Auditing Counterfire: Evaluating Advanced Counter-argument Generation with Evidence and Style

Anonymous ACL submission

Abstract

The ability of large language models (LLMs) to generate evidence-based and stylistic counter-arguments is crucial for enhancing online discussions. However, there is a research gap in evaluating these models' practical effectiveness in real-world applications. Previous studies often overlook the balance between evidentiality and stylistic elements necessary for persuasive arguments.

We created and audited Counterfire, a new dataset of 32,000 counter-arguments generated by non- and finetuned-LLMs with varying prompts for evidence use and argumentative style. We audited models like GPT-3.5, PaLM 2, and Koala, evaluating their rhetorical quality and persuasive abilities. Our findings showed that while GPT-3.5 Turbo excelled in argument quality and style adherence, it still fell short of human standards, emphasizing the need for further refinement in LLM outputs. Code and data are available at https://anonymous.4open.science/r/Style_control-2018/.

1 Introduction

Counter-argument generation refers to systematically creating opposing viewpoints or arguments in response to a given statement, hypothesis, or position as a rebuttal, undercut, or undermining of the original claim (Walton, 2009). Generating compelling counter-arguments grounded in evidence is a critical aspect of natural language processing, with applications in argument refining, argument mining, and text evaluation.

Our work addresses the research gap in evaluating generative text versus its application in real-world scenarios. Previous studies on counter-argument generation have focused on various methods: Bilu et al. (2015) used rule-based techniques, Hidey and McKeown (2019) employed data-driven strategies, and Alshomary et al. (2021) aimed at undermining the weakest claim. The Project Debater system (Bar-Haim et al., 2021; Slonim et al., 2021) engages in competitive debates using an argument mining framework with a corpus of about 400 million articles. Additionally, Hua et al. (2019) and Jo et al. (2021) incorporated evidence in counter-arguments. However, these studies often overlook the stylistic component of arguments – a critical strategem for their practical effectiveness for attitudinal change or persuasion.

Following the call for controllable composition in natural language generation (Chen and Yang, 2023; Kumar et al., 2023), particularly in scientific summarization (Ding et al., 2023), we argue for the need to evaluate LLMs' abilities for the controlled generation of stylized counter-arguments, accommodating preferences for evidence and style, auditing evidence integration, and prompt adherence while focusing on ultimate effectiveness. Therefore, this study audits the controlled counter-argument capabilities of three LLMs—GPT-3.5, PaLM2, and Koala—and two fine-tuned variants, focusing on adherence to evidence and style instructions. We create and evaluate a novel dataset for the political domain, annotated with human and automatic metrics, examining how top-performing counter-arguments compare to human ones in persuasion. We offer two key contributions to the counter-argument generation literature:

- A new style dimension to control arguments through evidentiality and reciprocity.
- Comparative insights into human- vs. LLM-generated fine-grained style and counter-argument structures.

2 Background

Generating stylized text with LLMs is feasible along those dimensions which have been previously studied in depth, such as readability (Pitler and Nenkova, 2008; Collins-Thompson, 2014), formality (Chawla et al., 2019; Chhaya et al.,

2018) and politeness (Yeomans et al., 2018; Althoff et al., 2014; Danescu-Niculescu-Mizil et al., 2013a). However, the state-of-the-art in characterizing argumentative style (Lukin et al., 2017; El Baff et al., 2020; Ben-Haim and Tsur, 2021; Al Khatib et al., 2020) needs more nuance to study political discussions. Furtherore, no prior paper has compared three LLMs - simple and finetuned - for argumentative style generation. While an excellent benchmark (auto- and human-) evaluation paper on news summarization by Goyal et al. (2022) exists, it does not include argument generation, finetuning, or style evaluation.

Herein lies the novelty of our approach: applying concepts from social science for LLM prompts offers a theoretically grounded method to enhance argumentation. Political communication research conceptualizes social media platforms as a space for 'internal reasoned dissent'(Rinke, 2015), where users engage with a "number of publicly available ideas, opinions, and arguments (and) different points of view"(Rinke, 2015) in the form of mediated deliberation. Recent work on political discussions in social media distinguishes analytical arguments from social arguments (Esteve Del Valle et al., 2018; Friess and Eilders, 2015; Jaidka, 2022; Rowe, 2015):

Analytical arguments: facilitate a fact-oriented discussion through

- **Justification**: Employing tangible evidence to support claims.
- Constructiveness: Using logic and rational arguments to move towards a consensus, and

Social Aspects: foster a constructive and inclusive dialogue through

- **Reciprocity**: Demonstrated through the interactivity of a discussion and whether participants invite engagement from each other, and
- Empathy and Respect: Demonstrated through the tone of responses that acknowledges and respects different viewpoints.

An examination of the actual distribution of these styles in the annotated CLAPTON corpus (Jaidka, 2022) suggests that, on the Reddit ChangeMyView platform, authors overwhelmingly prefer to write counter-arguments that follow a Justification (30%) or a Reciprocity (25.8%) style rather than Constructiveness (6.6%). Therefore, as our study also uses the Reddit ChangeMyView dataset, we have chosen to curate the Counterfire counter-argument dataset with **Justification and Reciprocity styles** styles of counter-arguments.

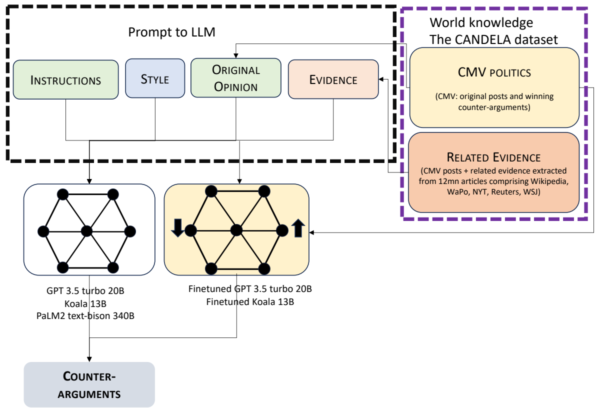

Figure 1: Experimental framework for generation.

3 Method

Figure 1 illustrates our framework, which uses facts from seq2seq intermediate outputs to create domain-injected prompts, generating relevant, logical, and grammatical counter-arguments from off-the-shelf and fine-tuned LLMs.

First, we used the retrieval system of Hua et al. (2019) that provided credible counter-evidence for the dataset of Original Posts. Second, we applied zero-shot style-focused prompts (reported in Table 1 and exemplified in Figure 2) to generate counter-arguments to CMV (Change My View) posts. Example counter-arguments generated by GPT-3.5 Turbo are reported in Figure 3. Third, we reran step 2 on fine-tuned GPT-3.5 Turbo and Koala 13B models that were trained on the argument and evidence dataset from CMV, to examine if adherence to style-related instructions changed in fine-tuned models designed to recognize and emulate the predominantly reciprocal and justification styles in the training set. PaLM2 often generated supporting arguments instead of counter-arguments, leading us to exclude it from fine-tuning. The configuration parameters for the LLMs are reported in the Appendix. The training loss plot and the hyperparameter settings for fine-tuning are reported in the Appendix.

4 Data Collection

The data collection process is twofold: initially using zero-shot prompting and refining through finetuning techniques.

Style	Prompt
Plain	Use a writing style that focuses on using the evidence and being convincing.
Reciprocity	Use a writing style that asks questions designed to elicit opinions or information from the user.
Justification	Use a writing style that focuses on fact-reporting or fact-checking, finding common ground, and providing personal or statistical evidence with references.

Table 1: The three variants of the style specifications added to the LLM prompt, adapted from the original definitions (Jaidka, 2022).

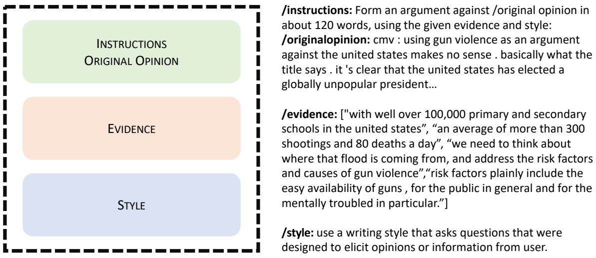

Figure 2: Example prompt for generating a reciprocal counter-argument.

Original opinion: parents who refuse vaccination of their children must sign a form of accountability so if their child dies from medical complications (...) **Evidence:** ["about 1 percent of u.s. parents get exemptions for their children , but the figure is higher in some areas","the fear is that parents who don't vaccinate or who delay vaccinating their children put their own kids as well as others at risk for developing preventable disease , such as whooping cough"]

Candela output: If you want to change your view, do you really think we should force them to do things they don't want to do? (...)

GPT-3.5 turbo output with no style: Firstly, it is not always easy to determine the exact cause of a child's illness or death, and therefore it may not be fair to blame it solely on the lack of vaccination. Secondly, some parents may not have access to vaccinations in their area or cannot afford them, and punishing them would be unfair. (...)

