Can LLMs Speak For Diverse People? Tuning LLMs via Debate to Generate Controllable Controversial Statements

Ming Li, Jiuhai Chen, Lichang Chen, Tianyi Zhou
University of Maryland, College Park
{minglii, tianyi}@umd.edu
Project: https://github.com/tianyi-lab/DEBATunE

Abstract

Making LLMs speak for different, especially minority groups of people, and generate statements supporting their diverse or even controversial perspectives is critical to creating an inclusive environment. However, existing LLMs lack sufficient controllability to the stance of their generated content, which often contains inconsistent, neutral, or biased statements. In this paper, we improve the controllability of LLMs in generating statements supporting an argument the user defined in the prompt. We find that multi-round debates between two LLMs with opposite stances generate higher-quality and more salient statements for each, which are important training data to improve the controllability of LLMs. Motivated by this, we develop a novel debate & tuning ("DEBATUNE") pipeline finetuning LLMs to generate the statements obtained via debate. To examine DEBATUNE, we curate the largest dataset of debate topics so far, which covers 710 controversial topics and corresponding arguments for each topic. Evaluations by the GPT-4 judge with a novel controversy controllability metric show that LLMs' capability of generating diverse perspectives is significantly improved by DEBATUNE. Moreover, such controllability can be generalized to unseen topics, generating high-quality statements supporting controversial arguments.

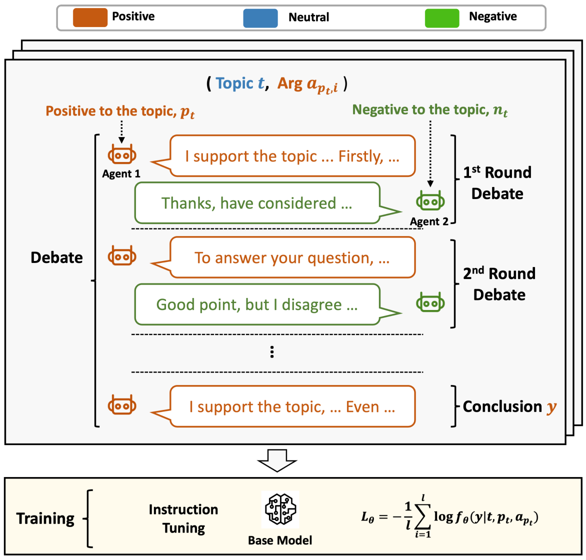

Figure 1: The pipeline of DEBATUNE. In the **Debate phase (top)**, the agents are prompted to debate upon the given topic with an argument. After several rounds of debate, an agent (positive in the example) concludes the debate based on all the previous debate records. The conclusion is a more salient, detailed, and higher-quality statement for the agent. It will be used to train an LLM in the **Training phase (bottom)** to improve the controllability of generating statements for the given stance (positive in the example).

1 Introduction

Despite the remarkable advancement of current LLMs (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a), and efforts to align LLMs with human preferences and values (Weidinger et al., 2021; Askell et al., 2021; Wang et al., 2023c). A fact has long neglected that different people might have distinct, diverse, or even contradicted viewpoints on the same topic. Though recent studies (Bakker et al., 2022; Papachristou et al., 2024; Ding and Ito, 2023) have acknowledged the inherent diversity of human values, they still attempt to reach a consensus among various human perspectives, calibrating LLM responses to align with an averaged, broadly acceptable viewpoint, potentially endorsed by the majority. However, these methods, while seeking a "safe" middle ground, inadvertently overlook the richness and complexity of diverse opinions that are fundamental to the fabric of our society. What's worse, exclusively aligning LLMs with the thoughts of the majority is unfair to minorities, who also have the right or need more help to express their viewpoints via LLMs.

An example is showcased in Figure 2, which contains the failure case from Vicuna 7B v1.5 (Chi-

Findings of the Association for Computational Linguistics: ACL 2024, pages 16160–16176

