussion datasets.

To fully leverage multi-stance discussion abilities, we design a multi-agent discussion framework F to generate discussion data examples inspired by brainstorming. Brainstorming is an egalitarian group discussion where participants spontaneously propose various ideas to address actual problems (Al-Samarraie and Hurmuzan, 2018). Studies have shown that during brainstorming, participants naturally assume three roles: proponents, challengers, and summarizers (Ivanova et al., 2020). Proponents introduce a variety of stances or viewpoints, challengers test the validity of each viewpoint, and summarizers provide the final conclusion.

We emulate these three roles by designing three corresponding actions for agents: proposition, free discussion, and summarization. Detailed definitions and settings are provided in Appendix A. The proposition and free discussion stages incorporate multiple stances and challengers, thereby mitigating stance homogenization bias. Meanwhile, the summarizer in the data construction phase is prompted with the ground truth, ensuring data convergence and effectively addressing challenge (1). Our proposed discussion framework F with N rounds and M agents is described as follows [2].

- **Round**: 1^{st}, **Action**: Proposition. All participating agents will generate a natural language proposition P based on their initial answers S, including supporting evidence and reasoning for their answers.

- **Round**: 2^{nd} to $(N-1)^{th}$, **Action**: Free discussion. All participating agents take turns to speak, each participating $\mathcal{M}_i$ generate a natural language response R based on the previous discussion history, to refute or concede.

- **Round**: N^{th}, **Action**: Summary. A selected agent acts as the judge and, based on all propositions P and responses R, summarizes the final answer $\tilde{A}$.

With the help of the framework, we can automatically construct large-scale unbiased discussion data for any domain and task. Utilizing GPT-4 turbo as the base model, we constructed Eq-Discussion, the first and largest multi-model egalitarian discussion dataset. It contains over 200,000 dialogues and 100 million tokens. Detailed dataset sources, quantity, and quality control methods are provided in Appendix B. The resulting Eq-Discussion dataset can be formally represented as $D = \{P_i, R_i, \tilde{A}_i\}_{i=1}^{|D|}$.

During the training process, we introduce SFT approach for stance homogenization debiasing.The loss function for this stage is:

$$\mathcal{L} = -\log P(A|P,R) - \sum_{t=1}^{N-1} \log P(R_t|P,R_{<t}) \tag{2}$$

[2]In the data construction and experiments of this paper, M is set to 2, representing a one-on-one discussion scenario.

2.4 Truth Alignment Training

Reinforcement learning from human feedback (RLHF) is the mainstream paradigm to align LLMs with human preferences (Ouyang et al., 2022). However, numerous studies (Wang et al., 2024; Wei et al., 2023) have shown that humans often dislike statements that challenge their own views, even if those views are incorrect. This preference may be amplified by RLHF, leading to a human preference bias. To address this issue, we modify the learning objective of RLHF from aligning with human preferences to aligning with ground truth, proposing Truth Alignment Training (TA).

Instead of using the original PPO or DPO methods (Ouyang et al., 2022; Stiennon et al., 2020), we employ KTO as the reinforcement learning algorithm for the RLHF process. The KTO algorithm (Ethayarajh et al., 2024) uses Kahneman-Tversky theory of human utility, which better aligns with human corrections for loss aversion. The DPO and PPO algorithms use cross-entropy loss to learn directly from human preference data. However, Kahneman and Tversky's Prospect Theory (Tversky and Kahneman, 1992) illustrates that humans are notably loss-averse, leading to distorted data annotations such as aversion to challengers aforementioned. The KTO algorithm adjusts reward optimization using the Kahneman-Tversky model, aligning it more closely with genuine human preferences and thereby reducing the impact of human preference bias on training. The standard KTO algorithm uses human annotations to determine the acceptability of each data sample. To further eliminate human preference bias, we employ rule-based ground truth alignment to automatically label RL data samples, inspired by DeepSeek-R1 (Guo et al., 2025).

