

# Both Sides of the Coin - An Explorative Investigation in LLM Debates

By

Linus Krause & Leander Streun

## Abstract

In today's media landscape, the proliferation of diverse perspectives via the internet has led to both opportunities and challenges. The rise of opinion diversity has also brought about an increase in misinformation and the formation of echo chambers. As technology continues to evolve, there is ongoing debate among researchers about whether future advancements will exacerbate or mitigate these issues. Natural Language Processing (NLP) and Large Language Models (LLMs) offer potential solutions to address misinformation and enrich debates by providing different perspectives.

This paper presents a project aimed at investigating the feasibility of using LLMs to tackle opinion singularity in media articles through debates. Specifically, the study employs a Multi-Agent Debate (MAD) approach, wherein two LLMs engage in a debate over polarized stances on various societal topics. The research explores the initialization of personas for the LLMs, the dynamics of debate convergence, and the interaction between debaters over multiple rounds.

Using both an open-source model (Llama2-Chat-7b) and Chat-GPT (gpt-3.5-turbo-0125), the study conducts discussions on eleven topics spanning societal issues such as climate change, gun control, and taxation. The research methodology involves creating polarized stances for each topic, initializing personas for the LLMs, and structuring debates over multiple rounds.

Results indicate that both models successfully engage in debates over polarized stances, with differences observed in convergence behavior and agreement levels. While Chat-GPT exhibits stable convergence in the majority of debates, Llama2 shows less divergence but also lacks stability across rounds. Additionally, the study presents an exploratory approach to measure opinion distances using a Likert scale, providing insights into the dynamics of argumentation between LLMs.

The dataset generated from this study serves as a valuable resource for future research in NLP and debate analysis. Overall, this paper contributes to the understanding of LLMs' capabilities in facilitating diverse perspectives and fostering informed discussions in media articles.

## 1. Introduction

Today's media landscape is characterized by a variety of different perspectives. Through the internet the expression of one's opinion isn't limited to journalists or media institutions (opinion singularity). Nowadays nearly everybody has the possibility to open a blog to spread their perspectives, which amplifies the development of opinion diversity in the internet. But this diversification also led to increased misinformation and fake news as well as the emergence of so-called "echo-chambers", where individuals surround themselves mainly with opinions similar to their own opinion (source). If future technology developments will account for a mitigation or worsening of these problems is currently debated among researchers (Anderson and Rainie 2017).

Natural Language Processing [NLP] and Large Language models [LLMs] may account for possible solutions to surveil or address misinformation in the internet (Leite et al. 2023) or to enrich debates, contributing different perspectives. Already, tools have been developed to detect check-worthy statements in media articles, like for example ClaimBuster (Arslan et al. 2020). But LLMs may also be interesting in the context of reflecting over a given context/opinion and providing feedback to it. Researchers have already investigated in the potential abilities of self-reflection or multi-agent debates [MAD] of LLMs. Regarding self-reflection, a core problem constitutes the *Degeneration-of-Thoughts* (DoT), once a solution has been put forward by the LLM, it establishes confidence in it and is unlikely to deviate away from it (Liang et al. 2023). Whereas MAD it has been found for certain models to possess the ability to improve their performances over multiple rounds, therefore being able to correct and beneficially contribute to a debate (Fu et al. 2023).

This project paper searches to investigate in the ground-work for a potential tool to tackle opinion singularity in media articles, by letting two Large Language Models

[LLM] participate in a debate over a specified topic. By encountering a diverse discussion, the end user should be able to know and understand different points of view and potentially find common ground with positions he before would have thought of as non-acceptable, idiotic or in any other way excluded from the hypothesis space of solutions to the given problem or topic. This could be of use for phenomena of opinion singularity.