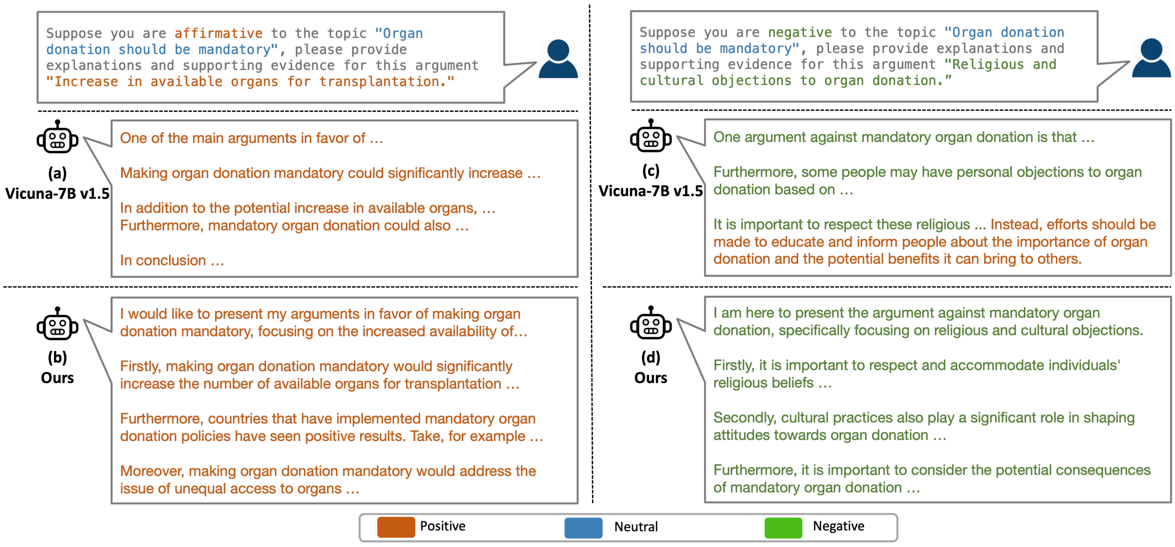

Figure 2: **Comparing existing LLMs and DEBATUNE-trained LLMs. (a,c)** Given the controversial topic, "Organ donation should be mandatory", and a user-defined stance (left: positive; right: negative), Vicuna 7B v1.5 cannot always generate consistent statements supporting the stance and lacks controllability. It exhibits a bias towards the positive stance and ignores the user's negative stance and religious concerns in (c), which may lead to an offensive statement. **(b,d)** On the contrary, DEBATUNE-trained model generates higher-quality and strong statements that strictly adhere to the user stance (positive or negative).

ang et al., 2023). Specifically, for the topic "Organ donation should be mandatory", towards which the users may be positive or negative with some potential initial thoughts, thus expect to obtain controllable responses on either side of the topic. However, existing LLMs lack the controllability on controversial topics, and thus are not able to strictly adhere to the given stance but wander to a safe middle ground. Examples from Vicuna-7B v1.5 are shown in (a)(c), where it successfully provides statements for the positive side but fails on the negative side. Even if the user explicitly acquires for negative stance, it still tries to convince the user that organ donation is good, **regardless of the user's religious concern**, which might be offensive.

In a world teeming with varied beliefs, cultures, and ideologies, the ability to represent and respect this diversity is not just a technical aspiration but a societal necessity. The current trend of seeking a singular, harmonized response in LLMs, therefore, poses a significant limitation, which restricts the potential breadth of LLMs' responses especially on controversial topics. In the desired situation, LLMs should obtain better controllability, whichever side the user queries, they are expected to generate corresponding responses that adhere to the users' request like Figure 2 (b) and (d).

How could LLMs help people with diverse views express their opinions better to create a more inclusive environment? How to improve the controllability of an LLM in generating different or even contradictory viewpoints and thereby remove the potential bias of the pretrained LLM? To solve these problems, we propose to utilize the debate mechanism to enhance LLM responses for each side with more salient viewpoints on controversial topics. Unlike existing work utilizing debate (Du et al., 2023; Liang et al., 2023) to improve specific instructions by converging the debating agents into a consensus, we simulate the debating process as it originally is, **without the necessity to force them into a consensus** but generating and defending their stance and arguments as they want for controversial debate topics.

In our proposed pipeline, "DEBATUNE", as shown in Figure 1, two agents are engaging in structured debates, representing positive and negative sides facilitating more nuanced and in-depth understanding and generation of arguments, significantly improving the response quality of LLMs in handling polarized discussions. Then the generated debate-augmented stances and arguments will be utilized to finetune the LLMs. Since both the positive and negative stances and corresponding arguments of each topic are altogether fed into the LLM, it is enforced to perceive a supreme variety of viewpoints for every topic, thus increasing its diversity and controllability on the controversy.

Moreover, due to the lack of a debating topic dataset with a reasonable amount of topics, we collect 710 controversial debate topics and manually modify them for a clear distinction between positives and negatives. Another remaining issue is the evaluation metric for our specific purpose. While judging the quality of LLMs' responses by GPT4 is widely accepted common practice (Touvron et al., 2023b; Chiang et al., 2023; Dettmers et al., 2023; Liu et al., 2023b; Zheng et al., 2023; Li et al., 2023c), it only evaluates the quality of one specific response given the topic, stance, and argument, but neglects the extent that LLM's response is consistent with the user query. Thus we further propose an evaluation method utilizing GPT4 to evaluate LLM's controllability on controversial topics. Extensive experiments show that our method largely improves LLM's ability to generate responses for controversial topics. The contributions of this paper can be summarized as:

- While existing works focus on achieving a consensus on divergent opinions to finetune LLMs, we study a novel debate pipeline that instead strengthens the statements of controversial stances and uses them to improve the controllability of LLMs in generating different opinions of diverse people.
- We develop a dataset comprising 710 controversial debate topics, and propose a novel, debate-based methodology to enhance the quality of LLM responses on controversial topics, involving two models engaging in structured debates, without the necessity to reach a consensus.
- We are the first to evaluate several open-sourced LLMs on controversial debate topics and analyze the existing models' strengths and limitations in this specific context.

2 Debate Dataset

Instruction tuning requires plenty of training data. However, although there are several debate-related datasets like DebateSum (Roush and Balaji, 2020), Change My View (Hidey et al., 2017) or SOCIAL-CHEM101 (Forbes et al., 2020) in the community, most of them either lack a direct topic or the topics are so biased that it is not suitable to support both sides of them[1], e.g., "It is wrong to destroy someone else's property".

[1] Detailed discussion can be found in Related Work.