GPT-3.5 turbo output with reciprocity: While I understand the frustration towards parents who refuse to vaccinate their children, do you think holding them accountable for their child's potential death is the best approach? What if the child had an allergic reaction to the vaccine or experienced complications that were rare but severe? How would this legal contract work in those situations? (...)

GPT-3.5 turbo output with justification: The proposed idea to hold parents accountable for not vaccinating their children is misguided and unfair. First and foremost, compulsory vaccination violates personal freedom. Parents have the right to make decisions for their children, including when it comes to medical procedures. (...)

Figure 3: An example input statement and the generated counter-arguments for the CMV dataset from GPT-3.5 turbo.

Baseline: The primary baseline uses the Candela dataset by Hua et al. (2019), sampling (a) 70,000 English posts and winning counter-arguments related to politics from r/ChangeMyView, and (b) associated evidence (deduplicated and ranked) from a database of 12 million articles, including sources like Wikipedia, Washington Post, New York Times, Reuters, and The Wall Street Journal. An additional random sample of 2000 rows of posts and evidence was created for further analysis.

Generating Stylized Counter-Arguments: Five off-the-shelf and fine-tuned LLMs were prompted with the original post and evidence from the subsampled Candela dataset. Prompts included stylistic variations based on prior work (Steenbergen et al., 2003; Jaidka, 2022). We also created a set of prompts without curated real-world evidence. Figure 2 shows a sample prompt for generating a reciprocal counter-argument, with style instructions listed in Table 1. We generated 32,000 counter-arguments from the five LLMs using these prompts (2000 inputs x 3 styles x 5 LLMs).

GPT-3.5 turbo: GPT-3.5 turbo is a language model based on GPT (Brown et al., 2020) capable of generating human-like text. The GPT-3.5 turbo is the latest and most capable model in the GPT-3.5 turbo series. We engineered prompts for style control and provided the same passages as we do to our baseline for the better factual correctness of generations.

Koala 13B: Koala-13B (Geng et al., 2023) has been created by finetuning LLaMA (Touvron et al., 2023) using EasyLM on high-quality deduplicated public datasets, such as a high-quality dataset curated with responses to user queries from larger, more capable, and close-sourced ChatGPT. Recent results have suggested that high-quality training data helps overcome problems faced by smaller models such as LLaMA and sometimes also gives competitive performance to larger models for specific tasks.

PaLM2 Text-Bison: Google's Pathways Language Models 2 series offers the text-bison generation model (henceforth referred to as PaLM2), trained on 340 billion parameters. PaLM2 models are notable for their improved multilingual, reasoning, and coding capabilities. They are trained on multilingual text in over 100 languages, and their datasets include scientific papers, web pages, and public source code, enabling better logic, common sense reasoning, mathematics, and programming language proficiency.

Finetuned variants of GPT-3.5 turbo and Koala GPT-3.5 turbo was finetuned using OpenAI's Application Programming Interface (API) for three epochs. Finetuning for Koala-13B was done on the training set of 70,000 input and counter-argument pairs from our primary dataset using Colab Nvidia A100 GPU (a different random sample of 2,000 pairs was used for subsequent counter-argument generation). The model was loaded in memory with 4-bit precision and double quantization using 4-bit NormalFloat and paging (Dettmers et al., 2023). After quantization, we added LoRA adapters (Hu et al., 2021) for each layer. For inference on our sample, the model was partially dequantized, and computations were done with 16-bit precision.

5 Analyses

Our analytical approach comprised four steps. First, we obtained human assessments across five quality dimensions. Second, we performed validation tasks to audit the ability of LLMs to adhere to instructed prompts in terms of fact and style. Third, we conducted an automatic rhetorical analysis to compare the generated counter-arguments along many dimensions of readability, rhetorical intent, and discourse. Finally, we compared human perceptions of the effectiveness of the counter-arguments generated by GPT-3.5 Turbo. Crowdsourced quality assessments were launched on Amazon Mechanical Turk, with all the detailed instructions provided in the Appendix.

5.1 Human Evaluation

We conducted a comprehensive manual evaluation of the quality of the generated counter-arguments to identify the most-preferred LLM. Therefore, we launched an Amazon Mechanical Turk task to obtain eight annotations per argument on facets such as content, grammaticality, logic, relevance, and overall effectiveness (Goyal et al., 2022). Results are reported in subsection 6.1.

5.2 Validation Tasks

We performed two validation tasks to audit the ability of LLMs to adhere to the instructed prompts: (a) Fact integration, assessing the incorporation of evidence, and (b) Style validation, evaluating whether the outputs reflect the expected discussion style.

Fact Integration For fact integration validation, we analyzed whether our prompts effectively guided the LLMs to incorporate the provided evidence into the generated counter-arguments. This involved comparing the similarity and absolute overlap of evidence with the outputs from the off-the-shelf LLMs, using metrics such as BERTScore (Zhang et al., 2019) and ROUGE-1 (Lin, 2004). Results are reported in subsection 6.2.

Style Integration For style validation, we examined whether the LLMs could integrate the expected style into the outputs. This was conducted using both automatic and human methods. Automatically, we fine-tuned OpenAI Ada models on the CLAPTON dataset (Jaidka, 2022) to label the presence of justification and reciprocity in the generated outputs. Human assessments of style were obtained through an Amazon Mechanical Turk task, where a random sample of 100 counter-arguments generated by each LLM variant received five annotations per argument on the discussion facet labels of justification and reciprocity. We then measured their θ, which overcomes many of the challenges of evaluating inter-annotator agreement on a five-point scale with chance-based metrics and was proposed by Passonneau and Carpenter (2014) and applied by other scholars (Jaidka et al., 2023; Davani et al., 2022; Jaidka et al., 2024). Unlike chance-based metrics, which have wide error bounds, model-based measures consider the actual categories of items in the corpus and the prevalence of each label to report the accuracy of reporting the correct answer through an expectation maximization approach. Based on recommended thresholds (Passonneau and Carpenter, 2014), we considered the inter-annotator reliability satisfactory as $\theta \geq 0.65$. Results are reported in subsection 6.3.

5.3 Rhetorical Analysis

We conducted an automatic rhetorical analysis to compare the generated counter-arguments along many dimensions of readability, rhetorical intent, and discourse.

Readability Metrics: We used the textstat package to calculate various readability metrics, such as Flesch-Kincaid grade, Flesch Reading Ease, Gunning Fog index, and Smog index.

Rhetorical Intent: We characterized the generated counter-arguments according to the presence of rhetorical moves related to argument alignment, authority, and persuasion. The Alignment and Authority in Wikipedia Discussions (AAWD) corpus (Bender et al., 2011) provided a basis for counter-argument analysis, with annotated phrases indicating agreement or disagreement. Authority moves express credibility, while persuasive moves include features like politeness, contingency, expansion, claims, and premises. Additionally, we utilized Debater API scores to evaluate stance and quality, with scores ranging from 0 to 1.

Discourse Analysis: Persuasive moves were measured using the Python convokit toolkit, which searches for various lexical features reflective of different categories. These features have been applied to study online persuasion and model politeness and trustworthiness in social media posts (Danescu-Niculescu-Mizil et al., 2013b; Niculae et al., 2015). Results are reported in subsection 6.4.

5.4 Human preference analysis

Finally, to evaluate the effectiveness of the counter-arguments, we once again surveyed Amazon Mechanical Turk workers to rank the best-performing counter-arguments against human-written ones, following the design of similar user experiments reported in prior work (Goyal et al., 2022). The survey aimed to examine whether users prefer justification- or reciprocity-style counter-arguments. We collected 10,000 counter-argument rankings from 1879 respondents. Results are reported in subsection 6.5.

6 Results

6.1 Human quality assessments

Figure 4 shows the human evaluation of counter-arguments generated by various models, focusing on content, grammaticality, logic, relevance, and overall effectiveness. Each boxplot represents the median, interquartile range, and range, with outliers indicated as dots outside the whiskers. The different colors and box styles correspond to various models and style prompts. GPT-3.5 Turbo models consistently outperformed Koala 13B and PaLM2 across all parameters. Notably, GPT-3.5 Turbo "No Style" counter-arguments scored higher in grammaticality and logic than those generated with justification and reciprocity styles, which had a broader spread in overall effectiveness and relevance. This suggests that while stylistic variations can enrich counter-arguments, they may also introduce variability in quality. There appears to be no significant advantage of fine-tuning LLMs. On the other hand, the baseline Candela outputs were perceived to be less grammatical, relevant, coherent, and less preferred than the counter-arguments generated through GPT-3.5 turbo, and the differences were statistically significant after Bonferroni correction for multiple comparisons ($p < 0.001$).

