An Empirical Analysis on Large Language Models in Debate Evaluation

Anonymous ACL submission

Abstract

In this study, we investigate the capabilities and inherent biases of advanced large language models (LLMs) such as GPT-3.5 and GPT-4 in the context of debate evaluation. We discover that LLM's performance exceeds humans and surpasses the performance of state-of-the-art methods fine-tuned on extensive datasets. We additionally explore and analyze biases present in LLMs, including positional bias, lexical bias, order bias, which may affect their evaluative judgments. Our findings reveal a consistent bias in both GPT-3.5 and GPT-4 towards the second candidate response presented, attributed to prompt design. We also uncover a lexical bias in both GPT-3.5 and GPT-4, especially when label sets carry connotations such as numerical or sequential, highlighting the critical need for careful label verbalizer selection in prompt design. Additionally, our analysis indicates a tendency of both models to favor the debate's concluding side as the winner, suggesting an end-of-discussion bias.

1 Introduction

Prior research in automatic debate evaluation has predominantly relied on pre-trained encoders and the modeling of argument relations and structures (Hsiao et al., 2022; Li et al., 2020; Ruiz-Dolz et al., 2022; Zhang et al., 2023). A significant drawback of these approaches is their dependency on feature engineering and extensive data training, limiting their generalizability across diverse datasets.

The advent of advanced large language models (LLMs) such as GPT-3.5 and GPT-4 (Achiam et al., 2023) has marked the beginning of a new era in automating a wide spectrum of complex tasks (Wei et al., 2022; Thirunavukarasu et al., 2023; Lin* et al., 2023; Wang et al., 2023a; Tang et al., 2023; Zhang et al., 2024; Jiang et al., 2023). These models have been increasingly utilized as automatic evaluators (Chiang and Lee, 2023a,b; Lin and Chen, 2023; Chan et al., 2023; Zeng et al., 2023; He et al., 2023). Leveraging LLMs for debate evaluation presents more challenges, including the extended duration of debates, evolving argument dynamics, and the necessity for evaluators to rely on comprehensive knowledge and reasoning that extend beyond the immediate scope of the debate. Our research delves into the utilization of LLMs for debate evaluation, uncovering their zero-shot capabilities now parallel human evaluators and surpass all existing state-of-the-art (SOTA) methods fine-tuned on ample data (Li et al., 2020; Hsiao et al., 2022).

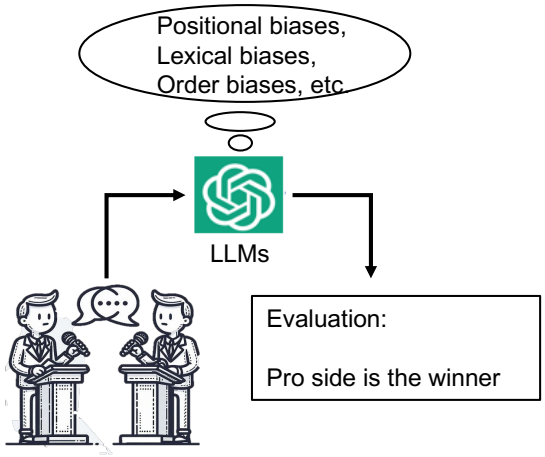


Figure 1: Large Language Models presents various biases during the evaluation of long debates.

We further investigate potential biases in GPT-3.5 and GPT-4 within the context of debate evaluation. While previous research has identified various biases in LLMs, such as persona bias (Wan et al., 2023), political bias (Feng et al., 2023), and positional bias (Wang et al., 2023b), our investigation uniquely concentrates on biases affecting debate evaluation performance, a relatively unexplored domain.

Specifically, upon comparing outcomes between scenarios where the positions of candidate responses are switched, persistent bias has been observed in both GPT-3.5 and GPT-4 toward the sec-

ond candidate response presented, a *positional bias* induced the prompt design. Beyond this, both models also display a significant *lexical bias*, particularly when label sets carry connotations such as sequential or magnitude, underscoring the importance of careful selection of label verbalizers in prompt design to mitigate unintended biases (Liu et al., 2023). Moreover, our study reveals that both GPT-3.5 and GPT-4 exhibit a tendency to favor the concluding side of a debate as the winner, pointing to a potential end-of-discussion *order bias*. These results highlight the nuanced nature of biases in LLMs and the challenge of developing fair, unbiased methodologies for debate evaluation.

2 Methodology

LLMs' capability for debate evaluation. As illustrated in Table 1, we utilize an evaluation template T with two placeholders, $T(\text{Side1_label}, \text{Side2_label})$, to solicit debate assessments from LLMs. For each debate topic, two labels are assigned to represent the Pro and Con sides, respectively. These labels are then inserted into the designated slots of the evaluation template, creating a prompt T, e.g. $T(\text{Side1_label} = \text{P}, \text{Side2_label} = \text{C})$. This tailored prompt is subsequently employed to query the LLM for its judgment on the debate.

LLMs' biases in debate evaluation. Analyzing errors and biases in LLMs is complex due to their multifaceted decision-making processes, which are influenced by inherent orders, stances, and the nuances of prompt design. To separate LLM-induced biases from dataset biases, we use a balanced dataset, ensuring class representation equality and mitigating skewness from overrepresentation. This method focuses on LLM-specific biases within a controlled environment, minimizing dataset composition effects. We explore LLM sensitivity to label choice through various formats (A/B, P/C, 1/-1, Pro/Con), providing a comprehensive bias analysis. Our study also examines the applicability of these findings in real-world, original distribution settings, extending the relevance of our insights.

3 Experiments

Dataset. We utilize DDO dataset (Durmus and Cardie, 2019), which comprises 77,655 debates from 23 topics on debate.org, structured into rounds with a single utterance from each of the Pro and Con side. We focus on debates of 3 to 5 rounds, defining winners by audience vote differences exceeding two, and exclude debates with forfeits to maintain analysis integrity, following previous works' setting (Li et al., 2020; Hsiao et al., 2022).

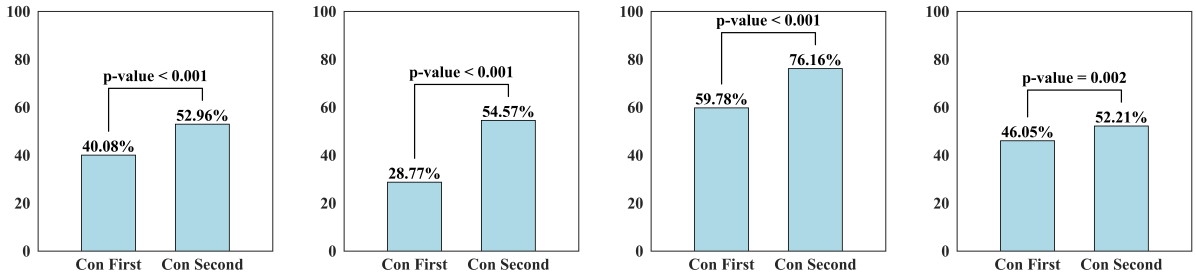
Content
The content of the whole debate:
The current speech in the debate is from the user {Side1_label}:
[The content of the side1]
The current speech in the debate is from the user {Side2_label}:
[The content of the side2]
The current speech in the debate is from the user {Side1_label}:
[The content of the side1]
...
Prompt
Assume you are a debate evaluator, there are two participants in this debate. Given the above context of the whole debate. Please give the decision on which participant is the winner, you only need to give the character(number) of either {Side1_label}, or {Side2_label}. {Side1_label} means user {Side1_label} wins. {Side2_label} means user {Side2_label} wins. Please only give the result without any other words.

Table 1: Our evaluation template for determining debate outcomes for Large Language Models. We provide a specific 3-round debate example in Appendix A.5

The dataset exhibits a win bias towards the Con side across 3 to 5-round debates (36.89% vs. 63.11%, 44.9% vs. 55.1%, 37.9% vs. 62.1%, respectively), likely due to a concluding side bias with Con frequently concluding debates. To evaluate LLMs in debate assessment, we propose two settings: balanced and unbalanced. The unbalanced setting replicates the original dataset's distribution, sampling 500 debates for each round count (totaling 1500). Conversely, the balanced setting aims to examine LLMs' inherent bias by ensuring equal representation of four scenarios—Pro or Con initiating and winning or losing—with 125 debates each for 3 and 4 rounds, and due to data constraints, 75 debates each for 5 rounds, resulting in 500 debates for 3 and 4 rounds and 300 for 5 rounds.

Evaluation metrics. In addition to the accuracy reported by previous works, we additionally measure weighted F-1 score to accommodate the imbalance between Pro win and Con win in the original data distribution.