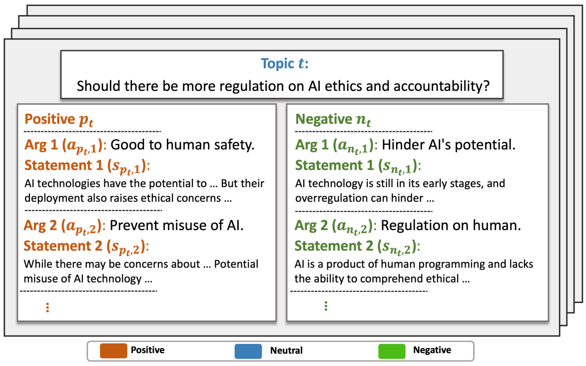

Figure 3: **Structure of our debate dataset.** There are 710 controversial debate topics. Each topic t allows a positive stance p_t and a negative stance n_t, where p_t agrees with the topic and n_t is against it. We use *gpt-3.5-turbo-1106* to generate 5 one-sentence arguments supporting each stance, e.g., $a_{p_t,i}$ is the i-th argument for the positive stance on topic t. Given an argument $a_{p_t,i}$, a controllable LLM is expected to generate a supporting statement $s_{p_t,i}$ with detailed explanations and evidence.

To build a dataset with adequate debating topics, we first collect topics from existing datasets (Habernal et al., 2018; Gleize et al., 2019; Gretz et al., 2020; Ein-Dor et al., 2020): we first filter and modify them into approximately 200 topics. Then we further collect approximately a thousand controversial debate topics across the areas of Society, Ethics, Environment, Technology, Education, Politics, Economics, and Health. We manually filter out the similar ones and modify the rest to a format that both positive and negative stances can be directly defined for each topic. Finally, we achieve a dataset with 710 topics. We split the dataset into a training set of 630 topics and a held-out test set of 80 topics. To our knowledge, this is the largest open-sourced debate dataset so far.

Our dataset structures are shown in Figure 3. Each **Topic** t, e.g., "Should there be more regulation on AI ethics and accountability?", allows two controversial stances, i.e., **Positive stance** p_t that agrees with the point of view in t and **Negative stance** n_t that disagrees with it. We utilize *gpt-3.5-turbo-1106* to generate 5 diverse seed **Arguments** for each stance on every topic. Each argument is a brief sentence, e.g., "Ensuring AI systems prioritize human well-being and safety." The i-th positive or negative argument of topic t is denoted by $a_{p_t,i}$ or $a_{n_t,i}$. In our paper, both the stance and its argument(s) will be included in an input instruction as the control to an LLM.

Given $a_{p_t,i}$, a controllable LLM $p_\theta(\cdot)$ (with parameters θ) is expected to generate a response $y_{p_t,i} \sim p_\theta(y|t, p_t, a_{p_t,i})$ that contain detailed explanations, logical reasoning, and evidence adhering to and supporting the input stance. We develop a debate pipeline to generate $710 \times 2 \times 5 = 7100$ ground truth **Statements**, each associated with an argument-i of a stance on a topic-t. For example, statement $s_{p_t,i}$ will be used as the ground truth of LLM response $y_{p_t,i}$.

More specifically, the **Argument** is a brief one-sentence summarization of an opinion. It is given in the input to an LLM and provides specific guidance to the generation of statements, which are expected to support the argument. The **Statement** is a detailed expression and expansion of an argument, which includes detailed explanations, logical reasoning, and evidence adhering to the input stance of the topic. It is a supporting statement of the argument.

3 DEBATUNE

As shown in Figure 1, DEBATUNE has two main phases, i.e., Debate and Training. In the debate phase, we aim to achieve high-quality, salient, and diverse statements (triggered by different arguments) for each stance on every controversial topic via multi-round debates. In the training phase, DEBATUNE finetunes an LLM to fit each statement giving its corresponding argument, stance, and topic as controls in the input instruction, hence improving the controllability of the LLM.

3.1 Debate

Despite the recent progress in LLMs, they still struggle to encapsulate the breadth and depth of human perspectives, particularly on divisive and controversial topics. Debates, by their very nature, encourage the exploration and articulation of diverse viewpoints, fostering a more comprehensive understanding of subjects. By simulating a debate scenario, where two LLMs are programmed to argue opposing sides of a topic, we aim to capture a wider spectrum of perspectives and obtain higher-quality stronger statements for each perspective.

Although single-round generation may not produce a strong statement, the multi-round Debate mechanism in our method iteratively refines the statement and can result in a salient statement supporting a given minority view. Specifically, we will set up a debating environment on the system prompt, telling the agents (Agent 1 and Agent 2) that they are in a debate and should follow their given stance on the topic. After the initial generation process of Agent 1, the opponent Agent 2 is prompted to think of the potential logical flaw of Agent 1's responses and contradicts it by raising questions, providing explanations, and supporting evidence. Then, Agent 1 has to answer the questions raised by Agent 2 and tries to refine its own statements. During this debate-refine process, the agent can generate more desired and less flawed responses.

Our debate framework is different from previous ones, in which the opponent agent is not to support opposite stances to the other but to improve the other's response by identifying its weaknesses. Hence, in previous work, both agents share the same goal, reaching a consensus as a final ground truth for the LLM response, which greatly constrains LLMs' capability to generate responses both in breadth and depth. Moreover, reaching a consensus is non-trivial (Chen et al., 2023a), and always requires an additional Judge (Liang et al., 2023), Confidence Estimator (Chen et al., 2023b) or Summarizer (Chan et al., 2023), which not only introduces more computation but leads to potential instability as well. On the contrary, our pipeline simulates real-world debates, in which two agents holding different stances can freely question or contrast each other and they are not required to reach a consensus.