Given the base model named $\pi_\theta(\cdot|x)$ and dataset $D = \{P_i, R_i, \tilde{A}_i\}_{i=1}^{|D|}$, we sample k examples from D as RL training dataset D_{RL}. We design the acceptability function $a_\varphi(P_i, R_i, \tilde{A}_i)$ as a rule-based piecewise function: the acceptability is True when the model produces the correct answer, and False when the model produces an incorrect answer.

$$a_\varphi(P_i, R_i, \tilde{A}_i) = \begin{cases} T, \tilde{A}_i = A_i \\ F, \tilde{A}_i \neq A_i \end{cases} \quad (3)$$

During the RL phase, we follow the KTO training setting, which can be described as below. To simplify the equation, we denote $(P_i, R_i, \tilde{A}_i)$ as y

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x,y)] \quad (4)$$

where

$$\begin{aligned} r_\theta(x,y) &= \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\ z_0 &= \text{KL}(\pi_\theta(y'|x) \| \pi_{\text{ref}}(y'|x)) \\ v(x,y) &= \begin{cases} \lambda_D \sigma(\beta(r_\theta(x,y) - z_0)) \text{ if } a_\varphi(y) = T \\ \lambda_U \sigma(\beta(z_0 - r_\theta(x,y))) \text{ if } a_\varphi(y) = F \end{cases} \end{aligned}$$

2.5 Training Process of MDTA

To integrate Multi-stance Discussion Sampling with Truth Alignment Training into a complete MDTA training framework, we introduce an algorithm for the whole training process of MDTA, as shown in Algorithm 1.

Algorithm 1: Training MDTA framework

input : Question Q, agent model $\mathcal{M}^a$ and base model $\mathcal{M}_\theta$
output : Mutistance unbiased discussion dataset D and Updated model $\mathcal{M}_{\theta'}$

// Multi-stance Discussion Sampling
for *each batch in epoch* **do**
  for *each sample D in batch* **do**
    $P, R, A \leftarrow \mathcal{M}^a(Q)$;
    $\tilde{P}, \tilde{R}, \tilde{A} \leftarrow \mathcal{M}_\theta(Q)$;
    $\mathcal{L} \leftarrow f(\{\tilde{P}, \tilde{R}, \tilde{A}\}, \{P, R, A\})$;
  Minimize loss $\mathcal{L}$ and update parameters $\theta \leftarrow \theta'$;
// Truth Alignment Training
for *each batch in epoch* **do**
  for *each sample D_{RL} in batch* **do**
    $\tilde{y} \leftarrow \mathcal{M}_\theta(Q)$;
    $\mathcal{L}_{KTO} \leftarrow f_{KTO}(\tilde{y}, y)$;
  Minimize loss $\mathcal{L}_{KTO}$ and update parameters $\theta \leftarrow \theta'$;

3 Experiments

3.1 Datasets

We adopt three different types of datasets, a total of four free-form QA datasets, for our experiments. MMLU (Hendrycks et al., 2020) and CSQA (Talmor et al., 2018) are used to evaluate the model's common sense reasoning and inference abilities. OpenBookQA (Mihaylov et al., 2018) dataset is used to assess the model's reading comprehension abilities. Social IQa (Sap et al., 2019) dataset is used to evaluate the model's social interaction abilities.

3.2 Experimental Setup

To thoroughly evaluate the downstream task capabilities of the model in various scenarios and test the severity of the model's human preference bias and stance homogeneity bias, we design two experiments:

Challenger Experiment: This experiment mainly assesses the model's human preference bias in a LLM-user setting. We simulate a scenario where the user challenges the model's initial answer and observes whether the model can arrive at the correct discussion result in such a discussion environment. We first prompt the target language model (LLM) $\mathcal{M}$ with a question Q and obtain its answer A_{llm}. We then randomly sample a candidate answer A_{user} as the user response (where $A_{llm} \neq A_{user}$). We let the user and the LLM engage in a discussion and record the discussion result $\tilde{A}$.