Models. We utilize OpenAI's GPT-3.5 and GPT-4 models to conduct our experiments, namely gpt-3.5-turbo-1106 and gpt-4-1106-preview.



(a) A/B label set (b) P/C label set (c) 1/-1 label set (d) Pro/Con label set

Figure 2: The observed positional bias in GPT-3.5 is evident through the alteration in the proportion of Predicted Con outcomes, which increases when Con is positioned as the second candidate response compared to its placement as the first. This consistent preference across all label configurations suggests a systematic positional bias favoring the second candidate, underscoring the model's sensitivity to the order in which options are presented.

Human annotation. To assess the effectiveness of LLMs, two authors manually annotated the "win/lose" outcomes of randomly selected debates independently for 75 debates. Unlike the collective voting in multi-audience settings, this annotation was independently completed by a single annotator.

4 Results and Analysis

4.1 LLMs' Performance

Table 2 reveals that GPT-3.5 and GPT-4 match human evaluators in assessing debates, highlighting their effectiveness. Using 75 debates labeled by two of the authors enables a direct comparison with GPT-3.5 and GPT-4. They achieve significant accuracy and F1 scores—82.04% and 81.85% for GPT-3.5, and 86.22% and 86.01% for GPT-4, respectively, outperforming previous SOTA models.

The choice of words in prompts has a profound impact on LLM performance (Table 3). For GPT-3.5, using 1/-1 labels decreases performance, while Pro/Con labels lead to GPT-4's lowest results. GPT-3.5 shows a notable drop when asked to identify the loser, indicative of its sensitivity to negative terms, unlike GPT-4, which exhibits much less performance drop under the same conditions.

4.2 Biases Analysis

Our study explores biases present in GPT-3.5 using a balanced dataset. Additional analyses of the GPT-3.5 on the original unbalanced data and analysis of GPT-4 are in Appendix A.3 and A.2, respectively.

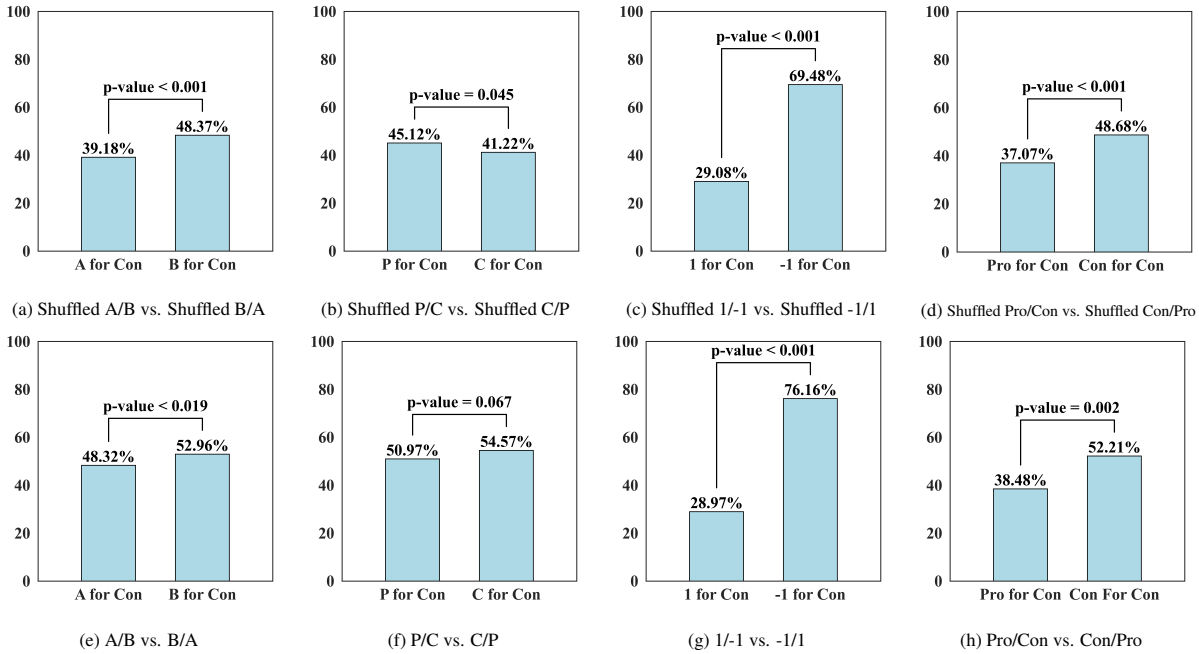
Positional Bias. Fig. 2 illustrates the disparity in "Con" predictions between scenarios where the corresponding Con_label is presented as the first candidate response and where is positioned as the second. It reveals GPT-3.5's systematic preference for responses in the second position across various verbalizer configurations. The significance of this positional bias is corroborated by two-proportion z-test P-values, aligning with Wang et al. (2023b)'s findings regarding GPT-3.5's second-position bias. More discussion can be found in Appendix A.2.

Evaluators	Size	Acc	F1
Rule-based	6058	67.53	46.68
BERT + Structure	-	78.89	–
BERT + Relation	1964	80.04	–
GPT-3.5	1500	82.04	81.85
GPT-4	1500	86.22	86.01
Human 1	75	77.33	77.39
Human 2	75	78.67	78.15

Table 2: GPT-3.5 and GPT-4's performance are on par with human performance and outperform the existing state-of-the-art BERT-based methods with fine-tuning. The rule-based model predicts the winner as the side that concludes the debate.

Lexical Bias. GPT-3.5 demonstrates a lexical bias, favoring the label 'B'('-1') over 'A'('1'), as it predicts 'Con' as the winner more often when 'B'('-1') represents 'Con' illustrated in Fig. 3. There is no significant lexical bias found within the P/C label set. The Con/Pro configuration might lead to confusion in LLMs due to the content of debates clarifying the actual stance of each side, potentially explaining the observed decrease in performance with this label setup. The analysis of lexical bias is further detailed in Appendix A.2.

Order Bias. GPT-3.5 exhibits a significant order bias, favoring the side that concludes the debate, as shown in experiments where the Pro_label consistently ranked as the primary response (Table 4). This bias is statistically significant across all ver-



(a) Shuffled A/B vs. Shuffled B/A (b) Shuffled P/C vs. Shuffled C/P (c) Shuffled 1/-1 vs. Shuffled -1/1 (d) Shuffled Pro/Con vs. Shuffled Con/Pro

(e) A/B vs. B/A (f) P/C vs. C/P (g) 1/-1 vs. -1/1 (h) Pro/Con vs. Con/Pro

Figure 3: GPT-3.5 has a performance fluctuation in Predicted Con proportions upon altering Pro and Con verbalizers. Each subfigure's legend delineates the Pro/Con label set across different verbalizer configurations. GPT-3.5 demonstrates a consistent lexical bias, which persists across shuffled positions aimed at counteracting positional bias, and in settings where Pro consistently precedes Con, except for the insignificant bias within P/C.

Evaluators	Verbalizer	Outcome	Acc	F1
GPT-3.5	A/B	Winner	**82.04**	**81.85**
	P/C	Winner	81.39	81.02
	1/-1	Winner	*72.08*	*68.24*
	Pro/Con	Winner	81.86	81.60
	A/B	Loser	37.72	24.74
GPT-4	A/B	Winner	84.49	84.49
	P/C	Winner	85.11	84.78
	1/-1	Winner	**86.22**	**86.01**
	Pro/Con	Winner	*79.72*	*78.16*
	A/B	Loser	80.94	81.11

Table 3: The "Verbalizer" column lists Pro_label and Con_label sets, and the "Outcome" column shows whether GPT-3.5 and GPT-4 are tasked with identifying debate winners or losers. Bold formatting indicates the top-performing verbalizer choice, while italics highlight the least effective choice.

Verbalizer	End-Side	# P-Pro	# P-Con	P-Value
A/B	Pro	**389**	253	< 0.001
	Con	215	**427**	
P/C	Pro	**408**	245	< 0.001
	Con	184	**460**	
1/-1	Pro	238	**409***	< 0.001
	Con	70	**575**	
Pro/Con	Pro	**399**	218	< 0.001
	Con	248	**426**	

Table 4: Analysis of GPT-3.5 predictions correlating with debate outcomes, using Chi-square tests for significance. "# P-Pro" and "# P-Con" indicate the counts of Pro and Con sides predicted as winners, respectively. The results reveal a significant association with order for all verbalizer choices. * here highlights the strong lexical bias for '-1' that dominates the others.

balizer options. The results suggest an inherent tendency in LLMs to give more weight to the final arguments.

5 Discussion

Our research demonstrates that LLMs outperform current SOTA models in evaluating debates but are influenced by specific word choices, affecting their efficacy. We highlight LLMs' embedded biases—positional, lexical, and order—offering insights for future LLM training enhancements.