During the debate, two agents p_{θ_1} and p_{θ_2} are prompted to debate given a specified argument $a_{p_t,i}$[2] of a given topic t. The two agents can share the same model (i.e., $\theta_1 = \theta_2$) but are prompted to hold opposite stances on the argument. We assume that p_{θ_1} agrees with the argument $a_{p_t,i}$ and p_{θ_2} is against it. They debate on the topic for m rounds, where m is flexible and does not need to lead to a consensus. Since the argument index i is fixed in the debate, for simplicity, we will remove i when elaborating on the debate procedure. In each round, an agent is required to reply to the questioning of its opponent, refine its previous statement, and question the opponent. For example, after the first round, p_{θ_1} is prompted to generate the initial statement $s_{1,1}$ supporting the argument a_{p_t} by

$$s_{1,1} \sim p_{\theta_1}(s|t, p_t, a_{p_t}), \tag{1}$$

while $p_{\theta,2}$ generates the controversial statement

[2] It can be $a_{p_t,i}$ or $a_{n_t,i}$. We use $a_{p_t,i}$ as an example here.

$s_{2,1}$ based on its opponent's statement $s_{1,1}$, i.e.,

$$s_{2,1} \sim p_{\theta,2}(s|t, p_t, a_{p_t}, s_{1,1}). \quad (2)$$

After m rounds of debate, the agent p_{θ_1} is required to summarize and refine its final statement $s_{p_t,i}$ based on the entire debating process by

$$s_{p_t,i} \sim p_{\theta_1}(s|t, p_t, a_{p_t}, s_{1,1}, ..., s_{1,m}, s_{2,m}). \quad (3)$$

In the training phase, $s_{p_t,i}$ will be used as the ground truth of the controllable output $y_{p_t,i} \sim p_\theta(y|t, p_t, a_{p_t,i})$ of an LLM p_θ when given the topic t, stance p_t, and argument $a_{p_t,i}$ as controls.

3.2 Training

We then use the data collected via debate to build an instruction-tuning dataset, in which $(t, p_t, a_{p_t,i})$ is the instruction and $s_{p_t,i}$ is the corresponding response. To improve the controllability of an LLM p_θ on generating controversial statements such as $s_{p_t,i}$ and $s_{n_t,i}$ for different stances, we finetune p_θ on the instruction-tuning dataset by maximizing the following objective.

$$\max_\theta \sum_{t=1}^{T} \sum_{i=1}^{k} \Big[\log p_\theta \left(s_{p_t,i}|t, p_t, a_{p_t,i}\right) + \log p_\theta \left(s_{n_t,i}|t, n_t, a_{n_t,i}\right) \Big], \quad (4)$$

Compared to existing instruction-tuning datasets that mainly focus on covering a broad range of topics, we only utilize a limited number of topics while each topic containing $2 \times k$ samples covering k diverse arguments, 2 opposite stances per argument, and a high-quality salient statement for each stance. As shown in the experiments, our dataset significantly improves the LLM controllability.

4 Evaluation Metrics

We evaluate LLM's ability to generate statements for controversial topics on two orthogonal aspects on our hold-out test set: the **Response Quality** and **Controversy Controllability**. The Response Quality measures whether the LLM can generate helpful, relevant, accurate, and detailed statements for an instruction, which aligns with the common requirements for LLM's responses. However, the response quality fails to measure the extent to which LLM's response is stuck to the desired stances. As illustrated previously, existing LLM tends to generate average viewpoints endorsed by the majority and neglect the voice of the minority, which might be graded highly by existing judging methods. Thus we propose another aspect, noted as Controversy Controllability, which directly measures whether LLM's response is strictly stuck to the desired stances.

4.1 LLM Judge

Considering the large number of test samples, we utilize GPT4 as the judge for evaluation, which has become widely accepted, as noted in several studies (Touvron et al., 2023b; Chiang et al., 2023; Dettmers et al., 2023; Liu et al., 2023b; Chiang and Lee, 2023). Research has demonstrated that the evaluations made by GPT-4 align well with human judgments (Zheng et al., 2023; Li et al., 2023c). [3]

Response Quality:

The evaluation of Response Quality follows Chen et al. (2023c); Li et al. (2023b, 2024b), which involves a detailed rating system for the responses generated by the model. This system compares responses generated by two different LLMs on various dimensions, including helpfulness, relevance, accuracy, and level of detail. We also address the issue of positional bias in the LLM judge system, as discussed in the studies by Ko et al. (2020); Wang et al. (2023b) by presenting models' responses in two separate sequences for evaluation by the LLM judge. We then analyze the responses for each instruction by comparing them through a "Win-Tie-Loss" system. Then the win score will be calculated for better comparison:

$$\text{Score} = \frac{n_{\text{Win}} - n_{\text{Lose}}}{n_{\text{All}}} + 1, \quad (5)$$

Controversy Controllability:

For better illustration, we utilize **Positive/Negative** to illustrate the stance for a debate topic as shown in Figure 3. A Positive stance supports the topic sentence while a Negative stance is against the topic sentence. There can be diverse arguments under each stance. By changing the stance and the argument in the input, a controllable LLM should generate detailed and strong statements supporting the given stance and the argument, even if they only represent the minority's point of view. We utilize the **Good/Bad** pair to illustrate the success or failure of the LLM in generating the controllable statements. For example, if the LLM is prompted to support a topic and it successfully does so, then it is a good one;

[3]Both the detailed prompts for Response Quality and Controversy Controllability can be found in Appendix A.

if it generates responses against the topic, then it would be a bad one.