To do so we decided on a variety of society dividing topics and created two polarized stances for each topic. By initializing two LLM personas, following these stances, we tried to create two LLM-agents depicting each side of a debate. The two polarized perspectives were juxtaposed in a debate context, spanning three rounds, where in each round one side tries to put forward their perspective. Ultimately, we exploratively tried to measure the distance of the two arguments stated by each side. Furthermore, we created a dataset of opposing debate statements, following our selected topics and developed stances.

## 2. Methodology

### General research setup

The structural setup of the discussion is the following:

- 11 topics (see Appendix)
- 2 models
- 2 or 3 rounds of discussion
- 3 runs
- 2 prompt versions (few vs zero shot)

We decided on these 11 topics with the goal to span a wide range of topic categories being relevant to society and potentially leading to polarization. Continuing from these topics we decided for two opposite stances and manually created an initial description of the characteristics of the stances. Afterwards we used ChatGPT (GUI) to create personas by describing our desired goal.

First, we created two personas for the debate about “Climate Change” and used the results for one shot prompting the following topics. Each persona was then manually evaluated and checked if the stance is accurately represented by the persona, decided by the researchers themselves. Finally, we use the topic, stance, and personas to create a system prompt which will be given to the LLM. For the Llama 2 model we use two setups of system prompts, a zero shot and a one-shot approach, giving an example of how the debate should look like. In both of these approaches we first describe the task, then clarify the belief systems and

the participants in the discussion (debaters = stances). We also provide discussion rules to encourage not only bringing the own belief system forward but also counter arguing the arguments of the opposition. All the material created, and prompts used can be accessed in the Appendix (Link to google docs).

Each discussion goes for multiple rounds (2 for open source 3 for paid service) and for multiple runs. A run refers to rerunning the exact same discussion multiple times for the full number of rounds. Runs are done from start to finish and not for each round of discussion.

Finally, we will use the paid version (Chat-GPT, for model version see below) to create additional runs to provide a dataset for future research.

### Models

#### Open-source model (Llama2-Chat-7B)

Although the training methodology for LLM is simple, high computational requirements have limited the development to a few players (Touvron et al. 2023). In order to provide an alternative to the currently dominant player OpenAI and its flagship model Chat-GPT (at this point GPT 3.5 turbo) we also implemented a debate system using an open-source model. We decided on Llama2, a well performing model on multiple Benchmarks created by Meta. In the following we will go over some of the characteristics of Llama2 which led to our decision.

First of all, Llama2 is an open-source model, even though in order to get access to it via the official Huggingface website one needs to register and state the intended use case.

Llama2 comes with a fine-tuned chat version which has been optimized to return safe and helpful responses (Llama2-Chat), using Reinforcement Learning with Human Feedback (RLHF). This characteristic is important to us, because the model shouldn’t be misused for initializing racist or extremist personas and therefore normalizing malicious belief systems. The optimization for safety and helpfulness is a tradeoff optimization, which means restricting speech to the fullest might lead to very safe responses but also excluding very helpful responses regarding any sensitive or personal topic. In this work we will touch on some very emotionally discussed topics and are interested if the different stances are accepted or rejected by guardrails.

Llama2-Chat is also trained on being context aware by introducing a technique called Ghost Attention. The researchers use this technique to control dialogue flow over multiple turns. The vanilla Llama2 model would simply “forget” some instructions that should be applied for all the conversation turns (e.g., to “act as” some public figure) af-

ter a few turns of dialogue. If such instructions are provided to Llama 2-Chat, the subsequent response should always respect the constraint.

We decided to use the 7B parameter version of Llama2-Chat, because the idea of using an open-source model in the first place was to make this paper more reproducible and accessible to a wider audience. More parameters would lead to a further increase of necessary resources and therefore lead to be more exclusive.

Llama2-Chat has a context window of 4096 tokens. While there have been new versions of Llama2 created to have a bigger context window of 32k, we decided that Llama2-Chat's benefits explained above are of bigger value to us. Due to the reduced window size the discussion using Llama2-Chat-7B was shortened to 2 runs.