6.2 Fact integration

Table 2 presents a comparative evaluation of different large language models (LLMs) and their prompting strategies using various automatic evaluation metrics. The first part of the table compares the fact integration capabilities of GPT-3.5 Turbo, Koala-13B, and Palm-2 using BERTscore (F1 value) and ROUGE-1 (Recall), suggesting that there is not much to choose between the three based on adherence to content. Yet, note that the average BERTScore F1 value across the three LLMs is 0.725, and the average ROUGE-1 recall is 0.313, suggesting that while LLMs are effective at paraphrasing evidence into counter-arguments, they have a lower absolute overlap in the words used.

As we found that GPT-3.5 Turbo was the most preferred by humans in our quality evaluations, we offer a deep dive into a comparison of its three different prompting strategies rather than try and compare all eight variants together. Therefore, the second part of the table evaluates the content and style adherence of just GPT-3.5 Turbo under different prompting strategies, compared to a baseline (Candela). The metrics used are ROUGE-1, ROUGE-2, ROUGE-L, and BLEU. The "No Style" prompt for GPT-3.5 Turbo achieves the highest scores across all the metrics, suggesting that no styling instructions allow better content adherence that outperforms the Candela baseline and other style-specific prompts. This indicates that while stylistic prompts can tailor the generated content to specific styles, they might compromise the overall content quality and factual accuracy.

6.3 Style integration

Results for the style validation are reported Table 3. The first half of the table reports an automatic evaluation of style, where PaLM2 demonstrates the highest accuracy (0.49), significantly outperforming GPT-3.5 Turbo (0.17) and Koala-13B (0.09). While PaLM2 excels in integrating the reciprocity style, GPT-3.5 Turbo is more effective in incorporating the justification style. For justification adherence, GPT-3.5 Turbo leads with an accuracy of 0.42, followed by Koala-13B (0.26) and PaLM2 (0.22).

Next, for the human evaluation of style integration, we have reported the inter-annotator reliability (θ) (Passonneau and Carpenter, 2014) for the

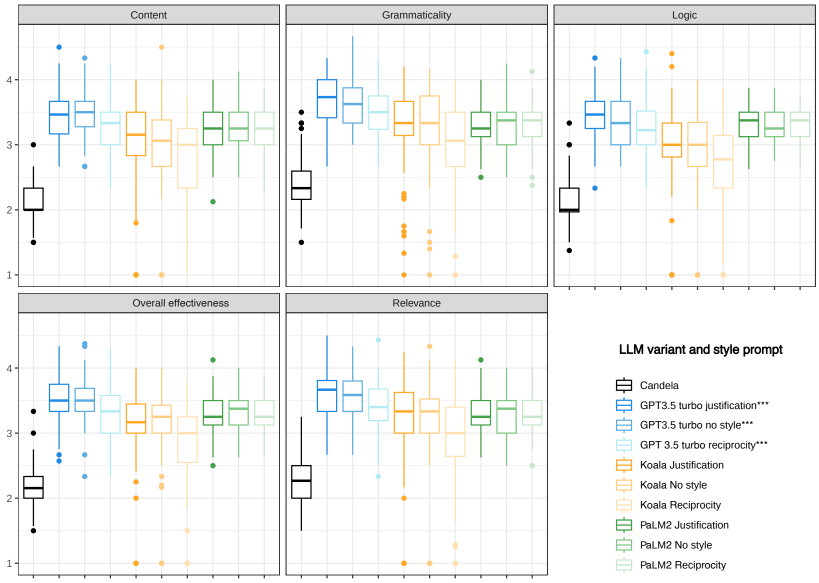

Figure 4: Results from the human evaluation of generated counter-arguments. GPT-3.5 Turbo outperforms Koala 13B and PaLM2 and their fine-tuned variants on Content, Grammar, Logic, Relevance, and Overall effectiveness.

Metric	**GPT-3.5 Turbo**	**Koala-13B**	**PaLM2**
Comparing LLMs			
BERTscore (F1 Value)	0.7312	0.7271	0.7175
ROUGE-1 (Recall)	0.3556	0.3631	0.3103

Metric	**Candela**	**GPT-3.5 Turbo No Style**	**GPT-3.5 Turbo Justification**	**GPT-3.5 Turbo Reciprocity**
Comparing Prompting Strategies				
ROUGE-1	0.24 (0.07)	**0.33 (0.07)**	0.17 (0.06)	0.17 (0.06)
ROUGE-2	0.03 (0.03)	**0.10 (0.06)**	0.02 (0.02)	0.01 (0.02)
ROUGE-L	0.21 (0.06)	**0.29 (0.07)**	0.15 (0.05)	0.14 (0.04)
BLEU	0.00 (0.01)	**0.06 (0.06)**	0.00 (0.01)	0.00 (0.01)

Table 2: Automatic evaluation metrics and fact integration scores for different models.

Style Integration			
Style	**GPT-3.5 Turbo**	**Koala-13B**	**PaLM2**
Automatic Evaluation: Style (Accuracy)			
Reciprocity	0.17	0.09	**0.49**
Justification	**0.42**	0.26	0.22
Human Evaluation: Style Integration			
Style	**θ (Inter-annotator Accuracy)**		
Reciprocity	0.9682		
Justification	0.7680		

Table 3: Evaluation of style integration through (a) automatic labeling with fine-tuned models, and (b) crowdsourced human labels. θ is the average annotator accuracy across true-positives and negatives (Passonneau and Carpenter, 2014).

reciprocity and justification variants of the GPT-3.5 Turbo counter-arguments (as they received the highest quality assessments in subsection 6.1). The θ values of 0.9682 and 0.7680 indicate strong agreement among annotators and also offer a stronger signal than the automatic metrics that the models effectively integrated the specified styles into their outputs. These values also surpass the recommended threshold of 0.65, confirming the reliability of the annotations.

6.4 Rhetorical insights

In Table 4, we report a comprehensive evaluation of readability and rhetorical intent across different models and prompting strategies.

Readability Metrics: The readability metrics indicate that GPT-3.5 Turbo No Style has the highest score (12.81), indicating higher complexity, followed by Justification (12.75) and Reciprocity (11.79). Candela has the lowest complexity (6.40). Similar trends are observed across other readability metrics, with GPT-3.5 Turbo No Style demonstrating the most complexity in terms of Flesch Reading Ease, Gunning Fog, and Smog Index.

Metric	Candela	GPT-3.5 Turbo No Style	GPT-3.5 Turbo Justification	GPT-3.5 Turbo Reciprocity
		Automatic Evaluation: Readability (0 to 1 scale)		
Flesch Kincaid Grade	6.40 (2.18)	**12.81 (2.07)**	12.75 (2.07)	11.79 (2.08)
Flesch Reading Ease	**83.10 (10.41)**	40.94 (11.31)	41.78 (10.62)	46.23 (11.37)
Gunning Fog	8.85 (2.05)	**15.05 (2.23)**	15.03 (2.23)	13.93 (2.17)
Smog Index	8.53 (2.39)	14.85 (1.89)	**14.87 (1.68)**	14.09 (1.72)
		Rhetorical Intent		
Move Type	**Human-written Reddit Counter-argument**	**GPT-3.5 Turbo No Style**	**GPT-3.5 Turbo Justification**	**GPT-3.5 Turbo Reciprocity**
		Alignment Moves		
Positive	12	0	4	2
Negative	12	0	4	6
		Authority Moves		
Experiential	10	0	0	6
External	10	0	4	2
Forum	10	0	4	4
Social Expectations	8	0	0	2

Table 4: Evaluation of readability and rhetorical intent of different models. The alignment moves identified in Counterfire outputs, based on AAWD features, highlight the discursive richness of human counter-arguments (Bender et al., 2011).

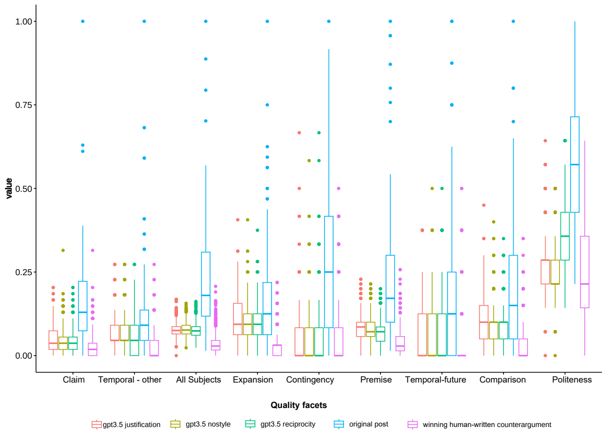

Figure 5: Results from the automatic evaluation of argumentation using Convokit highlights the discursive richness and low politeness of human counter-arguments.