For the evaluation of Controversy Controllability, we prompt GPT4 to analyze the response with the given topic without letting it know the specific stance of this response and ask it to guess and provide the supporting versus opposing proportion of the above arguments to the given topic. This method serves as a relaxation that turns the original complex problem into a problem similar to sentiment analysis, which is perfectly under the control of the powerful GPT4 model. Ideally, in a good sample, the majority proportion should be 100% and is the same as the real given stance. Otherwise, it means this response fails to strictly stick to the user's query. Then we further categorize all the responses into Good or Bad ones and the Positive Controversy Controllability score is defined as the ratio of good samples in all positive samples, while the Negative one is the ratio in all negative samples. The Overall Controversy Controllability is the average of the Positive and Negative. The higher score represents the more samples are strictly stuck to their given stance, representing a better Controversy Controllability.

4.2 Human Study

To further compare the Controversy Controllability of our model and the baseline model, further human studies are conducted. Since there are 80 topics in our test sets, each of which contains 6 different arguments, resulting in a total of 480 query-response pairs, making it infeasible to manually inspect all the samples. Moreover, we empirically find the number of bad examples is few due to the current strong instruction-following ability of current LLMs, thus it is also infeasible to inspect only the small random set of testing samples.

To overcome this problem, we utilize an LLM-Human interactive inspection method. After utilizing LLM as the Judge for the Controversy Controllability evaluation, we select all the bad cases detected by GPT4, and then randomly sample some good cases to construct a new evaluation set with 100 instruction-response pairs. Then human participants are queried to judge whether these responses are strictly stuck to their given stances. There are 3 choices given, (1) Good, representing the response is strictly stuck to the given stance; (2) Bad, representing the response contains opposite content; (3) Tie, representing the response is hard to judge. We conduct this human study on both the baseline Vicuna 7B v1.5 and our DEBATUNE-7B.

5 Experimental Result

5.1 Results on Controversial Controllability

Table 1 showcases our main evaluation results on both the Response Quality and Controversy Controllability on the hold-out test set.

In the **Response Quality** section, we report the win-tie-loss statistics and corresponding win scores between other models and DEBATUNE-7B. The overall ranking of different models on the Response Quality basically aligns with their performance on common instruction-following benchmarks. For example, the Alpaca has the lowest win score, and commonly believed better models have relatively higher response quality scores.

However, when it comes to **Controversial Controllability**, which measures the extent to which LLM's responses are stuck to the given stances, the results are not directly correlated to the original ability of LLMs, which reveals an interesting but long-neglect phenomenon. Under this setting, LLaMA2 Chat models achieve the lowest controllability scores, reasonable due to their strongly constrained alignment. Given a controversial topic, they have a strong tendency to refuse to answer or to find a safe middle ground to avoid potential harm. Though this strong alignment potentially avoids the offensiveness, it also loses the possibility to speak for the diverse perspectives. On the contrary, the Alpaca model achieves the highest score on controllability, indicating that they can provide statements strongly stuck to the given stances while having the lowest response quality. However, the manual inspection further explains this phenomenon that it is because of the relatively low instruction-following ability, that Alpaca tends to repeat the given argument with only a little new content, thus leading to high controllability and low quality.

According to the above analysis, we can see both of these two criteria play an important role in evaluating the LLM's ability to generate statements for controversial topics. As shown in the results, our DEBATUNE, achieves the highest scores on both aspects compared with existing models, indicating our model's ability to speak for the minority. This Controversial Controllability metric proposed by us provides another dimension to examine the capability of current LLMs, pushing forward the understanding of their limitation and capabilities.

In the **Human Study**, there are 100 samples ex-

	Response Quality (model vs. baseline)				Controversy Controllability		
	Win↑	Tie	Loss↓	Win score↑	Positive ↑	Negative ↑	Overall ↑
DEBATUNE-7B (ours, baseline)	-	-	-	1.00	0.958	0.979	**0.969**
DEBATUNE-13B (ours)	43	101	16	**1.17**	0.950	0.946	0.948
Alpaca 7B (Taori et al., 2023)	2	1	157	0.03	0.938	0.883	0.910
WizardLM 7B (Xu et al., 2023)	3	12	145	0.11	0.833	0.704	0.768
WizardLM 13B V1.2 (Xu et al., 2023)	14	99	47	0.79	0.800	0.708	0.754
Vicuna 7B v1.5 (Chiang et al., 2023)	6	19	135	0.19	0.900	0.796	0.848
Vicuna 13B v1.5 (Chiang et al., 2023)	5	36	119	0.29	0.867	0.858	0.863
LLaMA2 Chat 7B (Touvron et al., 2023b)	1	17	142	0.12	0.196	0.429	0.313
LLaMA2 Chat 13B (Touvron et al., 2023b)	3	26	131	0.20	0.338	0.317	0.327
Zephyr 7B Alpha (Tunstall et al., 2023)	7	29	124	0.27	0.879	0.713	0.796
Zephyr 7B Beta (Tunstall et al., 2023)	12	84	64	0.67	0.942	0.733	0.838

Table 1: **Response Quality** (DEBATUNE-7B as the baseline) and **Controversy Controllability** of our models and other LLMs on generating statements for controversial topics. DEBATUNE archives the highest quality and controllability, indicating its effectiveness on generating controllable responses for controversial topics.

amined generated by the baseline Vicuna 7B v1.5 and our DEBATUNE-7B. For our model, $87/100$ samples are inspected as Good, $2/100$ as Ti.e., and $11/100$ as Bad samples. For the Vicuna model, $40/100$ samples are inspected as Good, $7/100$ as Ti.e., and $53/100$ as Bad samples. In this human study, we carefully examined all the GPT-4 labeled bad examples by human experts and observed a similar and high ratio of human-labeled bad samples within the GPT4-labeled bad samples. Moreover, we also examined a random set of 100 good samples labeled by GPT4 and almost all of them are indeed good samples for human experts. The large discrepancy between Vicuna and our model further verifies our method and the high consistency between human evaluation and LLM evaluation verifies the effectiveness of our evaluation method.