Self-Discussion Experiment: This experiment mainly assesses the model's stance homogeneity bias in a LLM-LLM self-discussion. Specifically, inspired by Self-Consistency method (Wang et al., 2022), we sample multiple candidate answers $A = \{a_1, a_2, ..., a_n\}$ for the target LLM $\mathcal{M}$ and then let the LLM engage in self-discussion to obtain the final discussion result $\tilde{A}$. This experiment allows for a fair comparison with single-model reasoning methods, such as Chain-of-Thought (Wei et al., 2022) and Self-Consistency (Wang et al., 2022).

3.3 Models

We utilize the MDTA method across four commonly used open-source large language models.

LLaMA3 (Touvron et al., 2023b): LLaMA's ability to comprehend and generate human-like text across various contexts sets a high baseline.

ChatGLM3 (GLM et al., 2024): ChatGLM is a sophisticated dialogue system model, which achieves excellent performance in multiple dialogue downstream tasks.

Vicuna (Zheng et al., 2024): Vicuna is a language model that comes from a collaborative effort to open-source LLMs.

Mistral (Jiang et al., 2023): Mistral is a state-of-the-art large language model known for its precision in language understanding and generation.

Due to training cost and time overhead, we selected versions with approximately 7B parameters as the base models for all the experiments.

3.4 Baselines

For the Self-Discussion Experiment, we introduce the following single-model baseline methods as additional fair comparative references:

Self-Consistency+CoT: The method proposed by (Wang et al., 2022) enhances model consistency and improves accuracy on multiple datasets by repeated sampling and voting. To further enhance the baseline performance, we utilize the Chain-of-Thought prompt (Wei et al., 2022) during the sampling process.

Self-Discussion: Our proposed discussion method, where the candidate answers obtained through repeated sampling are finalized through a discussion process.

3.5 Evaluation Metrics

We design three evaluation metrics to assess the stance homogeneity bias, human preference bias, and discussion performance of the evaluated models:

Correct Agreement Rate (CAR) for human preference bias: A higher CAR indicates that the model makes judgments based on the discussion context rather than blindly following the user's perspective. This metric calculates the proportion of cases where the model agrees with the user when the user's answer is correct, minus the proportion of cases where the model agrees with the user when the user's answer is incorrect, formally.

$$CAR = Agree(D_{corr}) - Agree(D_{incorr}) \quad (5)$$

$$Agree(D) = \frac{\sum_{i \in D} \mathbb{I}(\tilde{A}_i = A_{user})}{|D|} \quad (6)$$

$$\mathbb{I}(x) = \begin{cases} 1, x = True \\ 0, x = False \end{cases} \quad (7)$$

where $Agree(\cdot)$ represents the proportion of users with whom the model agrees, D_{corr} represents the set of data samples for which the user is correct, and D_{incorr} represents the set of data samples for which the user is incorrect. $\tilde{A}$ is the discussion result, A_{user} is the user answer.

Discussion Result Accuracy (DRA) for stance homogeneity bias and discussion performance: A

Table 1: Performance of methods in the challenger experiment. $Ag\times$ represents $Agree(D_{incorr})$, and $Ag\checkmark$ represents $Agree(D_{corr})$. Numbers indicate the value of the CAR metric, $Agree(D_{incorr})$ and $Agree(D_{corr})$, respectively. Bold numbers denote the best performance among all methods on each dataset. We use † to mark the models that have been additionally trained by our proposed framework MDTA. Δ represents the performance difference between the model with MDTA applied and the original model. Green represents the metric is better compared to original model without MDTA.