Despite attempts to neutralize positional bias by shuffling labels in Figs 3a and 3d, GPT-3.5 still exhibits a Pro bias, contradicting its lexical preference for 'B'('Con'). This might suggest a confirmation bias-like tendency in GPT-3.5, favoring agreement with the debate topic. We further conduct experiments shuffling A/B with B/A and 1/-1 with -1/1 label sets, where each label randomly represents Pro or Con in 50% of cases, with positions also shuffled. Despite eliminating lexical and positional biases, results indicate a persistent Pro bias, detailed in Appendix Fig. 6., pointing to an underlying tendency warranting further investigation.

6 Limitations

The insights from our investigation, based on the examination of GPT-3.5 and GPT-4, indicate that the discerned behavioral patterns might be unique to these specific models and not necessarily extend to other language models with divergent architectures or training approaches. With the relentless advancement in language model technology and the anticipation of updated versions, the biases detected in GPT-3.5 and GPT-4 could become obsolete in subsequent iterations. Highlighting the significance of prompt types and training techniques on the efficacy of models, our research underlines the imperative for continued research to identify the optimal prompt types for various scenarios and the optimal training methods for reducing bias.

Although various biases may interact and potentially counterbalance each other, leading to improvements, the intensity of distinct bias types can vary significantly across different contexts. Consequently, a prompt that appears to exhibit balanced bias in one scenario may manifest more pronounced bias under slightly altered conditions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. *arXiv preprint arXiv:1906.11310*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Socreval: Large language models with the socratic method for reference-free reasoning evaluation.

Fa-Hsuan Hsiao, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Modeling inter round attack of online debaters for winner prediction. In *Proceedings of the ACM Web Conference 2022*, pages 2860–2869.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. *arXiv preprint arXiv:2010.03538*.

Jingyang Lin*, Hang Hua*, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, , and Jiebo Luo. 2023. Videoxum: Cross-modal visual and textural summarization of videos.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Ramon Ruiz-Dolz, Stella Heras, and Ana García-Fornes. 2022. Automatic debate evaluation with argumentation semantics and natural language argument graph networks. *arXiv preprint arXiv:2203.14647*.

Yunlong Tang, Jinrui Zhang, Xiangchen Wang, Teng Wang, and Feng Zheng. 2023. Llmva-gebc: Large language model with video adapter for generic event boundary captioning. *arXiv preprint arXiv:2306.10354*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

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. *arXiv preprint arXiv:2305.17926*.

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

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.

Gechuan Zhang, Paul Nulty, and David Lillis. 2023. Argument mining with graph representation learning. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 371–380.

A Appendix

A.1 More details of the Dataset

The dataset extends beyond textual debate content to audience votes across four evaluation criteria: making more convincing arguments, better conduct, use of reliable sources, and spelling and grammar proficiency. Consistent with prior research (Li et al., 2020; Hsiao et al., 2022), our analysis utilizes the criterion of "making more convincing arguments" for assessing debate outcomes. To ensure alignment with these studies and enhance comparability, we narrow our focus to debates with a definitive margin of victory—requiring a vote difference exceeding two—and limit our analysis to debates spanning three to five rounds, which represent the bulk of the dataset. Debates compromised by forfeits, identified either through explicit forfeit labels or instances of one side forfeiting a round, are omitted from consideration.

The debates within the dataset have an average length of 1574.93 words, with the majority fit within the input length constraints of contemporary LLMs. Regarding audience engagement, the average vote counts for 3-round, 4-round, and 5-round debates stand at 10.05, 7.02, and 7.03, respectively. Furthermore, the average vote differences for these debate formats are 5.52, 4.69, and 4.79, indicating a clear preference in outcomes that facilitate our focused analysis on convincing arguments.

A.2 Additional Results

The detailed confusion matrices with various settings we experiment on balanced datasets can be found in Fig 7 for GPT-3.5 and in Fig 9 for GPT-4.

Positional Bias. For GPT-3.5, McNemar's tests (McNemar, 1947) are also conducted for the settings with shuffled candidate response positions and fixed positions based on Table 5, and the results are all significant.

Verbalizers	$f_{\text{fixed_shuffled}}$	$f_{\text{shuffled_fixed}}$	χ^2	P-Value
A/B	25	86	33.52	< 0.001
P/C	16	205	161.63	< 0.001
1/-1	38	124	45.65	< 0.001
Pro/Con	34	79	17.92	< 0.001

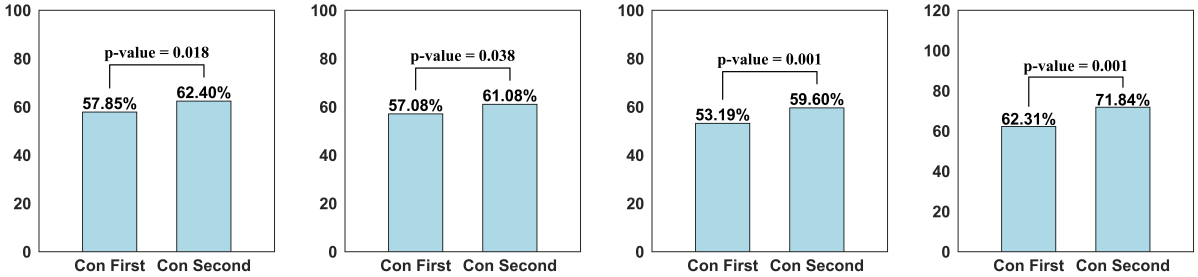
Table 5: McNemar's test demonstrates that all positional biases are significant within GPT-3.5. $f_{\text{fixed_shuffled}}$ indicates the number of debates predicted as Pro winning by the first verbalizer set but Con winning by the second verbalizer set. $f_{\text{shuffled_fixed}}$ indicates the number of debates predicted as Pro winning with shuffled positions but Con winning by GPT-3.5 with fixing Pro as the first candidate response.

The direction of the positional bias presented by GPT-4 is also shown towards the second position, contradicting the finding of the first position favorite illustrated by Wang et al. (2023b). The two-sided p-value from the two-portion z-tests demonstrates that the positional bias in GPT-4 as shown in Fig 4 is also statistically significant.

Lexical Bias. The difference in the significance of lexical bias within the A/B label set and P/C label set could be due to the alphabetical distance they have or due to their common usage. To discern the underlying cause, we further experiment with the M/N label set for they are alphabetically adjacent but not typically associated with sequential interpretation. The results, detailed in Fig 8, reveal minimal lexical bias within the M/N group, suggesting that the bias originates from conventional usage rather than alphabetic proximity.

To further quantitatively assess the lexical bias in GPT-3.5, we employ McNemar's test to analyze instances of concordances (both predict Pro or Con), instances of discordances (one predicts Pro and the other predict Con) of each flipping group as shown in Table 6. All results are statistically significant.

Similar to GPT-3.5, GPT-4 also exhibits lexical bias towards 'B', '-1' and potentially 'Con' within the A/B, 1/-1, and Pro/Con label set. However, GPT-4 favors 'C' over 'P' significantly. McNemar's tests of lexical bias for GPT-4 are shown in Table 8 for GPT-4.



(a) A/B label set (b) P/C label set (c) 1/-1 label set (d) Pro/Con label set

Figure 4: This figure illustrates the impact of positional bias on GPT-4 through the changes in the proportion of Predicted Con, shifting from when Con is fixed as the first candidate response to when it is positioned as the second. GPT-4 exhibits a positional bias towards the second candidate presented across all label set configurations.

Verbalizers	f_{12}	f_{21}	χ^2	P-Value
A/B vs B/A	59	178	59.751	< 0.001
P/C vs C/P	166	99	16.94	< 0.001
1/-1 vs -1/1	33	556	464.40	< 0.001
Pro/Con vs Con/Pro	147	298	51.24	< 0.001

Table 6: McNemar's test demonstrates that all lexical biases are significant within GPT-3.5. f_{12} indicates the number of debates predicted as Pro winning by the first verbalizer set but Con winning by the second verbalizer set. f_{21} indicates the number of debates predicted as Pro winning by the second verbalizer set but Con winning by the first verbalizer set. The positions of verbalizers in the prompt are shuffled.

Order Bias. The Chi-squared test to show the association between the GPT-4's predictions and the sides that conclude the debates are shown in Table 9. Same as GPT-3.5, across all verbalizer choices, the order biases presented by GPT-4 are also statistically significant. In addition, the magnitude of the order bias within GPT-3.5 is much stronger than GPT-4, as measured by the Phi Coefficient.