There are several other open-source models to consider, like open-chat 3.5, Mistral and Mixtral, Sora and Zephyr-beta. Multiple of these models outperform Llama2-Chat-7B on so-called "Benchmarks". Our primary focus for Benchmarks lies on the human perception of the given answer regarding helpfulness, preference, and safety (low toxicity). One Benchmark to consider is the LMSYS Chatbot Arena. It is a crowdsourced open platform for LLM evaluations with over 400,000 human preference votes collected. Even though other models "outperform" Llama2-Chat-7B on this benchmark it is important to notice that the Benchmark is calculated by pairwise comparisons with one user decided winner. The best performing model at this point is "gpt4-1106-preview" which still can "only" decide 70% of direct comparisons for itself. Therefore, it doesn't seem like there is a clear preference for any model regardless of topic or prompt.

We also want to emphasize the good documentation of Llama2 (and its chat version) which contributed to our decision and its ease of accessibility.

In the end a few test-tries lead to our final decision on personal preferences. We encourage others to try out other open-source models and state their results.

### **Chat-GPT (gpt-3.5-turbo-0125)**

In the following we will reference gpt-3.5-turbo-0125 by Chat-GPT. We access Chat-GPT via API call. We use python for this and the OpenAI python library version: 0.28. Using the API is a paid service provided by OpenAI. There are different models to select from. We decided to use gpt-3.5-turbo-0125 due to its tremendously lower cost compared to GPT4 and also GPT4-Turbo (see Appendix/docs\_link). GPT-3.5 Turbo models are capable and cost-effective and come at this point in two versions: GPT-3.5-turbo-0125 is the flagship model of this family, supports a 16K context window and is optimized for dialog.

GPT-3.5-turbo-instruct is an Instruct model and only supports a 4K context window.

We decided for GPT-3.5-turbo-0125 in the direct comparison due to its larger context window size to explore longer discussions compared to Llama2-Chat-7b.

## **Hardware**

Due to the computational demand of LLMs we tried different external openly accessible computation resources which use GPUs. Both Kaggle and GoogleColab were not sufficient in their free version to run Llama2-Chat-7b and the large and with every iteration growing input. Therefore, we accessed the paid Version of GoogleColab for 11\$ a month, offering us 100 compute units and "access" to different types of GPUs and TPUs. For our use case only the A100 with High RAM (40GB) was sufficient. At peak workload we used 23GB of RAM while still cleaning up the cache with every iteration. It must be noticed, while the A100 High RAM is theoretically available it can take several minutes (if one is lucky) to be granted access because computational resources are distributed by GoogleColab and the user has no guarantee to access the desired processing unit. With testing multiple open-source models for a few iterations plus generating data for our hypothesis we were left with 4.51 out of 100 processing units after finishing our discussion. 4.51 processing units would last for about 17 minutes using the A100 High RAM.

## **Distance metric and discussion analysis**

We decided to use manual evaluation and an experimental approach of using Chat-GPT for "main-points-of-difference" extraction and a judgment of distance based on a Likert scale. The idea arose from the fact that human beings can easily judge if two debaters disagree, while embedding distance-based approaches will fail due to the relatedness of words. We tested multiple approaches.

While there are good approaches using sentiment analysis to measure distance, the example of "Climate Change" shows that sentiment towards a problem is not sufficient to detect differences between opinions or solutions.

Also, semantic distance (search) models like SBERT do not work for our use case. While SBERT can detect some types of similarity, it's not very suitable for polarity.

There are optimized classifiers for stance detection but are fine-tuned classifiers for a specific set of stances, e.g. left and right. Many discussions might not be separable in the classes and therefore lacks real world application (at least in our case).

We will provide a dataset containing multiple runs of discussion rounds to allow and support the development of new approaches.