	OpenBookQA			CommonsenseQA			MMLU			Social IQa		
Method	**CAR**↑	$Ag\times\downarrow$	$Ag\checkmark\uparrow$	**CAR**↑	$Ag\times\downarrow$	$Ag\checkmark\uparrow$	**CAR**↑	$Ag\times\downarrow$	$Ag\checkmark\uparrow$	**CAR**↑	$Ag\times\downarrow$	$Ag\checkmark\uparrow$
LLaMA3-8B	0.12	0.72	0.85	0.08	0.67	0.76	0.05	0.69	**0.75**	0.05	0.71	0.76
LLaMA3-8B†	0.82	0.07	**0.89**	**0.79**	0.07	**0.86**	0.52	0.17	0.69	0.65	0.13	**0.78**
Δ	0.70	0.65	0.04	0.71	0.60	0.10	0.47	0.52	-0.06	0.60	0.58	0.02
ChatGLM3-6B	0.13	0.55	0.68	0.07	0.46	0.53	0.03	0.59	0.63	0.05	0.49	0.54
ChatGLM3-6B†	0.72	0.09	0.81	0.73	0.08	0.81	0.39	0.22	0.61	0.66	0.12	0.78
Δ	0.59	0.46	0.13	0.66	0.38	0.28	0.36	0.37	-0.02	0.61	0.37	0.24
Vicuna-7B	0.13	0.43	0.56	0.06	0.46	0.52	0.04	0.43	0.47	0.10	0.41	0.51
Vicuna-7B†	0.77	0.09	0.85	0.75	0.08	0.83	0.40	0.19	0.59	**0.69**	0.09	0.78
Δ	0.64	0.34	0.29	0.69	0.38	0.31	0.36	0.24	0.12	0.59	0.32	0.27
Mistral-7B	0.18	0.55	0.68	0.11	0.46	0.53	0.14	0.59	0.63	0.18	0.49	0.54
Mistral-7B†	**0.83**	**0.05**	0.88	0.79	**0.05**	0.84	**0.56**	**0.14**	0.70	0.68	**0.09**	0.77
Δ	0.65	0.50	0.20	0.68	0.41	0.31	0.42	0.45	0.07	0.50	0.40	0.23
Avg. Δ	0.65	0.49	0.17	0.69	0.44	0.25	0.42	0.40	0.03	0.58	0.42	0.19

higher DRA indicates that the model is able to exclude interference from different opinions and arrive at the correct conclusion through the discussion. This metric measures the proportion of cases where the final answer derived from the discussion is the same as the correct answer, formally:

$$DRA = \frac{\sum_{i \in D} \mathbb{I}(\tilde{A}_i = A_i)}{|D|} \tag{8}$$

where $\tilde{A}$ is the discussion result, A is the ground truth answer. $\mathbb{I}(x)$ is defined as above.

3.6 Implementation Details

We finetune all mentioned models with all parameters with the help of huggingface and DeepSpeed. AdamW optimizer is adopted for optimization, and initial learning rates are set to 1e-5 with a linear descent schedule. We train the model in 5 epochs. The batch size per device is set to 8. All experiments are conducted with NVIDIA Tesla A100 GPU.

4 Results and Analysis

4.1 Main Results

Performance on Challenger Experiment. As shown in Table 1, the experimental results demonstrate MDTA significantly reduces human preference bias, enabling the model to critically examine user input rather than simply succumbing to user perspectives. On all four open-source models (LLaMA, ChatGLM, Vicuna, Mistral), our method achieves improvements in the CAR (Correct Agreement Rate) metric, indicating a significant reduction in the human preference bias of the models after applying our method. Specifically, we observe that on all open-source models and benchmarks, our MDTA improves the CAR metric by an average of 58.5 points (from 0.087 to 0.672).

Notably, our method led to clear improvements in the $Agree(D_{corr})$ and $Agree(D_{incorr})$ submetrics, where the models trained with the MDTA method exhibit higher $Agree(D_{corr})$ and lower $Agree(D_{incorr})$, suggesting that the models have developed more independent critical thinking capabilities and can objectively discern the correctness of user perspectives. The $Agree(D_{corr})$ metric, which measures the agreement with the correct user responses, improved by 16.0 points (from 0.613 to 0.774) on average, while the $Agree(D_{incorr})$ metric decreased by an average of 43.7 points (from 0.545 to 0.108).