Stance Bias. Our hypothesis regarding stance bias is less evident in GPT-4, as it becomes overshadowed by lexical bias after positional bias is mitigated through shuffled positions. We conduct two experiments, employing shuffled label sets and positions under the A/B and 1/-1 configurations, as depicted in Fig 6. The findings reveal a contrasting residual bias in GPT-4 compared to GPT-3.5, after addressing positional, lexical, and order biases.

A.3 Extension to Unbalanced Dataset

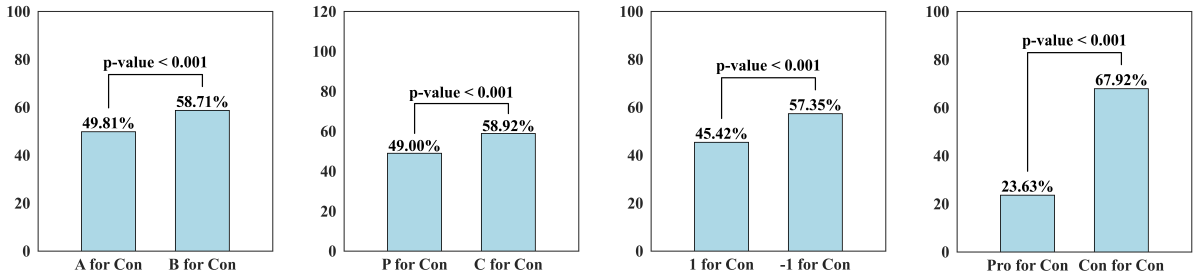
Positional Bias We additionally explore variations in "Pro" and "Con" predictions when alternating between shuffled and fixed candidate response placements in unbalanced data that reflects the original distribution. These observations, detailed in Table 10, highlight a consistent pattern.

Lexical Bias The same experiments applied to the unbalanced datset with the original distribution yield consistent results for the direction of lexical bias in GPT-3.5 (see Table 11), except for the non-significance P/C set.

A.4 Enhancing Bias Reduction through Prompt Engineering

LLM-Eval In a further step, we direct the LLMs to provide reasons for their judgments before they generate the outcomes using the prompt template shown in Table 12. Such a method is reported by Wang et al. (2023b) to be able to reduce the positional bias. We do a pilot experiment using GPT-3.5 with a single A/B label set to see if the effect comes from 'reducing' the bias or from providing a bias in the opposite direction and thus counteract it.

As the results shown in Fig. 10, GPT-3.5 exhibits a greater bias towards Pro when generating analysis compared to when positions are shuffled to eliminate the positional bias. Therefore, it is more likely that prompting GPT-3.5 to generate the analysis first introduces a new bias towards Pro, which is in the



(a) Shuffled A/B vs. Shuffled B/A (b) Shuffled P/C vs. Shuffled C/P (c) Shuffled 1/-1 vs. Shuffled -1/1 (d) Shuffled Pro/Con vs. Shuffled Con/Pro

Figure 5: This figure illustrates the impact of lexical bias on GPT-4 through the changes in the portion of Predicted Con from switching the verbalizers for Pro and Con.

Verbalizers	$f_{\text{fixed_shuffled}}$	$f_{\text{shuffled_fixed}}$	χ^2	**P-Value**
A/B	6	54	36.82	< 0.001
P/C	9	39	17.52	< 0.001
1/-1	15	45	14.02	0.002
Pro/Con	11	63	35.15	< 0.001

Table 7: McNemar's test demonstrates that all positional biases are significant within GPT-4. $f_{\text{fixed_shuffled}}$ indicates the number of debates predicted as Pro winning by the first verbalizer set but Con winning by the second verbalizer set. $f_{\text{shuffled_fixed}}$ indicates the number of debates predicted as Pro winning with shuffled positions but Con winning by GPT-3.5 with fixing Pro as the first candidate response.

opposite direction of the positional bias, since Con is consistently positioned as the second candidate response. However, arriving at a definitive answer necessitates further experimentation, which we defer to future research.

A.5 Debate Example

The debate example can be found in Table 13, 14 and 15.

Verbalizers	f_{12}	f_{21}	χ^2	P-Value
A/B vs B/A	4	36	24.03	< 0.001
P/C vs C/P	6	134	115.21	< 0.001
1/-1 vs -1/1	11	166	133.99	< 0.001
Pro/Con vs Con/Pro	6	581	561.29	< 0.001

Table 8: McNemar's test demonstrates that all lexical biases are significant within GPT-4. f_{12} indicates the number of debates predicted as Pro winning by the first verbalizer set but Con winning by the second verbalizer set. f_{21} indicates the number of debates predicted as Pro winning by the second verbalizer set but Con winning by the first verbalizer set. The positions of verbalizers in the prompt are shuffled.

Verbalizer	End-Side	# P-Pro	# P-Con	Phi Coeff.	P-Value
A/B	Pro	359	291	0.099	< 0.001
	Con	293	356		
P/C	Pro	277	372	0.076	0.006
	Con	227	419		
1/-1	Pro	286	362	0.074	0.007
	Con	238	411		
1/-1	Pro	203	445	0.069	0.001
	Con	162	486		

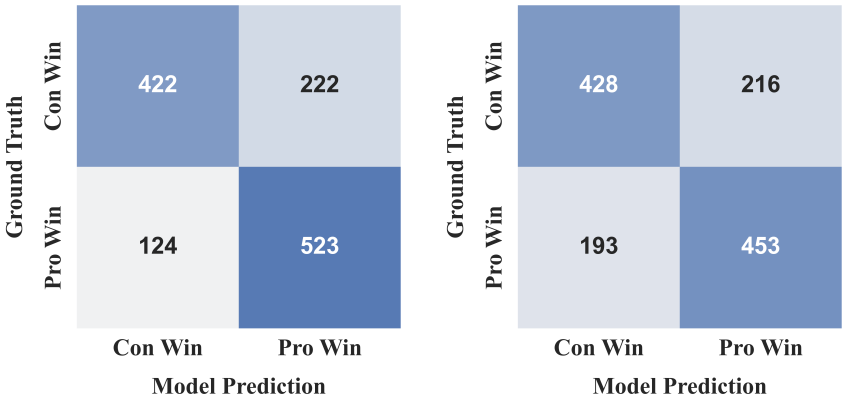
Table 9: GPT-4 predictions and debate conclusions association analysis with significance determined by Chi-square tests. # P-Pro and # P-Con denote the number of predicted Pro sides and Con sides as the winner by the model, respectively.

Verbalizer	Position	# Pred Pro	# Pred Con
A/B	Fixed	533	954
	Shuffled	610	881
P/C	Fixed	494	1000
	Shuffled	727	769
1/-1	Fixed	230	1267
	Shuffled	340	1155
Pro/Con	Fixed	517	977
	Shuffled	590	904

Table 10: Upon fixing and shuffling the positions of labels set as candidate responses in an unbalanced dataset that replicates the original data distribution, the analysis systematically reveals a positional bias towards the second position in GPT-3.5.

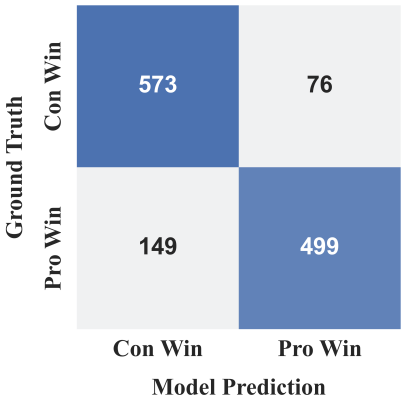
Verbalizer	# P-Pro	# P-Con
A/B	610	882
B/A	755	734
P/C	727	769
C/P	717	835
1/-1	340	1155
-1/1	943	553
Pro/Con	590	904
Con/Pro	877	619

Table 11: Upon flipping label sets and shuffling their positions in an unbalanced dataset which replicates the original data distribution, the analysis systematically reveals lexical biases in GPT-3.5 that align directionally with those identified in a balanced dataset. # P-Pro and # P-Con denote the number of predicted Pro sides and Con sides as the winner by the model, respectively.