### 3. Results

In this project paper we analyzed the behavior of LLMs debating over 11 topics, using a Multi Agent Debate approach. Since we focused mainly on the creation of debates between two polarized belief systems, the dataset of 100 debates over 2 (open source) / 3 (Chat GPT) rounds, featuring 11 topics, can be seen as the main contribution of this paper. As stated in the previous section we encourage further research on the determination of opinion/argument distances in NLP, we hereby provided with the dataset a suitable tool.

Having that said, we still want to provide an explorative approach to measure the opinion-distances for an extract of our dataset. The judge assessed the first 30 debates (runs) of the closed-sourced model and the first 3 debates (runs) for the open-sourced model. Using a Lickert scale spanning from 0 (very similar opinions) to 5 (very different) we used ChatGPT as a judge to compare the arguments stated by the two LLMs over the course of the debate. Furthermore, we told the judge to point out the main points of difference between the two statements. In this section we want to present our findings.

Our research questions are split up into two parts. Research questions regarding the Personas and questions regarding the debate.

#### Personas:

##### RQ1a & b:

The idea being that the specification for the different models potentially doesn't allow for a specific stance. That is, for example for "gun control" the two agents would immediately turn towards one specified stance. We find that there are no topics or personas explicitly rejected by either Llama2 or by Chat-GPT, indicating that the stances selected for our topics don't express positions too extreme to interfere with guardrail rules.

##### RQ2:

To see if the initialization of the two stances worked, we compared the initial distances of the two LLMs, that is what distance between arguments assigns the judge for the first round of debate. While the results were promising for Chat-GPT using one-shot prompting we used it as the default option. For Llama2, using the one-shot prompt sometimes led to a discussion about the topic given in the example part in the middle of the discussion. Therefore, for Llama2 we switched to zero-shot learning. We present our findings in Figure 1 & 2.

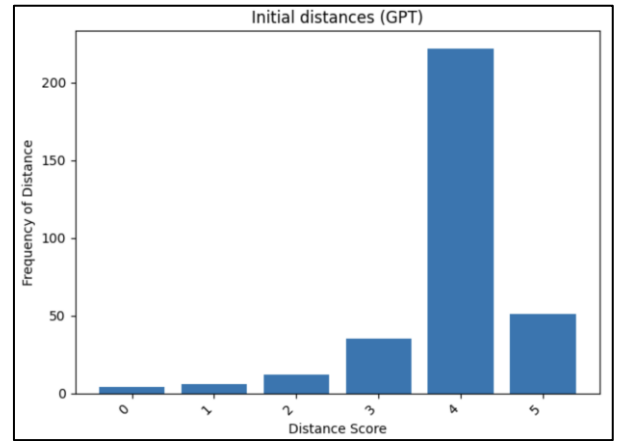


Figure 1: Initial distance (ChatGPT)

Figure 1 (Initial Distances ChatGPT) shows how often, for all 11 topics regarding only the two statements from the first debate round, the judge assigned a specific distance score to the argument distance. We can see that for all 330 cases (11 topics x 30 debates x 1 round), 308 times the judge assigned a distance of 3 or higher. That indicates that our stances initialized a debate starting for the huge majority of times from an opinion wise distant point of view. We therefore conclude that the initialization of stances did work out.

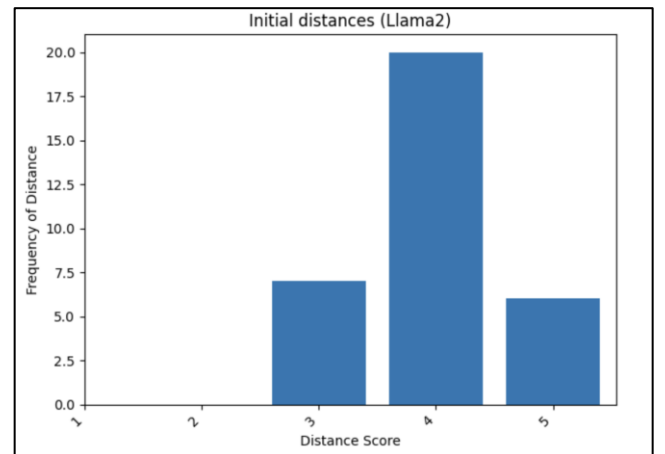


Figure 2: Initial distance (Llama2)