Rhetorical Intent:

- **Alignment Moves:** Human-written counter-arguments contain the most positive and negative moves (12 each), highlighting rich argumentative content. These moves are examples of social acts involving agreement or refutation in argumentation. In contrast, GPT-3.5 Turbo variants have fewer alignment moves. The Justification style includes more positive (4) and negative (4) moves than Reciprocity.
- **Authority Moves:** Authority moves are markers of social expectations, credentials, experiential claims, forum claims, and external claims. Human-written arguments lead across all categories. Specifically, the Reddit counter-arguments contain 12 positive and negative alignment moves, showing explicit agreement and positive alignment (e.g., praise thinking), as well as opposing alignment (e.g., criticizing or doubting). GPT-3.5 Turbo Reciprocity includes a notable number of experiential (6) and forum (4) moves, while Justification leads in external (4) moves.

Certain moves in the AAWD corpus, such as 'credentials' and 'experiential,' had no counts or low counts among the GPT-3.5 Turbo variants, highlighting domain differences compared to the AAWD corpus. The table suggests that human-written arguments are the most argumentatively rich and diverse, with more unique moves across different categories than the generated outputs.

Discursive features: With reference to Figure 5, GPT-3.5 Turbo-written counter-arguments are typically at par with each other concerning most discursive features. However, they significantly differ ($p < 0.001$) from human-written counter-arguments in covering more claims, temporal features, reference to subjects, premises, comparisons, and even politeness. Human-written counter-arguments have fewer claims with greater specificity, offering a more focused and less polite counter-argument.

Overall, the findings suggest that while human-written arguments are richer and more diverse in rhetorical intent, GPT-3.5 Turbo excels in readability metrics, especially in generating more complex texts. This highlights the balance between generating readable content and maintaining rhetorical richness in counter-arguments.

6.5 Human preference analysis

Figure 6 provides insights into the persuasiveness of GPT3.5-generated counter-arguments relative to the corresponding styles of human-written counter-arguments. Taken together with findings from Figure 5, the findings suggest that the highly focused, specific, and less polite human counter-arguments are more persuasive to humans than GPT3.5-generated counter-arguments.

More specifically, Figure 6 illustrates that in a comparison of 2000 original posts and counter-arguments sourced from ChangeMyView and the Counterfire corpus, humans find human-written reciprocal-style (Mean preference = 2.24 out of 5; lower score is better) and justification-style counter-arguments (Mean preference = 2.19 out of 5) more preferable to those written by GPT3.5 (Means 2.93 and 2.56 respectively) (Welch Two Sample t-Test, $p < 0.001$). The low preference for Justification implies that while these directives resulted in comprehensive, evidence-backed counterarguments, they may be less engaging than reciprocal counterarguments. The findings suggest an interesting tradeoff between fact integration and style while generating counter-arguments that inspire future research investigations.

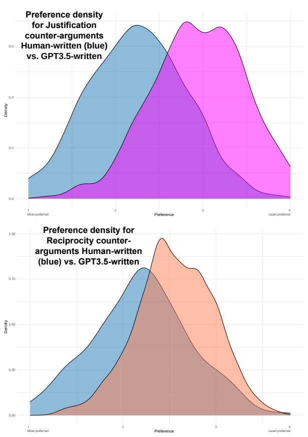

Figure 6: User preference analysis for human-written (blue) vs. GPT-3.5-written counter-arguments for justification (left) and reciprocity (right) highlights user preferences for reciprocal rather than evidence-based counter-arguments.

7 Error Analysis

A detailed error analysis is provided in the Appendix to better understand how LLMs fare on human evaluation. Furthermore, as some counter-argument generation tasks might favor greater content adherence, we have also reported the counter-arguments scoring high and low on ROUGE-L scores. In general, we observed that examples scoring highly on human quality assessments demonstrate practical and relevant arguments while low-scoring ones often suffer from incompleteness or lack of direct relevance. This was especially the case for counter-arguments from PaLM 2, which scored poorly because of broad statements and lack of direct relevance. Counter-arguments also scored poorly because of their repetitiveness and lack of clarity, especially on counter-arguments on specialized subjects. On the other hand, ROUGE-L F1 scores do not always align with our perception of the substantive quality or originality of a counter-argument.

8 Discussion and Conclusion

Our findings demonstrate interesting insights regarding (a) a classic trade-off in content versus style, where high-content arguments struggle to maintain quality expectations and vice versa, and (b) despite referencing the same evidence, GPT-3.5 turbo arguments succeed at overall persuasiveness and relevance compared to state-of-the-art seq2seq baselines. However, (c) human-written arguments are rhetorically richer and (d) usually preferred by users over the generated counter-arguments, which provides exciting avenues for future exploration.

The findings underscore significant implications for generating and evaluating counter-arguments using language models. On the one hand, along standard discussion quality dimensions, the models exhibit a notable proficiency in rephrasing content with relevant evidence, even with minimal lexical overlap, and demonstrate exceptional integration of argument styles, as evidenced by the high scores in style adherence, particularly in the 'reciprocity' category. GPT-3.5 turbo, in particular, stands out for its superior performance in argument quality evaluations, and the differences in the use of rhetorical moves and user preferences suggest that these counter-arguments comprise more innovative and convincing uses of evidence. Yet, on the other hand, the counter-arguments fall short of human preferences for effective counter-argument generation, elevating our understanding of LLM auditing frameworks to highlight the gap between what we measure and how we use them. While preferred to LLM outputs, human-generated counter-arguments also show more complexity and variety in argumentative tactics and herein may lie their persuasive advantage over LLM outputs.

9 Limitations

We focused on evaluating the style and quality of the arguments generated while presuming that the fact retrieval system adapted from Hua et al. (2019) was working perfectly. Furthermore, we are limited by the Candela dataset to focus only on English political posts. Before applying the dataset for further model-finetuning, we recommend annotating the generated counter-arguments to ensure veracity and pre-empt the selection or curation of irrelevant facts in the list of evidence (Mendes et al., 2023). Finetuning is a time-, memory-, and data-intensive process. In the case of GPT-3.5 turbo, our experiments were done using API calls with high latency. We observed inconsistencies in PaLM 2 outputs. In 10% cases, it generated an argument in support of the input instead of against it; therefore, we did not finetune it.

Our work was limited in scope because it does not develop dynamic models that accommodate a conversation partner's stylistic choices in generating a finely tailored counter-argument for greater persuasive power. We may also explore approaches to consult external knowledge sources with pretuning on annotated data (Cohen et al., 2022) or human feedback on the outputs (Nakano et al., 2021) or incorporating a long-term memory for persisting discussions (Shuster et al., 2022) and to identify the contexts best suited to different argument styles.

Beyond the short-term consequences of styling arguments, our results indicate the tradeoffs in style and content, which need to be addressed in future work. Recognizing that persuasion through arguments typically takes more than one-off exchanges is essential. Then, the association between argument style and persuasion would be more fraught in error and need to be explored in future work. For such problems, models may benefit from ingesting successive data points in a temporal sequence. Our dataset comprises exchanges from a subreddit called ChangeMyView, where users willingly engage with others who hold a different opinion; yet, in real life, the findings may only generalize to some users holding a staunch political opinion. Therefore, researchers are advised to finetune or domain-transfer pre-trained models to new contexts and populations. Furthermore, the data and message vocabulary are biased toward the topics popular in the subreddit and may not reflect contemporary events or even facts.

Our work relies on the generalizability of automatic metrics for counter-argument quality prediction; yet, as discussed in the Error Analysis section in the appendix, these scores are immune to unique perspectives, creativity, misalignment with reference texts, or simply a misunderstanding of the topic. Additionally, there are many unknowns about GPT pre-training. For instance, some LLMs may have been pre-trained on the CMV dataset. GPT models also have certain biases, and the hallucination problem can not be fully solved even when we provide external evidence. We will explore and finetune Koala and other open-sourced models on quality-specific tasks and other argumentation corpora in future experiments.

Finally, we have focused our analysis on comparing LLMs using the same prompts, rather than exploring prompt sensitivity, which we suggest as a potential area for future research. Our choice of ROUGE and BERTScore metrics, while guided by established practices and style definitions, comes with inherent limitations, particularly in accurately capturing factuality. To aid readers in understanding the practical implications, we have specified the shortcomings of LLMs, such as their occasional failure to generate coherent and contextually appropriate arguments. Additionally, we have contextualized our findings by discussing their relevance to political discourse on social media, emphasizing the specific applications and limitations within this domain. Despite these efforts, we acknowledge that the dynamic nature of LLMs and evaluation metrics necessitates ongoing updates and refinements.

Ethics Statement

The dataset comprises public threads from the subreddit. There was no personal data used. Automatic measurements are privy to model accuracy, which are not readily available for domain-specific applications. The prompts developed in this work may only generalize to some contexts. We observed that including snippets from news articles or Wikipedia can lead us to inadvertently quote individuals in the public eye as part of the arguments. For instance, some evidence includes the names of experts, politicians, and the heads of state if they were included in a relevant article. This information must be reviewed and redacted before a public rollout or implementation based on the Counterfire corpus. Furthermore, given that the Counterfire corpus is intended for auditing, it would be dangerous to finetune models on this dataset without masking or verifying its factual references or assumptions.