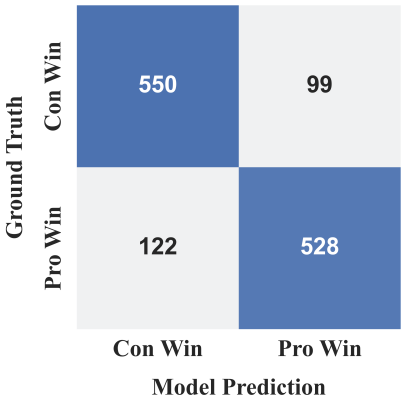


(a) Double shuffled A/B in GPT-3.5

(b) Double shuffled 1/-1 in GPT-3.5

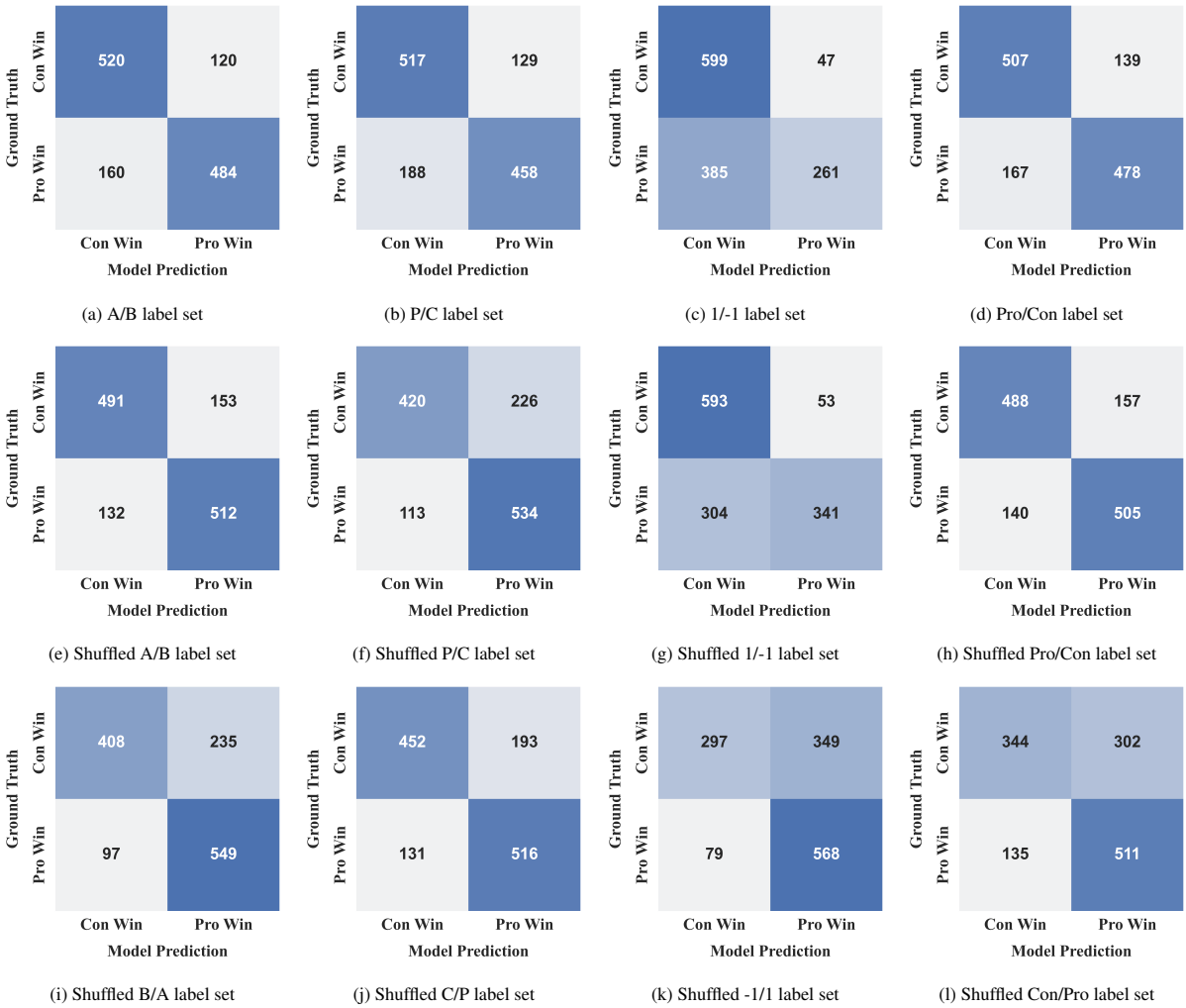


(c) Double shuffled A/B in GPT-4



(d) Double shuffled 1/-1 in GPT-4

Figure 6: Both the assignment of labels within each label set and the positions of labels are shuffled. These matrcies demonstrate that after eliminating the influence of the order bias, positional bias, and lexical bias, GPT-3.5 shows a stance bias towards the Pro stance, while GPT-4 shows a stance bias towards the Con stance.

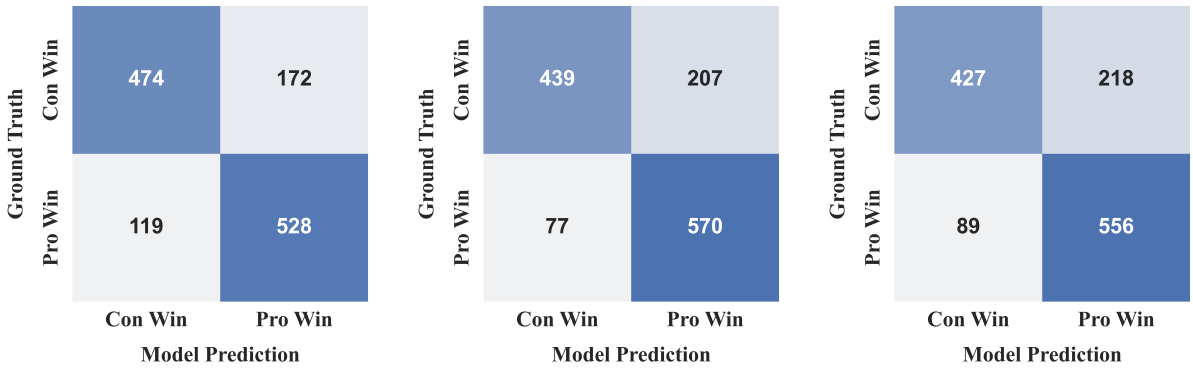


(a) A/B label set (b) P/C label set (c) 1/-1 label set (d) Pro/Con label set

(e) Shuffled A/B label set (f) Shuffled P/C label set (g) Shuffled 1/-1 label set (h) Shuffled Pro/Con label set

(i) Shuffled B/A label set (j) Shuffled C/P label set (k) Shuffled -1/1 label set (l) Shuffled Con/Pro label set

Figure 7: This figure displays confusion matrices for GPT-3.5 with various Pro_label/Con_label sets. The matrices in the first row correspond to scenarios where the Pro_label consistently occupies the leading position in the instruction prompt, potentially introducing a positional bias. In contrast, the second and third rows present matrices from experiments where the positions of Pro_label and Con_label are shuffled, aiming to mitigate this bias for pure comparisons between switching corresponding label verbalizers of Pro and Con.