In Figure 2 (Initial Distances Llama2) the distribution of the same investigation for the open-source model can be seen. For all the 33 cases (11 topics x 3 runs x 1 round) the judge assessed them to have a distance from 3 or above. Therefore, we also conclude for the open-source model that the initialization worked out.

## Debates:

### RQ3:

The idea being addressed by this research question is, how stable does a personas initialization remains over the course of several debate rounds.

### GPT

For the GPT closed model we observed in 74 debates (runs) a stable behavior, that is the distance metrics for the last and the first round was the same. In 53 debates (runs) the two personas diverged (the last round distance value > first round distance value) We observed a converging behavior (last round distance value < first round distance value) in 203 debates (runs) (~66%). On average (over all the runs) the debaters converged (given that they converged) 1.95 metric points. The divergence (calculated exactly the same way) was lower on average -1.54. Interestingly we can report that the only topic where no run resulted in a high agreement (score = 0) is speech control. All other topics have been labeled to highly agree at a certain point in a given run. A follow up analysis showed that the model converged often (20 out of 30) for “speech control”. In general, there are 26 debate rounds of high agreement and 93 of high disagreement.

### Llama2

For the open-source model Llama2 we report no divergence at all. Furthermore, we couldn’t observe stability over the two rounds. On average the open-source model converges 1.97 points. Contrary to the GPT model (Figure 3) the Llama2 model only once provides a case of high agreement (about “gun control”). There are 9 cases of high disagreement, none of these occurred in the following topics: [climate\_change, taxation, technology\_security speech, markets]

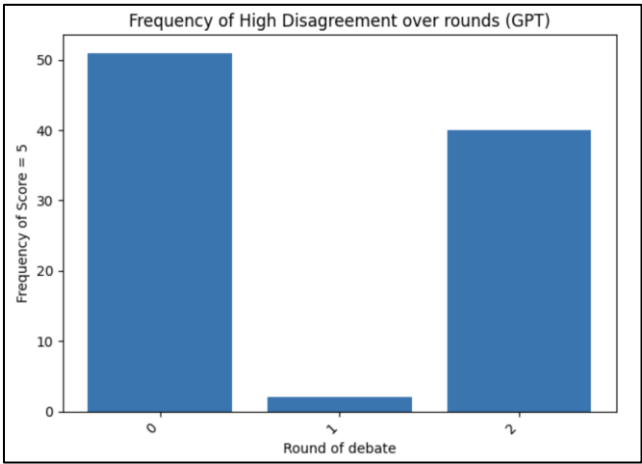


Figure 3: Frequency of High Disagreement over rounds (GPT)

In comparison it seems like the Llama2 model accepts polarized stances more often but tends to converge in a similar manner like the GPT model (identical rates of convergence). A potential interpretation of the nonexistence of high agreement scores for the Llama2 model, whereas the GPT model provides high agreement cases, is that GPT is highly trained to be customer oriented and non-conflictive.

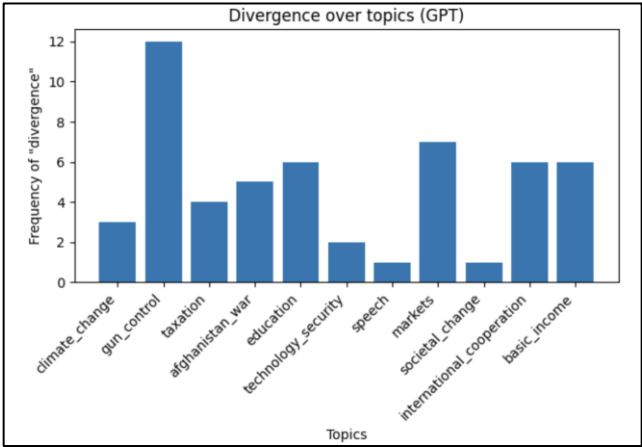


Figure 4: Divergence over topics (GPT): last round distance value > first round distance value