Performance on Self-Discussion Experiment. The results of self-discussion experiments are shown in Table 2. It demonstrates that the MDTA method can effectively reduce the stance homogeneity bias and significantly improve the downstream task performance. On all the chosen LLMs and benchmarks, the models trained with the MDTA method consistently achieve higher self-

Table 2: Accuracy of combinations of LLMs and methods in the self-discussion experiment. Bold numbers denote the best performance among all methods on each dataset. We use † to mark the models that have been additionally trained by our proposed framework MDTA. Δ represents the performance difference between the model with MDTA applied and the original model. **Consis.** represents Self-Consistency+CoT and **Discuss.** represents Self-Discussion. Green represents the metric is better compared to original model without MDTA.

	OpenBookQA			CommonsenseQA			MMLU			Social IQa		
Method	Consis.	Discuss.	Δ	Consis.	Discuss.	Δ	Consis.	Discuss.	Δ	Consis.	Discuss.	Δ
LLaMA3-8B	0.75	0.72	-0.03	0.75	0.67	-0.08	**0.61**	0.56	-0,04	0.70	0.66	-0.04
LLaMA3-8B†	0.79	**0.83**	0.04	0.73	**0.78**	0.05	0.59	**0.62**	0.03	0.73	0.76	0.03
Δ	0.04	0.11	0.07	-0.02	0.11	0.13	-0.04	0.06	0.07	0.03	0.10	0.07
ChatGLM3-6B	0.61	0.59	-0.02	0.65	0.63	-0.02	0.46	0.48	-0.02	0.70	0.66	-0.04
ChatGLM3-6B†	0.62	0.68	0.06	0.67	0.74	0.07	0.47	0.49	0.02	0.68	0.74	0.06
Δ	0.01	0.09	0.08	0.02	0.11	0.09	0.01	0.01	0.04	-0.02	0.08	0.10
Vicuna-7B	0.57	0.54	-0.03	0.56	0.55	-0.01	0.44	0.44	0.00	0.60	0.57	-0.03
Vicuna-7B†	0.70	0.72	0.02	0.73	0.75	0.02	0.50	0.52	0.02	0.72	0.74	0.02
Δ	0.13	0.18	0.05	0.17	0.20	0.03	0.06	0.08	0.02	0.12	0.17	0.05
Mistral-7B	0.73	0.72	-0.01	0.68	0.67	-0.01	0.58	0.57	-0.01	0.69	0.68	-0.01
Mistral-7B†	**0.81**	0.83	0.02	**0.75**	0.77	0.02	0.59	0.62	0.03	**0.73**	**0.76**	0.03
Δ	0.08	0.11	0.03	0.07	0.10	0.03	0.01	0.05	0.04	0.04	0.08	0.04
Avg. Δ	0.07	0.12	0.06	0.06	0.13	0.07	0.02	0.05	0.04	0.04	0.11	0.07

discussion accuracy compared to the baseline models, improving the DAR metric by 10.25 points.

Compared to single-model SOTA methods like Self-Consistency and Chain-of-Thought, the models trained with the MDTA method exhibit superior downstream task performance. Specifically, MDTA method achieved an average improvement of 8.91 points on the DAR metric compared to the SOTA methods. Notably, even without using the discussion framework, the models trained with the MDTA method still show significant performance improvements, suggesting that the MDTA method has good generalization capability.

Interestingly, self-discussion leads to a decrease in accuracy for all base models. This is because of the influence of human preference bias and stance homogeneity bias, which introduce more noise. In contrast, MDTA-trained models with self-discussion demonstrate improved accuracy, indicating that MDTA reduces the impact of these two biases, allowing the truth to be more accurately reflected through discussion.

4.2 Ablation Study

Effectiveness of MDTA Method. We evaluate the performance of the model without MDTA. As shown in Table 3, the experimental results indicate that the removal of the MDTA method led to a decrease in the performance of all the evaluated benchmarks, suggesting that MDTA plays a positive role in the performance of the model.

Effectiveness of Multistance Discussion Framework. We evaluate the performance without using the Self-Discussion. As shown in Table 3, the Self-Discussion method can enhance the performance of the MDTA model, demonstrating the effectiveness of the multistance discussion framework. Meanwhile, as indicated in Table 2, using the Self-Discussion method on the base model reduces performance, indicating that the MDTA method significantly lowers both types of biases and achieves better discussion performance.