(a) M/N label set (b) Shuffled M/N label set (c) Shuffled N/M label set

Figure 8: Lexical bias of M/N label set in GPT-3.5

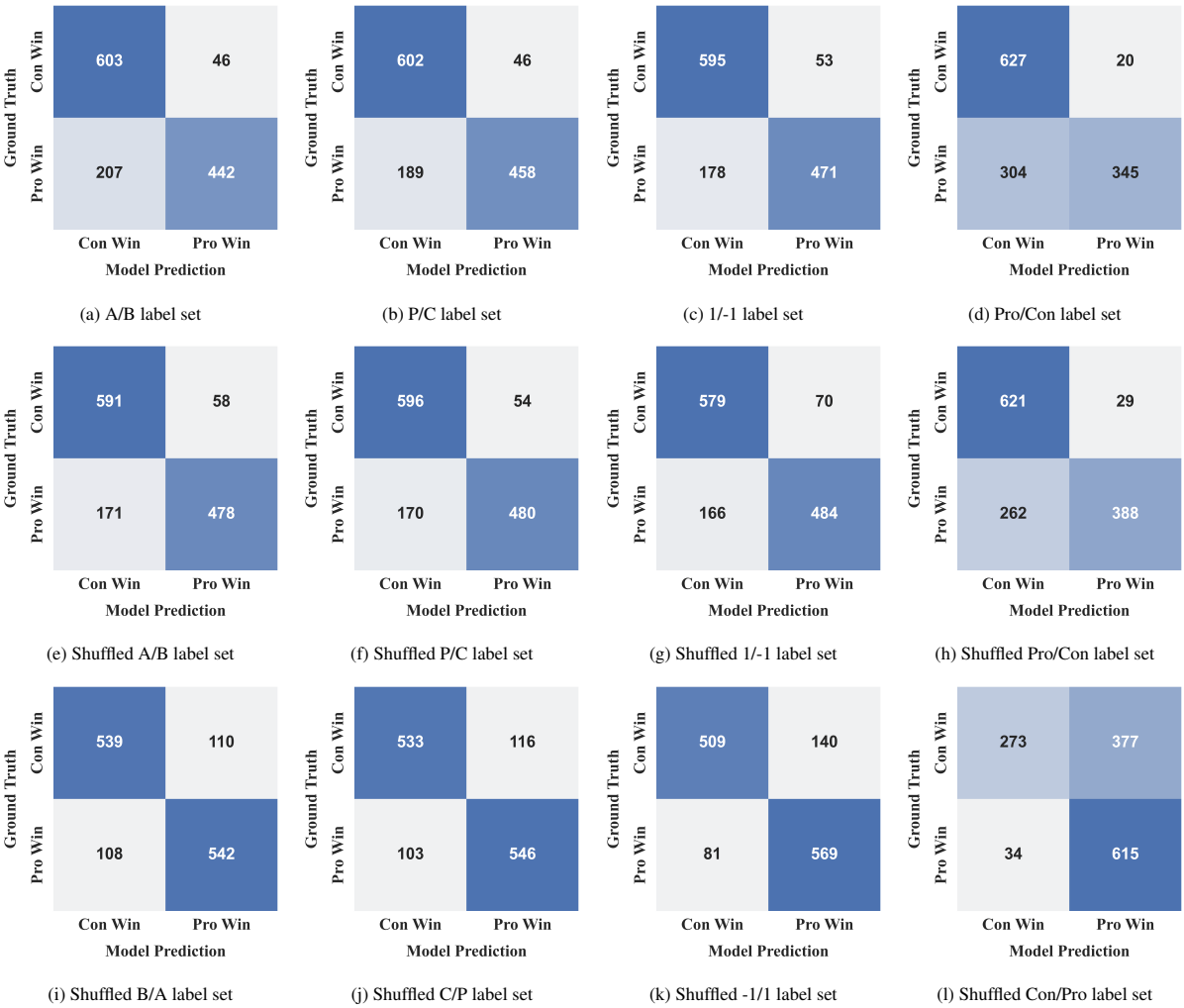


Figure 9: This figure illustrates the impact of lexical bias on GPT-4 through confusion matrices for various Pro_label/Con_label sets. The matrices in the first row correspond to scenarios where the Pro_label consistently occupies the leading position in the instruction prompt, potentially introducing a positional bias. In contrast, the second and third rows present matrices from experiments where the positions of Pro_label and Con_label are shuffled, aiming to mitigate this bias.

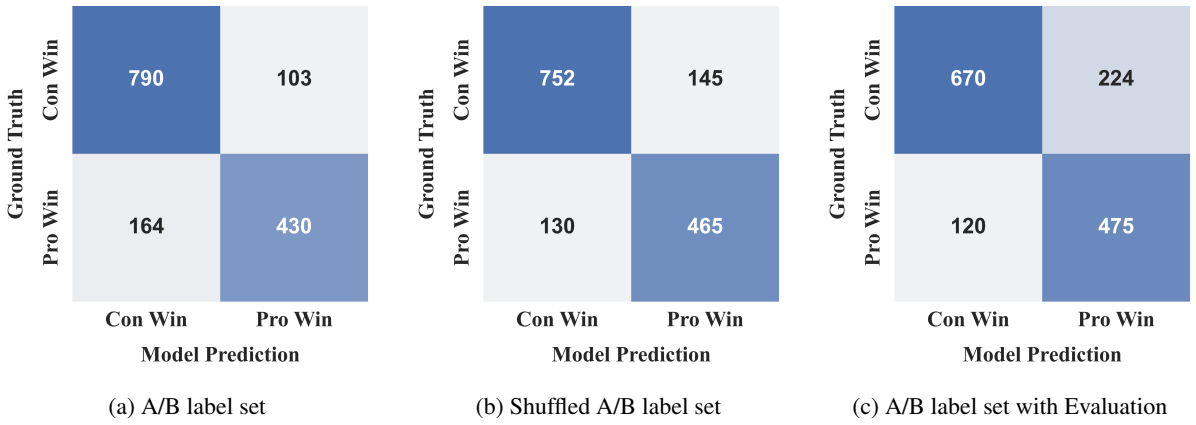


Figure 10: Analysis of the effect of generating analysis on reducing positional bias.

Content Prompt
The content of the whole debate:
The current speech in the debate is from the user {Side1_label}:
[The content of the side1]
The current speech in the debate is from the user {Side2_label}:
[The content of the side2]
The current speech in the debate is from the user {Side1_label}:
[The content of the side1]
...
Vanilla Prompt
Assume you are a debate evaluator, there are two participants in this debate. Given the above context of the whole debate. Please give the decision on which participant is the winner, you only need to give the character(number) of either {Side1_label}, or {Side2_label}. {Side1_label} means user {Side1_label} wins. {Side2_label} means user {Side2_label} wins. Please only give the result without any other words.
Eval Prompt
Assume you are a debate evaluator, there are two participants in this debate. Given the above context of the whole debate, 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. Finally, decide who wins the debate.
Output with the following format:
Evaluation:
<your comprehensive evaluation explanation here>
<winner ({Side1_label} or {Side2_label})>
The final line of your output should contain only one word: {Side1_label} if you conclude that user {Side1_label} wins, or {Side2_label} if you conclude that user {Side2_label} wins. No tie or inconclusive results are allowed.

Table 12: The "Vanilla Evaluation" prompts the model to predict results directly based on the content prompt. The "Eval Prompt" mandates the model to evaluate arguments for both sides and provide a holistic assessment based on the “Content prompt”.

The current speech in the debate is from the user {Side1_label}:
Thank you, to whoever accepts this challenge, I look forward to this debate.
Now, to start off, I will go over some definitions.
Morality: [conformity to the rules of right conduct] Evil: [morally wrong or bad; immoral; wicked] Right: [in accordance with what is good, proper, or just] Atheist: [a person who denies or disbelieves the existence of a supreme being or beings.] Theist: [the belief in one God (in this debate I am referring to the Christian God) as the creator and ruler of the universe, without rejection of revelation]
http://dictionary.reference.com...
To begin with, what makes something wrong or right? The law of a specific nation? Yourself? This question was simple when we were kids, for instance if John hit Sue then John was wrong and then gets in trouble. But as we get older this topic becomes more complicated. For instance, who said it was wrong for John to hit Sue? Who said it was wrong for someone to steal, cheat, lie, murder, torture, rape? The point is, in some cultures its acceptable and even encouraged to do these things. Just look at Hitler, Stalin or any other evil dictator/government. Anyone can read about the trattorias acts that have been accurately recorded through out history. But here's the thing, all these men committed terrible acts without believing that they themselves were 'wrong.' For example, Hitler murdered 10 million people for ethnic cleansing reasons, and through out his entire life as ruler over Germany, never once thought he was doing an immoral act. In fact, he believed he was doing just the opposite, Hitler thought, that through killing 10 million people he was "glorifying the Father Land" and doing the world a huge favor. Plus, Hitler not only was evil himself, but he had a whole nation behind him. Millions swore true allegiance to him, and his ideas.
Now, given the above paragraph, it is impossible to say that Hitler's actions were immoral under an Atheistic world view. Why? Because in an Atheistic world view there is no God to judge such acts. The only thing that can judge Hitler in an Atheistic world are other people, but what if every single person on the planet became a Nazi. So there must be an ultimate judge, or over seer, in order for Hitler's actions to be held accountable.
So, if one wants to debate that morality is defined by the law of a specific nation, or ones ability to justify there own actions, then the voters and my oppenent should be able to see clearly that an Atheistic world view can not account for morality.
Please answer the following questions in your next argument.
How can Atheism account for morality? And what will you base you morality off of, if not God?
The current speech in the debate is from the user {Side2_label}:
This should be an interesting debate.... I love this sort of debate, ie. what are morals and why do we have them sort of thing
"Because in an Atheistic world view there is no God to judge such acts. The only thing that can judge Hitler in an Atheistic world are other people, but what if every single person on the planet became a Nazi."
First If everyone was a Nazi there would be no problem with Nazism because they wouldn't have a WW2 Repeat due to the fact that everyone would agree....
Second the people/self being the judge is what I am arguing. The Ultimate judge is humanity. The concept of the Other best applies here. When we look at another acting, we judge them. When we look at ourselves acting the same way, we remember that judgment. We don't have to actually see someone else, but imagine that there is that Other judging us.
Also if we look to the roots of morality we don't find God, but humanity. Why is it immoral to kill? Because if it was allowed then people would freely kill us. If we look at what we would think had we seen the event happen, or been the recipient, we will agree that the event is bad. From All this we can take morality to really be a golden rule of sorts. Judge ourselves as we would judge others. Do to others as we would have done to ourselves. Neither of these concepts require God, in fact they function just as well with a God as without.
"How can Atheism account for morality? And what will you base you morality off of, if not God?" I have already sort of answered this but I will do it again for sake of order and clarity. Atheism accounts for morality via Humanity. The roots of our morals exist in an atheist society, they were created not by God but by human conscience and need for order and safety. I don't want to retype the explanation of the Other(which was admittedly pretty bad) but that is a general concept of how atheism can account for and provide a base for morality. The golden rule is another base for morality. Morals Exist for human safety primarily. Why is it immoral to kill? because we don't want to be killed.
God is not the source of Morals, and therefore an atheistic world view can account for morals just as well as a theistic world view can.

