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

Anonymous ACL submission

Abstract

We audited counter-arguments generated by large language models (LLMs), focusing on their ability to generate evidence-based and stylistic counter-arguments to posts from the Reddit ChangeMyView dataset. Our evaluation is based on Counterfire: a new dataset of 32,000 counter-arguments generated from large language models (LLMs): GPT-3.5 Turbo and Koala and their fine-tuned variants, and PaLM 2, with varying prompts for evidence use and argumentative style. GPT-3.5 Turbo ranked highest in argument quality with strong paraphrasing and style adherence, particularly in 'reciprocity' style arguments. However, the 'No Style' counter-arguments proved most persuasive on average. The findings suggest that a balance between evidentiality and stylistic elements is vital to a compelling counter-argument. We close with a discussion of future research directions and implications for fine-tuning LLMs.

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.

Prior work in counter-argument generation by Bilu et al. (2015) and Hidey and McKeown (2019) focused on generating contrastive claims, with the former blending rule-based techniques and the latter leveraging data-driven strategies, while Alshomary et al. (2021) focused on undermining the weakest claim. The Project Debater system (Bar-Haim et al., 2021; Slonim et al., 2021) engages in competitive debates and is centered on an argument mining framework that retrieves data from a corpus of about 400 million articles. On the other hand, Hua et al. (2019) and Jo et al. (2021) focused on incorporating evidence in counter-arguments. Following the call for controllable composition in other spheres of natural language generation (Chen and Yang, 2023; Kumar et al., 2023), most notably, scientific summarization (Ding et al., 2023), we also argue that for a well-rounded and compelling argument, the controlled generation of counter-arguments, customized to user-specified preferences of evidence and style, can further enhance the effectiveness of counter-arguments. ~~Accordingly, we introduce and compare LLMs on the~~

This study outlines its primary objective to audit the controlled counter-argument capabilities of three LLMs- GPT-3.5, PaLM2, and Koala, and two of their fine-tuned variants, which are prompted with evidence and style instructions. We evaluate their adherence to the instructions, ability to integrate evidence and style, and the implications for human preferences. We build the first dataset involving evidence and style as critical attributes for controlled counter-argument generation in the political domain. Our dataset ~~comprises high-quality counter-arguments~~ is annotated with human- and automatic evaluation metrics. ~~Aside from the new dataset, the Counterfire corpus, we make~~ We also offer two key contributions to the counter-argument generation literature:

- A new style dimension for counter-arguments to control their ~~intertextuality and engagement quality~~justification through intertextuality with supplied counter-evidence, and reciprocity through engaging with and eliciting responses from the reader.
- Insights on fine-grained counter-argument structure, such as phrase-level expressions of reciprocity, justification, argument alignment, and appeals to authority.

Our framework uses facts shortlisted from the intermediate outputs of a seq2seq baseline system to manufacture domain-injected prompts. Next, we evaluate their efficacy at generating relevant, logical, and grammatical counter-arguments from off-the-shelf and finetuned LLMs. We have employed standard automatic metrics and human evaluation to measure argument style and quality along five quality dimensions. 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.

2 Background

Carefully crafted natural language instructions may be used to steer generation and control style dimensions of the generated output. However, these approaches and the resultant outputs must be evaluated for factual integration, style adherence, or user preference.

Although LLMs excel in many downstream generation tasks, counter-argument generation proves to be much more complex since convincing arguments require external information for evidence. In the past, argument generation systems based on retrieval focused on selecting relevant passages or sentences from data sources and ordering them. Hua et al. (2019) combine a retrieval system with generation by feeding retrieved passages to a seq2seq architecture in Candela. The survey by Zhang et al. (2023) examines how LLMs capture world knowledge and identifies the major explicit approaches as memory-, retrieval-, and internet-enhanced. In general, these prior approaches focus mainly on integrity issues at the entity or document level (Wachsmuth et al., 2018), employing massive retrieval models that are computationally expensive, and no previous work has looked explicitly at argument generation with the retrieved information.

2.1 LLMs for stylized text generation

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.