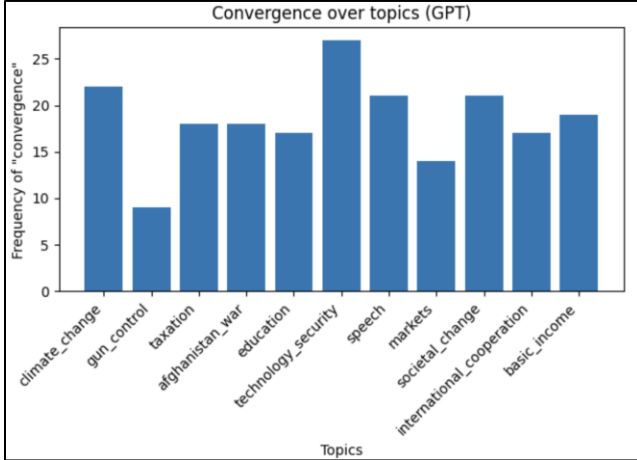


Figure 5: Convergence over topics (GPT)  
last round distance value < first round distance value.

#### 4. Future Research

Future research could use different open-source models and try different prompts to initialize the personas. Our provided dataset provides a large contribution for the development of distance measurements of natural language discussion. As seen in our work the relatedness to the same topic and paraphrasing of the opposition's arguments make it difficult to find a clear way to distinguish the results. Future research should focus on the development of a distance metric because an appropriate distance metric could allow for a higher degree of control on each round of debate and the distance debaters should keep.

#### 5. Sources

- Anderson, Janna, and Lee Rainie. 2017. 'The Future of Truth and Misinformation Online'. *Pew Research Center: Internet, Science & Tech* (blog). 19 October 2017.  
<https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/>.
- Arslan, Fatma, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. 'A Benchmark Dataset of Check-Worthy Factual Claims'. arXiv.  
<http://arxiv.org/abs/2004.14425>.
- Fu, Yao, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. 'Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback'. arXiv.  
<http://arxiv.org/abs/2305.10142>.
- Leite, João A., Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. 'Detecting Misinformation with LLM-Predicted Credibility Signals and Weak Supervision'. arXiv.  
<http://arxiv.org/abs/2309.07601>.
- Liang, Tian, 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.  
<http://arxiv.org/abs/2305.19118>.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. 'Llama 2: Open Foundation and Fine-Tuned Chat Models'. arXiv.  
<http://arxiv.org/abs/2307.09288>.

## 6. Appendix

- **Topics:**

Topics	Stance 1	Stance 2
Climate Change	Pro_Tech	Post_Growth
Gun Control	Pro_Guns	Anti_Guns
Taxation	High_Tax	Low_Tax
Afghanistan war	enforce_stability	Self_Autonomy
Education	Private	public
Technology vs. Security	National_security	Individual_privacy
Speech	Control_speech	Free_speech
Markets	Free_Markets	Control_Markets
Societal Change	Traditions	Progressive
Int. Cooperation	Globalisation	Nationism
Basic Income	Pro_UBI	Anti_UBI

### 1. Climate Change

The first topic is about how we should tackle climate change. If we should rely on technology advancements or if the climate change can only be addressed with a holistic rethinking of our economic system

### 2. Gun Control

This topic accounts for the debate if self defense with guns is a general societal liberty or a threat to the society.

### 3. Taxation

Should inequalities in society be addressed with higher taxation of enterprises and richer individuals or does lower taxation led to a better economic performance contributing everybody.