Table 13: The first round of a debate example.

The current speech in the debate is from the user {Side1_label}:
Thanks for your response.
Metz said, "Second the people/self being the judge is what I am arguing. The Ultimate judge is humanity."
To say that humanity is the ultimate judge is not saying anything. For instance, in one part of the world it may be morally acceptable to murder your wife if she disobeys her husband. In another part of the world that particular act may be unacceptable. But, which view of the issue is right? Who decides it? The point is, that to base what is considered right or wrong off humanity is ridiculous, since humanity can not agree on an absolute, universal view of what is considered moral or immoral. Since this is true anything could be acceptable, such as murder, rape, lying cheating, abusing, drugs ect... Why? Again, because morality is totally arbitrary under the jurisdiction of humanity, since all humans have different standards of morals. And since all humans have different standards on morals, then this just illustrates my point, there must be a God to judge people's actions. In an Atheistic world there are no absolutes for morals.
Also, if there are seven hundred billion people on the planet and half say gay marriage is right but the other half say gay marriage is wrong, then who decides? What makes one view right and the other wrong? This question can not be answered in an Atheistic universe, since all the opinions given by the people are different. So, humanity, can not, on it's own make a rational decision, dealing with morality. This is why there must be an objective standard for people to base their judgement off of. Again, under an Atheistic world view morals can not be accounted for.
"Why is it immoral to kill? Because if it was allowed then people would freely kill us."
What about the people who could care less about whether or not death is a reaction of killing another person. For instance, a man could be very enraged at a particular moment, so, what if he decides to kill everyone in the town regardless of wether he dies that day or lives, in the process of committing all the murders he can. Not only that flaw, but there are people who murder people all the time without getting caught, or getting killed back in the process. So for these murderers there is no incentive what so ever for them to not go out and murder another human being.
Plus, saying that its immoral to murder because you will get murdered back is not even answering the question of why it is immoral to murder another human being. You need to tell me why murder is wrong in the first place.
Metz said "Do to others as we would have done to ourselves."
Its amazing how Atheists think, they will always claim there world has morals, and do things such as feed the poor and help many in need ect... These are all good things, its just the principles in which these acts are found, are in the Bible. You see, Atheists take morals from the Christian world view but do not acknowledge the basis of which those morals came from, which is ultimately God. Now I'm not saying that all of the morals in an Atheistic world view are taken from Christianity, but a lot of them are, Along with many other religions that acknowledge the presence of a god.
Metz said, "Also if we look to the roots of morality we don't find God, but humanity."
Prove to me that we find humanity, don't just say it, prove it or at least tell expand on that reasoning. I do not agree with that statement at all and until you try to prove it it is just your word against mine. Which is exactly what an atheistic world view consists of, one man's word against another, which is no absolutes or universal ideas
I also encourage the voters to check out this link, it will help illustrate my point.
Thank you charles 15
Good Luck
The current speech in the debate is from the user {Side2_label}: "To say that humanity is the ultimate judge is not saying anything. For instance, in one part of the world it may be morally acceptable to murder your wife if she disobeys her husband. In another part of the world that particular act may be unacceptable. But, which view of the issue is right? Who decides it? The point is, that to base what is considered right or wrong off humanity is ridiculous, since humanity can not agree on an absolute, universal view of what is considered moral or immoral."
But this accounts for morality... it just doesn't account for my opponents version of morality. Also this really doesn't say why Theism can actually account for universal morals. People disagree on religion. If Morals were universal then the scenario my opponent laid out wouldn't exist. But yet he claimed it does.... So what my opponent is essentially arguing is that Morality doesn't work.
"In an Atheistic world there are no absolutes for morals."
Ok... Same thing in a Theist world. But lets look at the topic for a moment shall we? It never says Atheism needs to account for universal morals, just morals. This really doesn't attack my case at all. The Definition of Morality my opponent gives is "conformity to the rules of right conduct" But it never says these rules must be universal. If we have laws they do not hold everyone accountable worldwide, likewise morality doesn't have to be universal.
"Again, because morality is totally arbitrary under the jurisdiction of humanity, since all humans have different standards of morals."
That is how I argue we can account for morality. If we want to find acceptable morality we need people to disagree, this is how democracy works and how morality would inevitable work. And yet again, Theism is different how?
"Also, if there are seven hundred billion people on the planet and half say gay marriage is right but the other half say gay marriage is wrong, then who decides? What makes one view right and the other wrong? This question can not be answered in an Atheistic universe, since all the opinions given by the people are different. So, humanity, can not, on it's own make a rational decision, dealing with morality."
Oh yeah... and God is doing so much better? The reason so many people disagree is primarily religion...granted there are other factors but religion and tradition are massive players.
"So, humanity, can not, on it's own make a rational decision, dealing with morality"
Well actually we live in a largely theist world... so what you meant to say was " So, God and religion cannot make a rational decision dealing with morality"
"This is why there must be an objective standard for people to base their judgement off of"
yeah, its called survival mate.... people see other and judge themselves... People tell others that a certain action is wrong because they don't want what they see done to other done to themselves...
"You need to tell me why murder is wrong in the first place."
Its wrong because people say its wrong... you essentially made my argument for me there; "its immoral to murder because you will get murdered back" it isn't moral to Murder because you are ending that persons existence. I don't want to end my existence so I tell people that it is wrong to kill. If I wanted to be killed would I say it is wrong to kill?
"You see, Atheists take morals from the Christian world view but do not acknowledge the basis of which those morals came from, which is ultimately God"
Um... Alright... The First appearance of the golden rule was I believe in the Analects of Confucious... Not the bible. Also it really doesn't matter where the Morals came from as long as an Atheist world can account for them... I personally have a justification for all my moral opinions that has nothing to do with god but with how I perceive humans.
"Metz said, "Also if we look to the roots of morality we don't find God, but humanity." Prove to me that we find humanity, don't just say it, prove it or at least tell expand on that reasoning."
That I will be glad to do.... Name any generally accepted moral principle and I will show how it can be traced back to humanity. Also my opponent again makes the mistake of saying Atheism cannot account for UNIVERSAL MORALS, but sadly neither can Thiesm as we have seen and that is not the subject of this debate.
Lets do an example of morality being human using the Moral principle that killing is wrong.
1. Humans don't want to be killed 2. People, as a general rule, want to do what they feel is right. 3. Therefore people(in general), because they don't want to be killed, have said that killing is wrong 4. Therefore it is generally accepted among people killing is wrong 5. Hence killing is considered an Immoral Act.
Justification behind 1-5.
1. The Urge for Survival in all things is primary, it has been seen through the existence of life 2. The concept of the conscious tells us that we want to do the right thing. So people are deterred by the idea that what they may be doing is wrong. 3. Combination of 1&2 plus the fact that people made this decree to create the deterrence I mentioned in 2 4. A summary of 4 as a general rule 5. Putting the concept of right/wrong into Morals
Thank you, Matt
Good luck to my opponent, and I urge everyone to look critically at all arguments

Table 14: The second round of a debate example.