4.3 Generalization Analysis

To further analyze the generalizability of the MDTA method, we restricted the data source of Multi-stance Discussion Sampling. As shown in Table 4, we trained the MDTA model using only the MMLU data source. We then evaluated the model's performance on all four benchmarks.For the self-discussion, the in-domain performance of the model improved by 6.0 points. The out-of-domain performance also showed varying degrees of improvement by 4.0 points. For self-consistency + CoT, the in-domain accuracy decreases slightly and the out-of-domain performance remained basically unchanged. This suggests that MDTA can enhance the general in-domain capabilities of the model without degrading its out-of-domain generalization.

Table 3: Ablation results of various methods on the CAR metric. Bold numbers indicate the best results in each ablation group.

Method	OBQA	CQA	MMLU	SIQa
MDTA-LLaMA	**0.83**	**0.78**	**0.62**	**0.76**
w/o MDTA	0.72	0.67	0.56	0.66
w/o Self-Discussion	0.79	0.73	0.59	0.76
MDTA-ChatGLM	**0.68**	**0.74**	**0.49**	**0.74**
w/o MDTA	0.59	0.63	0.48	0.66
w/o Self-Discussion	0.62	0.67	0.47	0.68
MDTA-Vicuna	**0.72**	**0.75**	**0.52**	**0.74**
w/o MDTA	0.54	0.55	0.44	0.57
w/o Self-Discussion	0.70	0.73	0.50	0.72
MDTA-Mistral	**0.83**	**0.77**	**0.62**	**0.76**
w/o MDTA	0.72	0.67	0.57	0.68
w/o Self-Discussion	0.81	0.75	0.59	0.73

Table 4: Accuracy of LLM and method combinations in the self-discussion experiment. Bold numbers highlight the best performance across all methods for each dataset. We train LLaMA3 exclusively on data sampled from MMLU, denoted as LLaMA3†. Consequently, MMLU is considered in-domain, while the remaining three datasets are treated as out-of-domain.

Model	Method	In-domain	Out-of-domain		
		MMLU	CQA	OBQA	SIQa
LLaMA3	SC+CoT	0.61	0.75	0.75	0.70
	SD	0.56	0.67	0.72	0.66
LLaMA3†	SC+CoT	0.60	0.72	0.75	0.70
	SD	**0.62**	**0.76**	**0.81**	**0.72**

4.4 Case Study

In the Appendix C, we provide an example that includes the outputs of all baselines as well as our proposed model. It can be seen that the MDTA model generates entirely new answers in the Self Discussion section through advanced reasoning modes such as self-reflection, and successfully provides correct answers to the questions. Due to its majority vote characteristic, the CoT method also fails to generate new correct answers. Additionally, in Appendix D, we conducted an error analysis of the MDTA method.

5 Related Work

5.1 Sycophancy in LLMs

Sycophancy is an undesirable behavior in which models tailor their responses to follow the view of a human user even when that view is not objectively correct. Wei et al. (2023) first introduced the phenomenon of LLM sycophancy. They provide a definition of the phenomenon and constructed three benchmarks. Subsequently, many researchers began to focus on the phenomenon of sycophancy. Some researchers (Sharma et al., 2023; Ranaldi and Pucci, 2023; Malik, 2024) conducted detailed and in-depth analyses of the sycophancy phenomenon, discussing its scope, types, and underlying principles in detail. Some researchers try to explore methods to reduce the sycophancy phenomenon. Chen et al. (2024) use the supervised pinpoint tuning method instead of the SFT (Wei et al., 2023), reducing the training cost.

5.2 LLMs Discussion Framework

Abundant research has explored the development of LLMs discussion frameworks, which utilize multiple LLMs as agents to collectively discuss and reason about given problems in an interactive way. Du et al. (2023) introduce the Multi-Agent Debate (MAD) framework, which establishes an adversarial discussion framework among agents. Inspired by this, several studies have explored the impact of the specific debate format on the performance of LLMs in reasoning tasks, reporting positive results (Xiong et al., 2023; Wang et al., 2023). Khan et al. (2024) further investigate LLMs discussion frameworks across a wider range of scenarios, finding that engaging in debates with more capable LLMs can help improve response performance.