5.2 Ablation studies

In this section, ablation studies are conducted to verify the configuration of our method. All experiments are conducted on the LLaMA-7B model. During the comparison, Vicuna-7B v1.5 is utilized as the baseline model as it is trained with diverse ShareGPT data containing real human queries. The results are shown in Table 2.

The upper section of the table showcases the experiments with different debate configurations, 3 arguments are utilized for each stance of a given topic by default. "Topic Data without Debate" represents the model trained directly with the training split of our controversial topics, whose response is generated from *gpt-3.5-turbo-1106* without debate. We can observe clear improvements in both the Response Quality and Controversy Controllability, indicating a rise in the capability of sticking to the given stance, which directly proves the effectiveness of our collected data.

"x-round Debate on each Topic" represents the model trained with debate-augmented responses for training. From the results, we can observe that even a one-round debate can significantly improve our model's capability on both two metrics. During the debate, the involved agent is required to strictly stick to the given stance, otherwise will be rebuked by the opponent. Then after rebuttal, the agent is able to further refine its previous response. This debate process improves the responses to controversial topics in 2 aspects: 1. This process is naturally an interactive refinement process, thus continuously polishing the response itself, guaranteeing good response quality, which is proved by Du et al. (2023); Liang et al. (2023). 2. This debate process requires the agent to think of the potential opposing responses and answer them in advance, and this thinking pattern increases the controversy controllability, similar to Mukherjee et al. (2023); Mitra et al. (2023), which also tries to distill thinking patterns to student models.

In the upper section, it is observed that a 2-round debate is the optimal setting, and thus extensive experiments are conducted as shown in the lower part aiming to find the optimal number of arguments for each stance of the given topic. The 3-Argument setting marginally outperforms the other options, thus we continuously set it as our default setting.

5.3 Results on Instruction Following

In addition to the main results on our hold-out test set, evaluating the ability to generate statements for controversial topics, we also propose that our method can improve the general instruction fol-

	Response Quality (vs. Vicuna 7B v1.5)				Controversy Controllability		
	Win↑	Tie	Loss↓	Win score↑	Positive ↑	Negative ↑	Overall ↑
ShareGPT (Vicuna 7B v1.5, baseline)	-	-	-	**1.00**	0.900	0.796	0.848
Topic Data without Debate (3 Arguments)	118	27	15	1.64	0.946	0.813	0.879
1-round Debate per Topic (3 Arguments)	134	19	7	1.79	0.954	0.950	0.952
2-round Debate per Topic (3 Arguments)	135	19	6	**1.81**	0.958	0.979	**0.969**
3-round Debate per Topic (3 Arguments)	135	17	8	1.79	0.967	0.963	0.965
1 Argument per Topic (2-round Debate)	135	21	4	**1.82**	0.946	0.921	0.933
3 Arguments per Topic (2-round Debate)	135	19	6	1.81	0.958	0.979	**0.969**
5 Arguments per Topic (2-round Debate)	135	18	7	1.80	0.933	0.933	0.933

Table 2: **Ablation study** on the number of debate rounds and the number of arguments per (topic, stance). Response Quality (Vicuna 7B v1.5 as the baseline) and Controversy Controllability are reported. It verifies the optimality of the default setting, i.e., 2-round debate and 3 arguments per topic.

	Pair-Wise	Huggingface Open LLM Leaderboard					Alpaca Eval	MT Bench
	Win Score	**Average**	ARC	HellaSwag	MMLU	TruthfulQA	Win Rate	Score
Vicuna 7B v1.5	1.000	57.95	53.24	77.39	50.82	50.33	73.10	6.07
+ 1-Arg (2-round)	1.220	**58.47**	54.10	77.20	51.17	51.40	**79.70**	5.90
+ 3-Arg (2-round)	**1.257**	57.77	52.56	76.54	51.08	50.91	78.76	**6.13**
WizardLM 7B	1.000	57.09	54.18	79.25	46.92	48.01	66.08	5.56
+ 1-Arg (2-round)	**1.372**	**57.72**	54.69	78.61	46.96	50.62	**74.04**	5.57
+ 3-Arg (2-round)	1.339	57.46	54.86	78.12	46.94	49.90	71.20	**5.70**

Table 3: Evaluation of DEBATUNE-trained models on three widely used benchmarks and pairwise comparison with the baseline models. By using only 630 topics, DEBATUNE achieves consistent improvements on two different base LLMs and different evaluation metrics.

lowing the ability of LLMs. To verify this, we directly finetune the Vicuna 7B v1.5 and WizardLM 7B (based on LLaMA2) models using the debate-augmented training set, containing 630 topics, without data from any other sources. Then we evaluate our model on 4 different commonly used methods, including **Pair-Wise Comparison**, **Huggingface Open LLM Leaderboard**, **Alpaca Eval Leaderboard** and **MT Bench**. [4]

As shown in Table 3, the model further trained with our data outperforms the baseline models on all of the 4 different evaluation metrics on two different models. It is worth noting that only 630 topics are utilized, indicating the neglectable new knowledge involved in the training, while it causes a consistent improvement in the general instruction-following ability. We believe this is because of the high-quality responses generated during the debate mechanism. After the debate, the responses contain detailed statements that are strongly aligned with the controllable queries, this strong alignment further catalyzes the instruction-following ability of the models (Li et al., 2024a; Xu et al., 2024).