### 4. Afghanistan war

Do the US have a moral obligation to the stability in Afghanistan or would it be better to leave the country its self-autonomy.

### 5. Education

Would the introduction of competition in the education sector would lead to innovations and improvements in the quality of education or is public education the only way to ensure equality of opportunities.

### 6. Technology vs. Security

How should

### 7. Speech

Would surveillance of free speech foster more equality and less hate or does speech control leads to other risks mitigating creativity and societal advancements through debates.

### 8. Markets

Do free markets lead to innovation and optimal outcomes or do they manifest toward great inequalities in need of interventions and control.

### 9. Societal Change

Should we be open to changes when it comes to our way of living, or should we value traditions and customs we know for long.

### 10. International Cooperation

Should we work closer together when it comes to global challenges or are these problems better solvable when we address them in smaller, cultural closer, contexts.

### 11. Universal Basic income

Is UBI a right approach empowering individuals to unfold their full potential or will the lack of necessity to work disincentivize people.

- **Link to google.docs (all system prompts):**

<https://docs.google.com/document/d/1uuhUdiK1Aq8BW122rwZeehj8jMfWoY2nMH10ZpxLQLU/edit?usp=sharing>

- **Example for judge-assessment for high opinion similarity:**

Distance: 0

Main points of difference:

1. None, as both statements are apologizing for a misunderstanding and acknowledging the structured debate format.

{'role': 'assistant', 'content': "private\_ed : I'm sorry, it seems like there might have been a misunderstanding. The conversation is part of a structured debate where private\_ed presents arguments based on their beliefs and counters arguments made by public\_ed. Let's get back to the debate format so private\_ed can respond to the points raised by public\_ed and further elaborate on their beliefs. Thank you for your understanding."}

{'role': 'assistant', 'content': "public\_ed : Of course, I apologize for the confusion. Let's continue with the structured debate format so that private\_ed can provide their response to the points raised by public\_ed and further elaborate on their beliefs. Thank you for your understanding."}

- **Example for judge-assessment for low opinion similarity:**

Distance: 5

Main points of difference:

1. Individual liberties and right to self-defense
2. Effectiveness of strict gun control measures in ensuring public safety
3. Addressing the root causes of violence in society
4. Focus on responsible gun ownership vs. imposing additional regulations
5. Balance between gun control measures and upholding constitutional rights

```
{'role': 'assistant', 'content': 'pro_guns: \nI appreciate your thoughtful points, anti_guns, but I maintain my position that individual liberties and the right to self-defense should not be compromised in the pursuit of public safety through strict gun control measures. \n\n1. While it is true that the presence of firearms in a household can elevate risks in certain situations, it is also essential to acknowledge that firearms can serve as a deterrent to potential intruders and provide a means of protection for law-abiding citizens facing threats. The ability to defend oneself and loved ones is a fundamental right that should not be diminished based on statistical probabilities alone. \n\n2. Limiting the freedoms of responsible gun owners through additional regulations may not necessarily address the core issues that lead to violence in society. Criminals who seek to harm others will often find ways to access illegal firearms, bypassing legal restrictions. Instead of imposing more barriers on lawful gun owners, efforts should be focused on tackling crime at its roots, addressing the social, economic, and mental health factors that contribute to violent behavior. \n\n3. By prioritizing individual rights and encouraging responsible gun ownership, we can empower citizens to take charge of their own safety within the confines of the law. It is essential to strike a balance between implementing sensible gun control measures and upholding the constitutional rights of citizens, rather than resorting to blanket restrictions that may impede on these foundational freedoms.\n\nIn conclusion, anti_guns, while public safety is paramount, it is crucial to navigate the complexities of gun control in a way that preserves individual liberties and addresses the underlying causes of violence in society effectively.'}
{'role': 'assistant', 'content': 'anti_guns : The debate has concluded. Thank you for your participation!'}
```