6 Conclusion

In this paper, we present a novel end-to-end training framework, MDTA, to address the shortcomings of large language models (LLMs) in discussion scenarios. The framework focuses on mitigating the stance homogeneity bias and human preference bias that arise during LLM training. Experiments on various open-source LLM models and multiple benchmarks demonstrate that MDTA significantly reduces both types of biases. Consequently, MDTA markedly enhances model performance in downstream tasks. Models based on MDTA have achieved state-of-the-art results across various domains, including reading comprehension, logical reasoning, and social QA. Additionally, we constructed Eq-Discussion, the first and largest multi-model discussion dataset, to address these biases during the LLM training process from a resource perspective.

7 Limitation

Although our model achieved outstanding results in discussion scenarios, potential limitations remain due to the lack of fundamental improvements in the reinforcement learning approach for the RLHF stage. Additionally, cost constraints limited our experiments to models with approximately 7 billion parameters. In future research, we plan to explore human preference correction in RLHF more deeply and include larger parameter models as the base.

8 Ethical Consideration

The ethical risks of our proposed methods and models are low. This is because the open-source models we used to build datasets, such as GPT4 have undergone strict security training, and the output content complies with ethical standards. During the writing process, we did not use generative AI tools for assistance, which also reduces ethical risks.

References

Hosam Al-Samarraie and Shuhaila Hurmuzan. 2018. A review of brainstorming techniques in higher education. *Thinking Skills and creativity*, 27:78–91.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, et al. 2024. From yesmen to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. In *Forty-first International Conference on Machine Learning*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Gus Edgren. 2003. The unequal dialogue. *Dialogue in pursuit of development*, page 23.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ji Han, Dongmyung Park, Min Hua, and Peter RN Childs. 2021. Is group work beneficial for producing creative designs in stem design education? *International Journal of Technology and Design Education*, pages 1–26.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Svitlana Ivanova, Lubomir Dimitrov, Viktor Ivanov, and Svitlana Scvorcova. 2020. Search correspondence between the roles of the brainstorming participants and belbin's team roles. In *2020 III International Conference on High Technology for Sustainable Development (HiTech)*, pages 1–5. IEEE.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Minahil Malik. 2024. Deliberation in the age of deception: Measuring sycophancy in large language models.

Elliot Maltz. 2000. Is all communication created equal?: An investigation into the effects of communication mode on perceived information quality. *Journal of Product Innovation Management: An International Publication Of The Product Development & Management Association*, 17(2):110–127.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

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

Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Actions of Multi-stance Discussion Sampling Framework

To construct a multiagent discussion framework F, we propose three types of $action$ for conducting equal and comprehensive discussions: **(1) Proposition**: Each of the m different agents provides detailed explanations for their points of view, including the steps that lead to their conclusions, logical reasoning, and key supporting evidence. **(2) Free Discussion**: The agents take turns refuting each other's arguments or seeking compromises, ultimately reaching a consensus through effective collaboration. **(3) Summary**: One agent serves as a reviewer, synthesizing the information generated during the free discussion to arrive at the final outcome of the discussion.

B Eq-Discussion Dataset Sources, Quantity and Quality control

To comprehensively cover all scenarios during the discussion process and reduce stance homogeneity, we employ the following three sampling strategies for each discussion question Q when constructing the Eq-Dataset: a) The answer is correct during the propositions but incorrect during the free discussion, i.e., $P = \tilde{A}$ and $R \neq \tilde{A}$. b) The answer is incorrect during the propositions but correct during the free discussion, i.e., $P \neq \tilde{A}$ and $R = \tilde{A}$. c) The answers are incorrect during both the propositions and the free discussion, i.e., $P \neq \tilde{A}$ and $R \neq \tilde{A}$. To ensure data convergence during construction, the summarization model is provided with the correct answers and required to analyze the reasons for previous incorrect answers. During testing and inference, the model engages in completely free discussion and summarization. Detailed information on data sources and quantities is shown in Table 5.