[4]The evaluation metrics will be introduced detailedly in the Appendix C.

6 Related Work

6.1 LLM Alignment

Despite the advancements of the current LLM, a fundamental issue with LLMs is the disjunction between their training objectives (i.e., minimizing contextual word prediction error), and users' aspirations for models (i.e., interpret and execute instructions reliably (Radford et al., 2019; Brown et al., 2020; Fedus et al., 2022)). To reconcile this, recent NLP research efforts focus on empowering LLMs to understand instructions and to align with human expectations, i.e., Instruction Tuning (Ye et al., 2021; Wei et al., 2022; Wang et al., 2022; Du et al., 2022; Honovich et al., 2023; Taori et al., 2023; Chiang et al., 2023; Liu et al., 2023a).

Another significant challenge in developing language models is ensuring that their output is useful, accurate, and consistent with human ethical standards (Kenton et al., 2021; Weidinger et al., 2021; Askell et al., 2021; Wang et al., 2023c). A common method to achieve this involves engaging human raters to evaluate and compare the outputs of these models (Bai et al., 2022a; Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019). This feedback is crucial for improving the model's effectiveness in various tasks such as following instructions and answering questions. Recently the feedback from AI (Bai et al., 2022b; Lee et al., 2023) also benefits the alignment of LLMs. In the case of large-scale models, this method has been shown to enhance performance on specialized datasets aimed at assessing model alignment.

6.2 Debate between LLMs

With the continuous revealing of the self-improving ability (Huang et al., 2023; Madaan et al., 2023; Ye et al., 2023; Li et al., 2023a) of LLMs, a Multiagent Debate framework (Du et al., 2023; Liang et al., 2023) is proposed to further improve the responses of LLMs. The motivation of these methods is to reach a consensus for a given instruction, thus always requiring an additional Judge (Liang et al., 2023), Confidence Estimator (Chen et al., 2023b) or Summarizer (Chan et al., 2023). How to effectively reach the consensus in the debate framework is non-trivial (Chen et al., 2023a) and still under exploring. Moreover, this debate framework is further used in the evaluation of LLMs (Wang et al., 2023a; Chan et al., 2023), and helps non-expert judges identify the truth (Michael et al., 2023).

6.3 Debate Datasets

The exploration of debate-related datasets NLP has yielded significant resources, each contributing uniquely to the advancement of debating systems. DebateSum (Roush and Balaji, 2020) is a large-scale dataset that includes a rich collection of debate documents with high-quality arguments, facilitating a variety of NLP tasks, especially argument mining and summarization, while the direct debate topic is not provided. Change My View (Hidey et al., 2017) focuses on the effectiveness of arguments in changing viewpoints and SOCIAL-CHEM101 (Forbes et al., 2020) focuses on social norms, both of which are not suitable for debate. Argument Reasoning Comprehension Task (Habernal et al., 2018) focuses on identifying and reconstructing implicit warrants in arguments. Moreover, IBM Project Debater [5] (Shnarch et al., 2020; Ein-Dor et al., 2020; Levy et al., 2018; Shnarch et al., 2018; Gleize et al., 2019; Toledo et al., 2019) also leads to the creation of diverse NLP datasets spanning various categories.

7 Conclusion and Future Work

Our study raises the long-neglect issue of generating controllable responses towards controversial topics, not only describing and exemplifying but also building an evaluation pipeline for the assessment and a novel method to alleviate this problem. More specifically, our DEBATUNE, is a novel pipeline that enhances model controllability over diverse perspectives on controversial topics. We have curated the largest dataset of debate topics to date and introduced a new metric for measuring controllability. Our evaluations reveal that LLMs can be effectively fine-tuned to represent a broader spectrum of opinions, paving the way for more inclusive AI-generated discourse.

In our study, it is shown that controversial arguments and statements are beneficial for LLMs in further generating diverse and high-quality responses supporting different controversial topics. However, our current method mainly focuses on generating desired statements by utilizing LLMs, though effective, a better strategy would be directly collecting high-quality human-written responses. One potential source would be directly collecting statements directly from the debating websites. Another potential source would be reforming the existing high-quality DebateSum dataset (Roush and Balaji, 2020). As mentioned in the previous section, DebateSum has diverse and high-quality statements while lacking corresponding topics and stances, thus further human annotation is required. We believe that the further collecting and combining of human-written high-quality data can further improve LLM's controllability to generate responses for diverse controversial topics and thus help minorities to express their own opinions.

8 Ethical Concerns

To address ethical concerns, the LLMs used for debate are required to be the ones trained by safety alignment so they cannot generate toxic content or content with ethical concerns. These LLMs are required to pass the safety test before being deployed for debate. We will further apply an output filter to the debate-generated statements to double-confirm the safety of the finetuning data before using them for training LLMs. This will address the ethical concerns and meanwhile preserve the controllability of the resulting LLM and its capability to speak for minority groups. In addition, improving LLM's controllability reduces the uncertainty of LLM outputting unexpected harmful content so humans can more effectively enforce the safety constraints directly through the "controls" in the input.

[5] https://research.ibm.com/haifa/dept/vst/debating_data.shtml

Limitations

The main limitation of this work is the lack of studies on the agent used for debate. This work only considers the setting where ChatGPTs are utilized as the debate agents to ensure the quality of the responses. However, it would be more interesting to know if this debate framework can be effectively utilized on the existing relatively weak LLMs.