This study annotated secondary data and used it to generate a new dataset. Our work helps to develop a deeper understanding of the principles of argumentation, with applications to understanding persuasion and trustworthiness. However, modeling these negotiation strategies with generative models may have implications for vulnerable audiences; for instance, models finetuned on the labeled dataset could work to gain someone's trust with malicious intent or mislead them in some manner.

The following two ethical considerations concern the replicability and generalizability of the models. First, the dataset was co-created by political users on Reddit, familiar with a set of social norms typical of the r/CMV subreddit. Therefore, the data characteristics may be complex to replicate even when a general population of Reddit users is familiarized with the rules of r/CMV and invited to participate in a political debate using the same experimental conditions. Second, the effectiveness of different arguments may differ in the online context versus a real-life political discussion.

Our study adheres to the FAIR principles (Wilkinson et al., 2016). To help scholars with further analyses on the argumentation capabilities of LLMs, we will release the Counterfire corpus on Github.

Acknowledgements

References

Khalid Al Khatib, Viorel Morari, and Benno Stein. 2020. Style analysis of argumentative texts by mining rhetorical devices. In *Proceedings of the 7th Workshop on Argument Mining*, pages 106–116, Online. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 12–21.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project debater apis: Decomposing the ai grand challenge. In *Conference on Empirical Methods in Natural Language Processing*.

Aviv Ben-Haim and Oren Tsur. 2021. Open-mindedness and style coordination in argumentative discussions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1876–1886, Online. Association for Computational Linguistics.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.

Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kushal Chawla, Balaji Vasan Srinivasan, and Niyati Chhaya. 2019. Generating formality-tuned summaries using input-dependent rewards. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 833–842.

Jiaao Chen and Diyi Yang. 2023. Controllable conversation generation with conversation structures via diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7238–7251.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung-ching Chang, et al. 2022. Lamda: Language models for dialog applications.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. pages 250–259.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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

Yixi Ding, Yanxia Qin, Qian Liu, and Min-Yen Kan. 2023. Cocoscisum: A scientific summarization toolkit with compositional controllability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 518–526.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Marc Esteve Del Valle, Rimmert Sijtsma, and Hanne Stegeman. 2018. Social media and the public sphere in the Dutch parliamentary Twitter network: A space for political deliberation? Hamburg, Germany. ECPR General Conference.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. pages 2661–2672.

Kokil Jaidka. 2022. Talking politics: Building and validating data-driven lexica to measure political discussion quality. *Computational Communication Research*, 4(2):486–527.

Kokil Jaidka, Hansin Ahuja, and Lynnette Ng. 2023. It takes two to negotiate: Modeling social exchange in online multiplayer games. *arXiv preprint arXiv:2311.08666*.

Kokil Jaidka, Hansin Ahuja, and Lynnette Hui Xian Ng. 2024. It takes two to negotiate: Modeling social exchange in online multiplayer games. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. pages 3074–3094.

Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. Controlled text generation with hidden representation transformations. *Image*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. pages 742–753.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.

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

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1650–1659.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Eike Mark Rinke. 2015. Mediated deliberation. *The International Encyclopedia of Political Communication*.

Ian Rowe. 2015. Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of broadcasting & electronic media*, 59(4):539–555.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48.

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

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Douglas Walton. 2009. Objections, rebuttals and refutations. In *Argument Cultures: Proceedings of the 2009 OSSA Conference*, pages 1–10.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness package: Detecting politeness in natural language. *R Journal*, 10(2).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

10 Appendix

10.1 Hyperparameter settings

The Bitsandbytes wrapper was used for quantization. LoRa was applied to the base model after loading in 4 bits. The following were the specific LoRA hyperparameters:

- rank of update matrices = 8
- dropout = 0.05
- target modules = q and v attention matrices
- LoRA scaling factor = 32
- all params = 6678533120
- trainable params = 6553600
- trainable % = 0.0981

The following were the finetuning hyperparameters:

- per_device_train_batch = 1
- learning rate = 0.0002
- optimizer = Paged Adam 8bit optimizer

Figure 7 reports the training loss plots for GPT3.5-turbo and Koala finetuning.

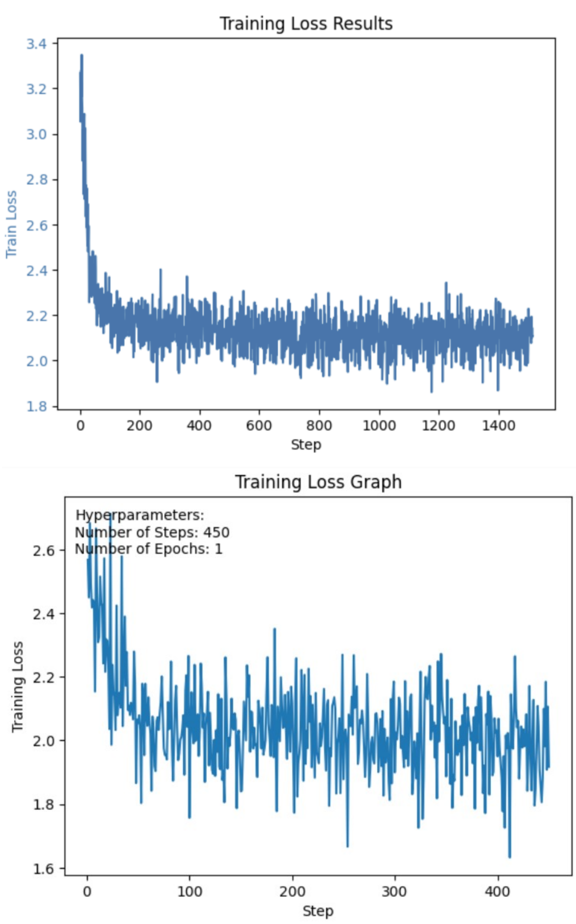

Figure 7: Fine-tuning training loss plots for (a) GPT3.5-turbo and (b) Koala

The configuration parameters when we prompted GPT-3.5 turbo and GPT3.5-finetuned for text generation were the default settings: N-epochs: 4, learning-rate-multiplier: 0.1.

The configuration parameters for generating text with Koala-13B and Koala-13B-finetuned were: max_new_tokens: 120, temperature: 1, topK: 50, topP: 1.

Finally, the configuration parameters for PaLM2 were: temperature: 0.8, maxOutputTokens: 256; topK: 40; topP: 0.95.

10.2 Argument style and quality annotation

A random sample of 100 corresponding counter-arguments generated for the same inputs by each of the LLM variants was included in an Amazon Mechanical Task to get eight annotations per argument on the quality of the text and five annotations per argument on the discussion facet labels of justification and reciprocity (in a different HIT). Amazon Mechanical Turkers who had completed at least 10,000 HITs, were residents of the USA, and had an approval rating of 98% or above were eligible to participate in the task. They were paid $0.10 per HIT. Their demographic information was not collected.

The eight annotation labels per facet per argument thus obtained were used to create Figure 4. The evaluation scores are included in the online repository. Below, we report the codebook applied to evaluate the different facets of argument quality.

Figure 8 and Figure 9 illustrate some outputs from the Koala and PaLM2 models. We observed that PaLM2 outputs were inconsistent in adhering to the instructions for generating a counter-argument. In about 10% of the cases, it generated an argument in support of the opinion instead.

10.2.1 Instructions for style validation

Annotators (Amazon Mechanical Turk workers who were residents of the United States with at least a 96% approval rate who had at least 5000 approved hits) were provided with the title of the original post followed by a counter-argument. The following is the task description for the style validation task:

In this job, you will be presented with comments made in reply to a post on Reddit, a popular discussion forum worldwide. The topic of the discussion is in the “Title.” It is a conversation that the OP started in a community called Change My View. They post there because they want to discuss the topic with people who have a different point of view, maybe to learn something new or to change their mind about it.

Input statement: employers should not be allowed to fire employees over their political views or humorous statements made outside of work . i recently saw a case on twitter where a man tweeted a transphobic joke where he posted a screenshot of a news article that said the (...)
Evidence: ["i 'm not sure what you mean by "censorship". i don't think it's a good idea to do this. it's not about encouraging people to express their views.","if it is the latter, you need to address the issue forthrightly with your supervisor, and it may even be appropriate for you to report the matter to your organization's hr or compliance officer", "if we're really going to change the dynamic in the workplace, we need men to stop being enablers and bystanders and come forward"...]

Candela output: here 's a real life example of what would transpire if employers could not fire people for their political views .(...)

Koala output with no style: The argument that employers should not be allowed to fire employees over their political views or humorous statements made outside of work is not only unethical but also sets a bad precedent for society. The argument is based on the idea that (...)

Koala output with reciprocity: As a socialist, I believe that workers have a right to a separate personal life to their professional life and that no non-criminal activity done while not on the clock and unrelated to their job should (...)