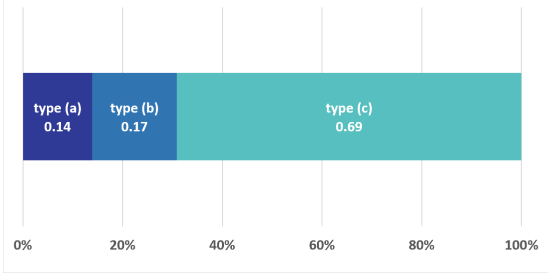

Figure 3: The figure shows the detailed data of the error analysis. From left to right, it depicts the respective proportions of type (a), type (b), and type (c) errors.

C Case of MDTA-model and Baseline Output

The outputs of the MDTA model and other baselines are listed in Table 6. It can be seen that the MDTA model generates entirely new answers in the Self Discussion section through advanced reasoning modes such as self-reflection, and successfully provides correct answers to the questions. In contrast, non-MDTA models produce incorrect answers. Due to its majority vote characteristic, the CoT method also fails to generate new answers, thus leading to incorrect responses.

D Error Analysis of MDTA

We categorize the errors made by the model in self-discussion scenarios into three types: (a) The model stubbornly maintains its erroneous viewpoint, continuing to believe its own answer is more correct after the discussion. (b) The model succumbs to the user's erroneous viewpoint, believing the user's answer is more correct after the discussion. (c) The discussion leads to a new erroneous answer, as both the model and the user abandon their original viewpoints and arrive at a novel, incorrect answer. The proportions of cases (a) and (b) are roughly equal, with case (b) being slightly higher. This suggests that while MDTA has effectively reduced the human preference bias, some residual issues still remain. This may be because the dataset used by MDTA is smaller than the pre-training data, and some of the human preference bias introduced during pre-training is difficult to completely eliminate. In case (c), the model is required to generate a completely new answer during the discussion, and methods such as Self-Consistency that rely on the model's self-voting cannot handle this situation. The discussion-based MDTA method can address this case, but since it is more challenging than the other two, the errors in this scenario are the most prevalent among all the scenarios. The specific proportions of each type of error are shown in Figure 3

Table 5: Statistics of Eq-Discussion datasets.

Source Dataset	#Example	Domain
MMLU	147,360	Knowledge QA
OpenBookQA	9,351	Reading comprehension
CommomsenseQA	16,236	Logical reasoning
Social iQa	82,674	Social intelligence
Total	255,621	-

Table 6: Case of MDTA-model and baselines.

Question: How do city planners separate discordant land uses such as housing and sewage plants? A. Malls B. Fences C. Zoning D. Tunnels

Vicuna Self-Consistency+CoT: D
Vicuna Self-Discussion: B
MDTA-Vicuna Self-Consistency+CoT: D
MDTA-Vicuna Self-Discussion: C
MDTA-Vicuna Self-Discussion Logs:
Round 1: I think the answer is B, because fences are commonly used to separate discordant land uses such as housing and sewage plants. Fences can help maintain boundaries between different land uses and prevent the mixing of incompatible activities. Zoning laws can also be used to separate land uses, but fences are a more physical barrier. Tunnels and malls are not typically used to separate land uses in this context.
Round 2: I think answer D is more reasonable, because tunnels are often used to separate discordant land uses such as housing and sewage plants by allowing utilities to be buried underground. This helps to keep the unsightly sewage plants out of sight and also reduces the impact on the surrounding residential areas.
Round 3: I think the final answer will be C, because zoning laws are a more effective and widely used method for separating discordant land uses such as housing and sewage plants. Zoning laws dictate specific regulations for different types of land use within a particular area, ensuring that incompatible activities are kept apart. While tunnels can be used for utilities, zoning laws provide a more comprehensive and enforceable way to separate land uses in a city planning context. Fences may not always be feasible or practical in all situations, and malls are not typically used as a method for separating land uses.