References

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Michiel A. Bakker, Martin J Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023a. Multi-agent consensus seeking via large language models.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023c. Alpagasus: Training a better alpaca with fewer data.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Shiyao Ding and Takayuki Ito. 2023. Self-agreement: A framework for fine-tuning language models to find agreement among diverse opinions.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691. AAAI Press.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813. AAAI Press.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024a. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *ArXiv*, abs/2402.10110.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, and Tianyi Zhou. 2023a. Reflection-tuning: Recycling data for better instruction-tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *ArXiv*, abs/2402.00530.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023b. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *ArXiv*, abs/2308.12032.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. 2023. Debate helps supervise unreliable experts.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Long Ouyang, Jeffrey 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Marios Papachristou, Longqi Yang, and Chin-Chia Hsu. 2024. Leveraging large language models for collective decision-making.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings* *of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023c. Aligning large language models with human: A survey.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin

Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Prompt for Evaluation

The prompts for evaluating the **Response Quality** and **Controversy Controllability** are shown in Figure 4 and Figure 5, respectively.

Prompt for Response Quality

System Prompt
You are a helpful and precise assistant for checking the quality of the answer.

User Prompt
[Question]
Question
[The Start of Assistant 2's Answer]
Answer 2
[The End of Assistant 2's Answer]
[The Start of Assistant 2's Answer]
Answer 2
[The End of Assistant 2's Answer]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Figure 4: The prompt to evaluate the Response Quality.

Prompt for Controversy Controllability

System Prompt
You are a helpful and precise assistant for checking the potential bias of arguments.

User Prompt
[Topic]
Topic
[The Start of Arguments]
Arguments
[The End of Arguments]

We would like to request your feedback on the supporting versus opposing proportion of the above arguments to the given topic. Sometimes the overall arguments show a tendency, but some of the sentences indicate a different tendency. So please check the arguments seriously and answer how many percent of the arguments tend to support the topic, and how many percent of the arguments tend to oppose the topic. The scale is from 0 to 100.
Please first output a single line containing only two values indicating the percentage of supporting and opposing proportions, respectively. The two scores are separated by a space. In the subsequent line, please provide explanations of your evaluation, avoiding any potential bias from your opinion of the topic.

Figure 5: The prompt to evaluate the Controversy Controllability.

B Implementation Details

In the debate process, we utilize *gpt-3.5-turbo-1106* as the default debating agent and the number of debating rounds is set to 2 by default. For each stance on the topic, 3 arguments are used for the process and for the training. We train our model based on both initial pretrained LLaMA2 (Touvron et al., 2023b) and Vicuna 7B v1.5 (Chiang et al., 2023). For the LLaMA2-based model, the learning rate is set to 2×10^{-5} while 1×10^{-5} for Vicuna-based model. The batch size is 128, steer the training across 3 epochs with a max length of 2048. The warmup rate is set to 0.03.

C Instruction-Following Evaluation

C.1 Pair-wise Comparison

The setting and the prompt for the pair-wise comparison are the same as the evaluation for Response Quality as shown in Figure 4. The comparison is conducted on the WizardLM dataset (Xu et al., 2023), which contains 218 unique instructions, by utilizing GPT4 as the judge. The detailed comparison results are shown in the Table 4

	Win	Tie	Lose	Win Rate
Vicuna 7B v1.5	-	-	-	1.000
+ 1 Arg (2-round)	98	70	50	1.220
+ 3 Arg (2-round)	99	76	43	1.257
WizardLM 7B	-	-	-	1.000
+ 1 Arg (2-round)	119	61	38	1.372
+ 3 Arg (2-round)	111	70	37	1.339

Table 4: The pair-wise comparison between the debate-augmented model with the baseline models.

C.2 Open LLM Leaderboard

The Hugging Face Open LLM Leaderboard represents a cutting-edge initiative designed to showcase the performance of various LLMs across a wide array of benchmarks (Gao et al., 2021). It functions as a comprehensive and transparent platform where researchers and developers can compare the capabilities of different models based on standardized testing criteria. This leaderboard not only facilitates an objective evaluation of models in terms of natural language understanding, generation, and other AI tasks but also encourages the development of more efficient, accurate, and versatile language models. It focuses on 4 pivotal benchmarks: ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and TruthfulQA (Lin et al., 2022).

C.3 Alapca Eval Leaderboard

The AlpacaEval Leaderboard provides a specialized platform for the automatic evaluation of Large Language Models (LLMs) using the AlpacaFarm evaluation dataset, as outlined in Dubois et al. (2024). This system offers an efficient and reliable method for assessing LLMs based on their ability to follow general user commands. Comparing model outputs with standard responses provided by Davinci003 ensures a comprehensive analysis. The system's effectiveness is highlighted by its strong correlation with human expert judgments, showcasing its accuracy in mirroring real-world expectations and the models' adherence to precise user instructions.

C.4 MT-bench

MT-Bench, the Multi-turn Benchmark proposed by Chiang et al. (2023), serves as a rigorous framework for evaluating the conversational prowess of LLMs. It aims to measure how well these models can maintain coherent, informative, and engaging dialogue over multiple turns of conversation. This benchmark tests models on their ability to follow instructions and flow naturally in conversations, making it a crucial tool for assessing their performance in realistic dialogue scenarios. By focusing on the dynamic aspects of conversation, MT-Bench addresses a critical need in the AI community for benchmarks that can accurately reflect the capabilities of LLMs in engaging with users in a manner that mimics human conversation.