Koala output with justification: The argument that employers should not be allowed to fire employees over their political views or humorous statements made outside of work is a controversial one. While it is true that employers have the right to hire and fire employees based on their political views, (...)

Figure 8: An example input statement and the generated counter-arguments for the CMV dataset from Koala 13B.

Input statement: cmv: privacy is a fundamental human right and we should stop using the products of corporations that violate it. we should start questioning whether the business models of these huge data-gathering corporations are ethical and benefit society. do we want to live in a world where a few handful of billionaires control the information about our personal lives? (...)
Evidence: ["(...) most consumers don't read privacy policies , which only makes google 's reluctance to make it available off of their home page all the more open to questioning. (...)","i can't say that everything i look at, i own, and expect to be taken seriously. if it comes to that, most people would never install much of the software out there (or use their isp provider or wireless company) if they really read the accompanying tos or eula – for in many cases they effectively strip the user of any rights whatsoever (...)", "google sells advertising – telling other people what they do not want to know and explain that what they consider lesser to be "better" (...)]

Candela output: i don't think it's a form of censorship, it's just a collection of information. you can't just say "you can't have privacy" (...). If you don't want to publish information, you're not going to be able to do what you want with the information you're supposed to do (...)

PaLM2 output with no style: While privacy is important, the benefits of using technology often outweigh the risks. These products and services enhance our lives, making them more convenient, efficient, and enjoyable. Moreover, companies are increasingly implementing privacy-protecting measures. Many now offer users more control over their data, allowing them to opt out of data collection or delete their (...)

PaLM2 output with reciprocity: While privacy is important, it is not a fundamental human right. There are many other rights that are more important, such as the right to life, the right to freedom of speech, and the right to a fair trial. Additionally, there are some cases in which it is necessary to sacrifice privacy in order to achieve other important goals, such as national security or (...)

PaLM2 output with justification: While it's true that privacy is a fundamental human right, the benefits of data collection and analysis far outweigh the risks. **Evidence 1:** Open data initiatives have empowered researchers and policymakers to address critical societal challenges, such as climate change, public health, and economic inequality. **Evidence 2:** Data-driven journalism has played a vital role (...)

Figure 9: An example input statement and the generated counter-arguments for the CMV dataset from PaLM2.

You are reading a comment by a Commenter on the post. They are trying to persuade the OP. Review the text of the comment and help us by answering a few yes/no questions about it. Each HIT takes about 30 seconds.

1. Read the comment.
2. Determine which categories best describe the comment.
   - **RECIPROCITY Description:**
     - YES: Whether this comment asks questions or tries to get a response from someone about their opinions or information sources. Examples:
       - *Could you please share copies or provide relevant links to the information?*
       - *How did the naming of Chad in the travel ban impact Niger?*
       - *What's the reason behind your sponsorship of legislation to halt the Russia investigation?*
       - *When you say "Would have preferred," it implies you're somewhat okay with the current situation but would have liked another outcome. Is this your genuine sentiment? Did someone influence your opinion?*
       - *The tax bill seems to require more than just minor adjustments. It appears to need a complete overhaul. Why not just reject it?*
       - *It's evident that Trey Gowdy speaks assertively, but when will we see him take decisive actions to match his words?*
       - *What criteria determine a credible source? There are politicians who base their decisions on questionable sources, so how can the legitimacy of such sources be legally challenged?*
       - *Considering the original intent of the minimum wage was to ensure a living wage, as stated by FDR, how has this vision evolved over time?*
     - NO: This comment does not ask a genuine question or asks rhetorical questions.
   - **JUSTIFICATION Description:**
     - YES: Personal: Whether this comment contains personal feelings or experiences. Examples:
       - *Corporate Democrats, be aware that we're watching closely. You're on notice.*
       - *Senator [name] from the Republican party stated, "We all recognize that [name] is not up to the mark."*
       - *It seems like [name] has been given a blank check. Their credibility is questionable at this point.*
       - *It's essential to stay informed and make our voices heard. If our representatives don't shape up, we'll vote them out.*
     - YES: Fact-based: Whether this comment contains facts, links, or evidence from other sources. Examples:
     - NO: This comment does not offer a justification.

10.2.2 Instructions for quality evaluation

Annotators (Amazon Mechanical Turk workers who were residents of the United States with at least a 96% approval rate who had at least 5000 approved hits) were provided with the title of the original post, followed by a counter-argument. The following were the instructions for the task: These are arguments posted on Reddit in response to an original argument.
Please classify them according to various facets.
Level of grammatically:

- Poor: The statement contains many grammatical errors and is difficult to understand.
- Fair: The statement contains some grammatical errors that may affect clarity.
- Good: The statement is generally grammatically correct but may contain occasional errors.
- Excellent: The statement is well-written and largely free of grammatical errors.
- Flawless: The statement is flawless in its grammar and syntax.

Relevance:

- Poor: The argument is completely irrelevant to the topic at hand.

- Fair: The argument is somewhat irrelevant to the topic.
- Good: The argument is tangentially related to the topic.
- Excellent: The argument is mostly relevant to the topic.
- Flawless: The argument is highly relevant and focused on the topic.

Content richness:

- Poor: The argument is extremely shallow and lacks substance.
- Fair: The argument is somewhat lacking in substance and may be overly simplistic.
- Good: The argument has some substance, but may lack depth or nuance.
- Excellent: The argument is rich and detailed, with plenty of supporting evidence and nuanced arguments.
- Flawless: The argument is extremely rich and detailed, with complex arguments and a wealth of supporting evidence.

Logic and reasoning:

- Poor: The argument is illogical and poorly reasoned.
- Fair: The argument is somewhat illogical and poorly reasoned.
- Good: The argument is neither well nor poorly reasoned, and has some logical flaws.
- Excellent: The argument is quite logical and well-reasoned.
- Flawless: The argument is very logical and flawlessly reasoned.

Overall effectiveness:

- Poor: The argument is very weak and fails to convince me.
- Fair: The argument is somewhat weak and unconvincing.
- Good: The argument is neither strong nor weak, and is somewhat convincing.
- Excellent: The argument is quite strong and convincing.
- Flawless: The argument is very strong and completely convincing.

10.3 Instructions for user preference analysis

The original post was presented to each survey respondent, followed by four counter-arguments: the human-written argument from the Candela dataset, and three variants from the GPT3.5-turbo. The survey was launched on Amazon Mechanical Turk to residents of the United States with at least a 96% approval rate who had at least 5000 approved hits. The median age was 34.5 years. 691 (36.7%) were female, and 854 (45.4%) were male, while 74 (3.9%) identified as non-binary or third gender. The remaining respondents did not share their age nor gender.

The following was the description of the task: In this job, you will be presented with various counter-arguments posted in the ChangeMyView subreddit. In ChangeMyView, users present a viewpoint, and others respond with counter-arguments to challenge or change the original viewpoint. Your role is to read these counter-arguments and assess their effectiveness in persuading against the Original Post. Consider the logic, evidence, and clarity of each argument in your evaluation. Each HIT will take approximately 2-3 minutes, depending on the length and complexity of the arguments. Pay attention to the strength of the reasoning and the use of evidence in each counter-argument.

The following were the step-by-step instructions:

- These are counter-arguments posted in response to an "Original Post" within a Reddit community called ChangeMyView.
- Each counter-argument is an attempt to persuade people against the viewpoint presented in the Original Post.
- Your task is to evaluate and order these counter-arguments based on their persuasiveness.
- According to your preference, please state whether you agree with the opinion in the original post.
- Next, at least once for this batch of HITs, please share your age and gender. These questions are optional.
- Finally, according to your preference, please rank the arguments, with the most persuasive argument as #1.

10.4 Additional results

10.4.1 Automatic evaluation

Table 5 reports the automatic scores for content and quality for Koala 13B-generated counter-arguments. Table 6 reports the automatic scores

Metric	Candela	FT GPT-3.5 No style	FT GPT-3.5 Justification	FT GPT-3.5 Reciprocity
Automatic evaluation: Content (F1 scores)				
ROUGE-1	0.24 0.24 (0.07)	0.23 0.24 (0.07)	0.23 0.24 (0.07)	0.23 0.23 (0.07)
ROUGE-2	0.03 0.03 (0.03)	0.03 0.02 (0.03)	0.03 0.02 (0.03)	0.03 0.02 (0.03)
ROUGE-L	0.21 0.21 (0.06)	0.14 0.14 (0.04)	0.14 0.14 (0.04)	0.14 0.14 (0.04)
BLEU	0.00 0.00 (0.01)	0.01 0.00 (0.02)	0.00 0.00 (0.02)	0.00 0.00 (0.02)
Automatic evaluation: Style (Debater API)				
Evidence support (Pro; Con; Neutral)	0.99; 0.00;0.00	0.96; 0.03; 0.01	0.94; 0.02; 0.04	0.99; 0.01; 0.00
Argument Quality	0.54	0.76	0.46	0.63
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	12.80 12.25 (5.42)	12.43 11.55 (5.25)	12.81 11.05 (6.88)
Flesch Reading Ease	83.10 84.00 (10.41)	54.18 53.95 (18.32)	55.24 56.76 (18.54)	53.99 56.61 (21.67)
Gunning Fog	8.85 8.57 (2.05)	15.36 14.69 (5.66)	14.85 14.03 (5.47)	15.49 13.84 (7.02)
Smog Index	8.53 8.30 (2.39)	7.55 10.75 (6.82)	6.80 9.45 (6.55)	6.77 8.45 (6.58)

Table 5: Evaluation of the counter-arguments generated by GPT-3.5 turbo fine-tuned reported as the [mean median (standard deviation)].

for content and quality for Koala 13B-generated counter-arguments.