The current speech in the debate is from the user {Side1_label}:
Metz said, "Also this really doesn't say why Theism can actually account for universal morals. People disagree on religion. If Morals were universal then the scenario my opponent laid out wouldn't exist."
I thought I made this clear in my opening statement: the God I am referring to is the Christian God. So, when you say that a theistic world view can not account for morality, because of all the different religions, then yes I would agree with you, because if there are many different religions that judge humans, then there would not be one standard to which morality is based upon. So, when I mention God I am only referring to the Christian God. Now that my view on the issue has been restated, any argument used by Metz (con), about why a theistic universe can not account for morality either; because of all the different God's derived from different religions, will be irrelevant. Since I am only referring to One religion, which is Christianity. And since there is just one God then there is only one moral standard, thus God can account for what is wrong or right.
"It never says Atheism needs to account for universal morals, just morals."
Okay, lets have it Metz's way, Atheism does not need to account for universal laws, just morals in general, very well. If there are no universal morals that prohibit certain acts of crime such as rape, murder, polygamy, theft, ect... then why am I obligated to obey those morals? Why can't I just abide by my own moral standards, since there are know Universal ones? For instance, I could think that its just fine to murder, rape, steal ect... because that's what I believe is right. So if there are just MORALS, to be defined by anybody, and no UNIVERSAL MORALS then who is to say that my morals are wrong? Whose to say anything is wrong for that matter? Once again the argument for an Atheistic world view on morals collapses on itself because it can not account for what is truly right or wrong.
Metz said, "Its wrong because people say its wrong..." this quote is in response to me asking why murder is wrong.
So are you saying that if the majority of the human population say that gays should not be aloud to marry, then that is automatically the moral standard? This is exactly my point, if a moral act is defined by what people say is moral then anything from the act of murder to a little white lie must be accepted by humanity. For instance, I could say murder is right because I said so. Also, a real life example is, 'people' started 'saying' that Jews should be considered sub human and thrown into concentration camps, but did this make it right? No, of course not. You see, I can say that Hitler was wrong because God commands it in the Bible, "thall shall not murder," its the 6th commandment. But, the best that my opponent can say, is, "Hitler's acts of genocide were immoral because the Jews were being murdered against there own will." Well, my opponent's statement just begs the question, So? Who says the Jews have a right to live in the first place? After all, millions were saying that Jews did not have the right to live. So which side is right, and why? For, to simply say that Hitler was wrong because he murdered Jews against there will is NOT answering the WHY? It only states a mere fact.
Another example, John Locke, a well known philosopher who came up with idea of the Social Contract, this contract was to ensure that every human being was born with the right to live. Now, the question I have for Locke's thinking, along with anyone else who agrees with him, is this, WHY? Why are humans born with the right to live? I do not see a logical answer without God in the equation.
1. Humans don't want to be killed. 2. People, as a general rule, want to do what they feel is right. 3. Therefore people(in general), because they don't want to be killed, have said that killing is wrong 4. Therefore it is generally accepted among people killing is wrong 5. Hence killing is considered an Immoral Act.
Again, not only does this example have nothing to do with WHY murder is wrong. But, what you have described here is Western Civilization for the past 200 years or so, ONLY. This certainly is not the case in the Philippines, the Middle East, or any other extremely violent area in the world. This totally disproves your point above. And not only that, but what about in past history, such as the Dark Ages where many people considered murder to be a normal act, in order to get food or money so they could fill there bellies. So, when murder became an act that was generally accepted among the people, such as in the Philippines, the Middle East, and any other extremely violent areas in the world or from times in the past, such as the Dark Ages, is it then morally acceptable to murder? I see no reason why not, under an Atheistic universe.
In conclusion, I still believe that my opponent has failed to answer the why for his reasoning? For instance, everyone knows that people don't want to be murdered. But the question I am asking is, why is it wrong for people to be murdered? To say because people don't want to be murdered is not answering the question. Because why should a murderer care about what his or her victim wants if its just a question of morals and not universal morals? Also my opponent argue that a theistic world view can not answer for this question either. Well, that isn't answering the question, that's just pointing fingers.
As I have said, I am a Christian and will be basing my arguments off a single religion and a single God. Now, my opponent may take this as a opportunity to criticize my religion like he did in his last argument, some what. If my opponent starts to argue that Christianity is not perfect and why should God be the ultimate judge this is still not answering the question of why anything is right or wrong to do anything. Again its just pointing fingers.
Now this is something I have only touched on a little, I can say something is wrong or right because I believe there is an ultimate judge, God. This means there are universal laws of morality, that are absolute, and everyone must abide by them. In an Atheistic universe the only thing that can judge morality is humanity which I have proved is inconsistent and ultimately can not account for morality at all.
Again I encourage the voters to listen to the video above it really illustrates my point.
My dad also had a personal relationship with Dr. Bahnsen (the man debating in the video). My dad told me that after the debate between Bahnsen (Christian) and Stein (Atheist) they continued debating each other through emails and letters, after a couple weeks of going back and forth with their arguments Stein eventually wrote "I don't really have any answers for you, but I'm just not ever going to agree with you."
Please answer the fallowing questions...
1)Why should a murderer care about what his or her victim wants if its just a question of morals? 2)Why are humans born with the right to live? I do not see a logical answer without God in the equation. 3)Are you saying that if the majority of the human population say that gays should not be aloud to marry, then that is automatically the moral standard?
Thank you, charles15
The current speech in the debate is from the user {Side2_label}:
I will start with the three questions my opponent proposed to me at the end of his last argument.
1)Why should a murderer care about what his or her victim wants if its just a question of morals? There are, obviously exceptions to my rule of moral deterrence. But remember my proof established it as a general rule. This has nothing to do with Atheism at all, when someone murders someone they are not in a state of mind that would disregard any moral background whatsoever. Even if we assume a theist stance, these people have committed a sin, so therefore God as much fails to uphold morals as would Atheism. Also the Psychological consequences would be felt later as philosopher and psychologist Fyodor Dostoevsky laid out in his book Crime and Punishment.
2)Why are humans born with the right to live? I do not see a logical answer without God in the equation.
I hate to say this but its the shocking truth... We are born with the right to live because we have a will to live. If nobody wanted to vote would it be considered a right? This will to live is also not traceable to god, but to the fact that humans are just animals with the ability to reason. Unless my opponent wishes also to deny evolution and biological fact then this has to be accepted. The most primal instinct of live is to preserve itself. This is where morals come from as I have repetedly argued. Humans judging others and therefore judging themselves.
3)Are you saying that if the majority of the human population say that gays should not be aloud to marry, then that is automatically the moral standard? This is Mob rule, not necessarily morality. But not to criticize to much but I have that the same would be said of God. If the Bible says it then its wrong, which seems to be a common belief about gay marriage. As I said The base of Morality is humans, Gay marriage does not threaten anybody, so it is therefore it is not sought to prevent like killing would be. People Judge others in Gay Marriage but it does not affect them so the link between natural morals is flawed. A society may come to the belief that gay marriage is immoral, but it is not intrinsically immoral, and this seems to be what is happening in the world today.
Now on to the remaining arguments:
"Since I am only referring to One religion, which is Christianity. And since there is just one God then there is only one moral standard, thus God can account for what is wrong or right."
This really doesn't mean that everyone would follow this God, so are these people immoral? People believe do different extents, and so therefore have different morals even assuming the same God and religious texts and Church structure. In order for God to be as great a source for morals as my opponent claims we would need to abandon any remaining Autonomy and become almost robotic in our beliefs, an act which is, ironically, immoral in either world.
"So if there are just MORALS, to be defined by anybody, and no UNIVERSAL MORALS then who is to say that my morals are wrong? "
Not defined by anybody, defined by humanity. Humans Judge, you are judged by your fellows, you judge others and so judge yourself. Every step of the way there are checks.
" If there are no universal morals that prohibit certain acts of crime such as rape, murder, polygamy, theft, ect... then why am I obligated to obey those morals? Why can't I just abide by my own moral standards, since there are know Universal ones? For instance, I could think that its just fine to murder, rape, steal ect... because that's what I believe is right."
First, I Never said Atheism CAN'T account for universal morals merely that it was not my burden to prove that it did. Also, you can have your own moral standards, I know many people that have there own and are not killers, for example I think we have a moral obligation to fairness and to help people, I have friends that have a more sink or swim attitude. These morals can be relative, this is part of what shapes humanity, to accept that all morals are dictated to us really destroys that humans element. However when we get into killing, people judge more carefully, people are afraid. For the sake of protection and for moral order HUMANS establish moral rules, such as that against killing. Atheism can account for Morality because it was humans all along that accounted for morality.
"Who says the Jews have a right to live in the first place? " They do... They have a will to live that is as strong as that of any other. This turns Life into a right intrinsic of humanity. Thus when the Jews were killed Hitler was taking an intrinsic right and the act was thus, immoral. I already addressed the other problem at the beginning.
"everyone knows that people don't want to be murdered. But the question I am asking is, why is it wrong for people to be murdered?"
You gave me the answer right there. This bring me back to the same Will to Live argument. It is wrong because people have a will to live. Because they have this will it becomes a recognized right to live. Thus when someone violates this right the act is immoral in most circumstances(there are exceptions to every moral idea).
Voters, When you are reading this debate you need to think about whether or not you would logically do some of the things my opponent has said in his examples, and whether you would want them done to yourself. You also must recognize that Murder's generally have an altered or disturbed state of mind that could be influenced by such things as Alcohol that means in Either world these people don't respect morals.
The key question here is: Did my opponent prove that without God morals COULD NOT exist? Or did I prove that morals COULD exist in such a world. Remember the resolution asks could, which means "is it possible"
Thanks, Metz

Table 15: The third round of a debate example.