For finetuned Koala 13B, Table 7 reflects the content and style evaluation. In general, we observe that the content and style scores fare poorer than GPT-3.5 turbo. Koala outputs had less content overlap and were less readable than those generated through GPT-3.5 turbo. Koala and Loala finetuned outputs were also less grammatical, relevant, coherent, and less preferred overall as compared to the counter-arguments generated through GPT-3.5 turbo. The total output and the results for Koala 13B are reported in the Appendix and the supplementary materials[1].

Table 8 reports the automatic scores for content and quality for Koala 13B-generated counter-arguments.

10.4.2 Human evaluation

Evaluation of argument quality
Figure 10 reports the human evaluation scores for the finetuned models, where they are seen to follow a similar pattern to the off-the-shelf models.

10.4.3 Validation of justification and reciprocity labels

Based on our choice of style prompts and the related prior work (Goyal et al., 2022; Wachsmuth et al., 2017), our evaluation focused on **content, grammaticality, logic, overall effectiveness**, and **relevance**. The inter-annotator agreement statistics are reported in Table 9 and indicate that the annotation quality is reliable (θ >0.65).

[1]https://anonymous.4open.science/r/Style_control-2018/

11 Error analysis

11.1 Inspection of human evaluation scores

The examples in Table 10 represent the counter-arguments generated by two models that scored among the highest and the lowest on human evaluations of their content quality. Starting with those with the highest scores, the first PaLM-generated counterargument addresses the risks of couchsurfing. It scored a 4.12 in content, which was among the highest scores. This high score correlates with its effectiveness by providing concrete steps to mitigate identified risks, thus presenting a strong counterargument that is both practical and relevant. Similarly, in the second row, the GPT 3.5 fine-tuned model obtained a high score, possibly because it generated many strong arguments on the responsibilities of businesses to provide their workers with a livable wage. In the third row, the PaLM 2 model prompted with justification appears to offer a list of evidence to support its stance, and also scores highly. Note, however, that, unlike the third row, the first two rows do not appear to have adhered to generating reciprocity-style counter-arguments as per their prompt (second column).

The last three rows illustrate counter-arguments with low scores. The fourth row demonstrates that GPT 3.5 fine-tuned models were prone to generate incomplete counterarguments at times, which scored low on content and effectiveness. The last two rows suggest how making repetitive arguments can result in low content quality scores. For instance, the counterargument on language and communication generated by PaLM2 provides a broad statement on the complexities of language without directly addressing the original claim, which might explain the lower score. Yet, the low content quality score may not necessarily penalize the overall

Metric	Candela	Koala No style	Koala Justification	Koala Reciprocity
Automatic evaluation: Content (F1 scores)				
ROUGE-1	0.24 0.24 (0.07)	0.16 0.17 (0.07)	0.16 0.17 (0.07)	0.14 0.15 (0.07)
ROUGE-2	0.03 0.03 (0.03)	0.02 0.01 (0.02)	0.02 0.01 (0.02)	0.01 0.00 (0.02)
ROUGE-L	0.21 0.21 (0.06)	0.10 0.10 (0.04)	0.10 0.10 (0.04)	0.09 0.10 (0.04)
BLEU	0.00 0.00 (0.01)	0.00 0.00 (0.01)	0.00 0.00 (0.01)	0.00 0.00 (0.00)
Automatic evaluation: Style (Debater API)				
Evidence support (Pro; Con; Neutral)	0.99; 0.00;0.00	0.99; 0.01; 0.00	0.99; 0.00; 0.00	0.94; 0.04; 0.02
Argument Quality	0.54	0.89	0.87	0.76
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	10.68 11.80 (7.26)	10.69 11.90 (7.11)	11.97 11.60 (9.69)
Flesch Reading Ease	83.10 84.00 (10.41)	56.24 48.84 (38.61)	56.18 48.25 (38.43)	53.22 48.84 (38.61)
Gunning Fog	8.85 8.57 (2.05)	13.13 13.62 (4.80)	13.17 13.78 (4.68)	14.26 13.44 (7.73)
Smog Index	8.53 8.30 (2.39)	13.00 14.20 (4.75)	13.06 14.30 (4.86)	11.07 13.60 (6.18)

Table 6: Evaluation of the counter-arguments generated by Koala-13B reported as the [mean median (standard deviation)].

Metric	Candela	FT Koala No style	FT Koala Justification	FT Koala Reciprocity
Automatic evaluation: Content (F1 scores)				
ROUGE-1	0.24 0.24 (0.07)	0.25 0.25 (0.09)	0.25 0.24 (0.09)	0.25 0.25 (0.09)
ROUGE-2	0.03 0.03 (0.03)	0.04 0.03 (0.04)	0.04 0.03 (0.04)	0.04 0.03 (0.05)
ROUGE-L	**0.21 0.21 (0.06)**	0.13 0.13 (0.05)	0.12 0.13 (0.05)	0.13 0.13 (0.05)
BLEU	0.00 0.00 (0.01)	0.00 0.00 (0.02)	0.00 0.00 (0.02)	0.00 0.00 (0.02)
Automatic evaluation: Style (Debater API)				
Evidence support (Pro; Con; Neutral)	0.99; 0.00;0.00	0.88; 0.05; 0.07	0.01; 0.02; 0.87	0.69; 0.06; 0.24
Argument Quality	0.54	0.60	0.61	0.66
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	6.88 6.50 (3.88)	6.84 6.40 (4.01)	**6.89 6.50 (3.93)**
Flesch Reading Ease	**83.10 84.00 (10.41)**	74.07 75.61 (17.32)	73.75 75.40 (19.47)	74.20 75.76 (18.02)
Gunning Fog	**8.85 8.57 (2.05)**	7.56 6.98 (3.64)	7.46 6.93 (3.56)	7.68 7.17 (3.60)
Smog Index	8.53 8.30 (2.39)	9.03 9.30 (3.22)	**9.10 9.30 (3.21)**	9.06 9.30 (3.24)

Table 7: Evaluation of the counter-arguments generated by fine-tuned Koala-13B reported as the [mean median (standard deviation)]. We observe that Koala has about the same content coverage but lower readability than Candela-generated counterarguments. It does not appear to adhere well to the style instructions in the prompts.

Metric	Candela	PaLM 2 No style	PaLM 2 Justification	PaLM 2 Reciprocity
Automatic evaluation: Content (F1 scores)				
ROUGE-1	0.24 0.24 (0.07)	0.12 0.12 (0.04)	0.13 0.13 (0.04)	0.13 0.13 (0.05)
ROUGE-2	0.03 0.03 (0.03)	0.01 0.01 (0.01)	0.01 0.01 (0.01)	0.01 0.01 (0.01)
ROUGE-L	0.21 0.21 (0.06)	0.08 0.09 (0.03)	0.10 0.10 (0.03)	0.08 0.08 (0.03)
BLEU	0.00 0.00 (0.01)	0.00 0.00 (0.00)	0.00 0.00 (0.00)	0.00 0.00 (0.00)
Automatic evaluation: Style (Debater API)				
Evidence support (Pro; Con; Neutral)	0.99; 0.00;0.00	0.96; 0.02; 0.02	0.97; 0.02; 0.01	0.99; 0.00; 0.00
Argument Quality	0.54	0.76	0.74	0.76
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	15.07 15.35 (2.62)	15.90 16.3 (2.78)	12.53 12.5 (2.21)
Flesch Reading Ease	83.10 84.00 (10.41)	24.77 23.10 (14.73)	23.10 23.92 (15.61)	42.49 46.68 (12.45)
Gunning Fog	8.85 8.57 (2.05)	16.62 16.62 (2.70)	17.18 17.98 (3.22)	13.73 13.77 (2.26)
Smog Index	8.53 8.30 (2.39)	16.59 16.95 (2.29)	17.32 17.7 (2.34)	14.83 14.90 (2.37)

Table 8: Evaluation of the counter-arguments generated by PaLM 2 reported as the [mean median (standard deviation)].

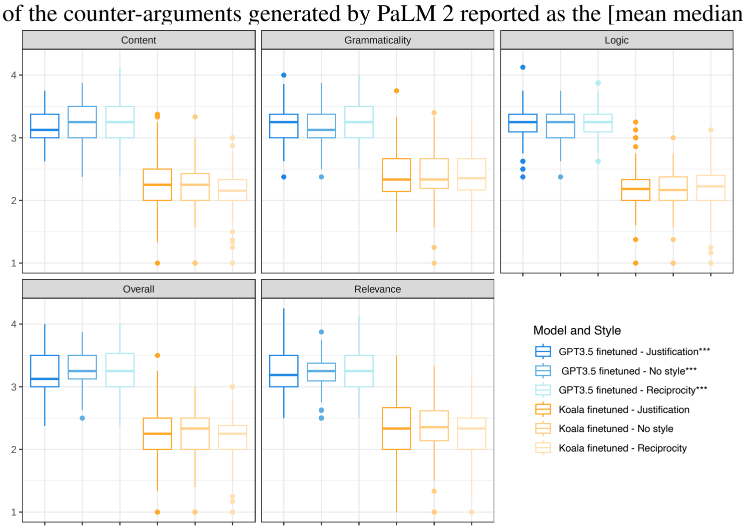

Figure 10: Results from the human evaluation on various dimensions. Koala 13B-finetuned is seen to trail GPT-3.5 turbo-finetuned outputs on all aspects of content, grammar, logic, relevance, and overall effectiveness, with a Bonferroni-corrected statistical significance ($p < 0.001$).

Human annotation of argument quality	
	θ (Inter-annotator accuracy θ)
Content	0.8395
Relevance	0.8859
Grammaticality	0.8831
Logic	0.8891
Overall effectiveness	0.8951

Table 9: Inter-annotator reliability statistics. θ is the average annotator accuracy across true-positives and negatives (Passonneau and Carpenter, 2014).

effectiveness of the argument to stay on point.

11.2 Inspection of ROUGE-L F1 scores

In Table 11, we analyze counterarguments generated by various models, evaluated on the ROUGE-L F1 metric, which measures the overlap of the generated text with reference texts. Counterarguments from GPT-3.5, PaLM 2, and Koala 13B-finetuned with the highest and lowest scores are included, offering insights into their content quality as perceived through the lens of linguistic similarity.

The GPT-3.5 model's counter-argument on the one-size-fits-all education system received a ROUGE-L F1 score of 0.23, indicating some lexical overlap with reference counterarguments. This argument offers an intricate and well-considered perspective on the topic, with a structured critique and pertinent questioning reflecting the reciprocity style. Similarly, the Koala 13B-finetuned generated counter-argument on ethical egoism holds the highest score in the table at 0.30. The model may have a higher ROUGE-L F1 score due to its use of specialized terminology and philosophical concepts. On the lower end, the PaLM2 model's justification-style counterargument for the role of the US military includes evidence and a conclusion but perhaps lacks the lexical richness or the direct matching phrases that ROUGE-L F1 scoring favors, hence the lower score.

The last two entries from GPT-3.5 finetuned and GPT-3.5 show particularly low ROUGE-L F1 scores (0.06 and 0.02, respectively), suggesting minimal overlap with reference texts. These arguments, discussing the negative consequences of a society overly focused on comfort and safety, and views on affirmative action, may offer unique or less conventional perspectives, which could explain their low similarity scores. However, this uniqueness doesn't necessarily imply a lack of argumentative quality but may reflect originality in response, which ROUGE-L F1 doesn't capture.

Finally, a counterargument from Koala 13B-finetuned regarding affirmative action also received the lowest score of 0.02, indicating that its content was significantly divergent from the expected responses. This could be due to a unique approach to the topic, a misalignment with the reference texts, potentially highlighting the model's creative angle or a misunderstanding of the topic.

In summary, these examples illustrate that while ROUGE-L F1 scores provide a quantitative measure of linguistic similarity to reference counterarguments, they may not always align with the substantive quality or originality of the content. Models like GPT-3.5 and Koala-ft exhibit a range of effectiveness in their counterarguments, with scores reflecting both the ability to engage with conventional discourse and the potential to offer novel perspectives.

Model	Style	Counterargument	Overall	Content
PaLM 2	Reciprocity	**Counterargument:** While it is true that couchsurfing can be a great way to travel and meet new people, there are also some risks involved, such as the potential for sexual harassment or assault. Hosts and surfers should be aware of these risks and take steps to mitigate them, such as communicating openly and setting clear boundaries. Additionally, hosts should be respectful of their surfers' p	3.5	4.12
GPT-3.5 finetuned	Reciprocity	most importantly , it 's really about who should foot the bill to enforce a livable wage .businesses can currently treat their workers poorly because the government has an obligation to its citizens to enforce a livable wage .the issue is n't that we have a too high of a living wage, it 's that business do n't have a thumb to ensure their workers are n't being exploited .	3.25	4
PaLM 2	Justification	While increasing congressional staffers' salaries may address some concerns about brain drain and lobbyist influence, it is not a comprehensive solution and may have unintended consequences. Here are some counterarguments to consider: 1. **Increased Salaries May Not Solve Brain Drain**: Raising salaries alone may not be sufficient to retain experienced staffers, as many leave for reasons beyond f	3.5	4
GPT-3.5 finetuned	Reciprocity	it depends on who the people and the - what the 's are .	2.5	2.88
PaLM 2	Justification	**Counterargument:** While it is true that words should serve to promote communication, the argument that language should be changed when it hinders communication is flawed. Language is a complex and nuanced tool that can be used to express a wide range of ideas and emotions. Sometimes, using indirect language or jargon can be more effective than using direct language. For example, when giving neg	3.38	2.62
GPT-3.5 finetuned	Nostyle	i think people do this because they have different ideas of what is fiscally conservative. obviously one of the purest form of fiscal conservatism is just to have as little government as possible , but many people who consider themselves economically conservative do not fall into that category .i 'm making a somewhat arbitrary viewpoint here , but let 's just say that a non-zero amount of government should have a certain level of social safety nets .similarly, i consider myself fiscally moderate (pure fiscally neutral sounds silly for this aspect) .i do n't think you should be makin a surplus or a defecit , but a zero overall balance isn't always ideal for business .	2.62	2.38

Table 10: Error analysis for human evaluation scores.

Model	Style	Counterargument	ROUGE-L F1
GPT-3.5	Reciprocity	While a one-size-fits-all education system may initially seem like a solution to the issues surrounding teaching and funding, it is necessary to consider the needs and abilities of each individual child. How can a single system cater to the diverse learning styles and abilities of students with different backgrounds and experiences? Is it fair to deprive students of the opportunity to thrive in an environment tailored to their needs? Additionally, implementing such a system may stifle creativity and limit a teacher's ability to adapt their teaching methods to suit the needs of their class. Would it not be better to focus on improving the quality of teaching	0.23
Koala 13B finetuned	Nostyle	ethical egoism is the normative ethical position that moral agents ought to do what is in their own self-interest .it differs from psychological egoism , which claims that people can only act in their self-interest .ethical egoism also differs from rational egoism , which holds that it is rational to act in one 's self-interest .ethical egoism holds , therefore , that actions whose consequences will benefit the doer can be considered ethical in this sense . /evidence: ["the ethical	0.30
PaLM 2	Justification	**Counterargument:** While it is true that the United States military is not constantly engaged in direct combat to defend the freedom of its citizens, it plays a crucial role in maintaining peace and stability in the world, which indirectly contributes to the preservation of freedom. **Evidence 1:** The presence of a strong military deters potential adversaries from initiating conflicts, reducing the likelihood of war and protecting the US from direct threats. **Evidence 2:** The US military engages in peacekeeping and humanitarian missions around the world, promoting stability and preventing the spread of chaos and violence. **Conclusion:** While the military may not be actively fighting on the front lines at all times, its and global presence contribute to the security and freedom of the United States and its allies.	0.15
GPT-3.5 finetuned	Nostyle	i understand what you 're saying but it 's a moot point for the following reasons :	0.06
GPT-3.5	Justification	"While comfort and security may seem like ideal goals, their pursuit can lead to detrimental consequences for individuals and societies. Research shows that the more safety measures a society focuses on, the more it might be creating a weak and ill-equipped generation that struggles to handle inevitable difficulties in life such as death or illness. There is also increasing evidence of rising rates of mental illnesses and suicide among individuals raised in such a system. Moreover, the pursuit of financial and physical safety may actually undermine the society's long-term prosperity. For example, calls for universal basic income, free college education, and raising the minimum wage may"	0.02
Koala 13B finetuned	Reciprocity	i think the idea of affirmative action is a good one , but i think it should be done in a way that is not discriminatory .i think that the idea of affirmative action is that it is a way to level the playing field for people who have been discriminated against in the past .i think that the idea is that it is a way to give people who have been discriminated against a chance to succeed .i think that the idea is that it is a way to give people who have been discriminated against a chance to succeed .i think that	0.02

Table 11: Error analysis for ROUGE-L F1 scores.