In our paper, applying concepts from social science for LLM prompts offers a theoretically grounded approach to better argumentation. Political communication research conceptualized social media platforms as a space for 'internal reasoned dissent' (Rinke, 2015), where social media 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 has distinguished analytical arguments from social arguments (Esteve Del Valle et al., 2018; Friess and Eilders, 2015; Jaidka, 2022; Rowe, 2015). First, the analytical aspects of arguments include the use of *constructiveness*, specifically logic and rational arguments, to move towards a consensus, and the use of *justification*, specifically tangible evidence to support claims. Second, the social aspects of arguments include the use of *reciprocity*, the interactivity of a discussion identified through whether participants invite engagement from each other. However, an examination of the actual distribution of these facets in the annotated CLAPTON corpus provided in prior work (Jaidka, 2022) suggests that at least in Reddit, authors overwhelmingly prefer to write counter-arguments that follow a Justification (30%) or a Reciprocity (25.8%) style rather than Constructiveness (6.6%), thereby motivating our focus on Justification and Reciprocity for auditing counter-argument generation in the Reddit ChangeMyView context.

No prior paper has compared three LLMs - simple and finetuned - in this manner for this task. 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. The following sections explore the methodology and findings of our study in three parts: firstly, the data collection process utilizing zero-shot prompting and finetuning; secondly, validation tasks involving fact integration and a dual approach of automatic and human evaluation; and thirdly, an in-depth analysis providing insights into the distribution of alignment moves and a user preference analysis of the generated counter-arguments.

3 Data Collection

This study audits zero-shot prompting and finetuning to collect counter-arguments to CMV (Change My View) posts. The data collection process is twofold: initially using zero-shot prompting and refining through finetuning techniques. Figure 3 illustrates our experimental framework. Our work applied the dataset curated by Hua et al. (2019) in a retrieval system to generate counter-arguments from various language models. We curated parallel corpora comprising the primary dataset of target argument and related evidence and the outputs generated from various primary and finetuned LLMs on 2000 inputs. The different generation pipelines are discussed below. Figure 1 illustrates the outputs from GPT 3.5 turbo, while the outputs from Koala 13B and PaLM2 are shown in the Appendix. 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.

Table 1: The three variants of the style specifications added to the LLM prompt.

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.

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", "the most severe cases of disease and death linked to whooping cough have been in infants under six months of age , and the source of that infection is most often an older child or adult "]

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

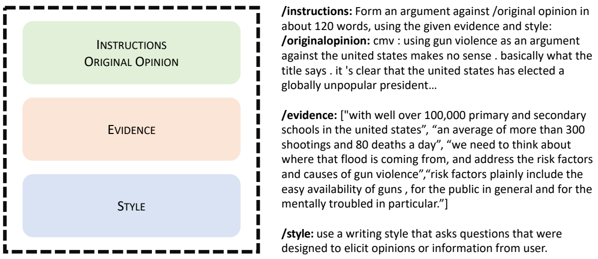

Figure 2: Example prompt for generating a reciprocal counter-argument.

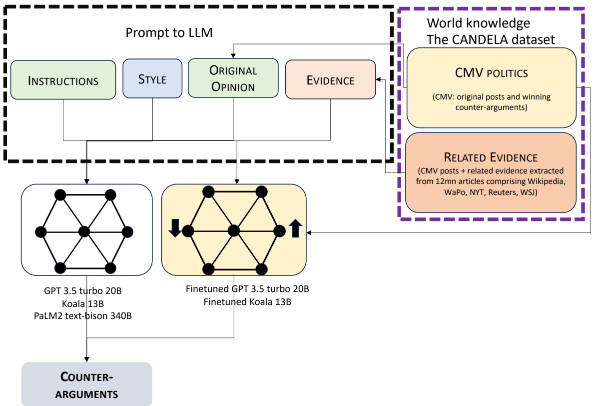

Figure 3: Experimental framework for generation.

Baseline: The primary baseline comprises the Candela dataset by Hua et al. (2019), ~~which is~~ of which we sampled our training set of (a) 70,000 ~~randomly sampled~~ English original posts and winning counter-arguments related to politics from the subreddit r/ChangeMyView, and (b) their associated evidence retrieved from a database of 12 million articles from Wikipedia, and four major

English media wires of different ideological leanings - Washington Post, New York Times, Reuters, and The Wall Street Journal - are queried. When queried using the text of a Reddit post, each input's size-constrained related passages retrieved from diverse sources are deduplicated, ranked, and returned as "evidence~~".We randomly sampled~~ ." Beyond the training set, we created an additional random sample of 2000 rows of original posts and evidence from this dataset for further analysis.

Generating stylized counter-arguments: Five off-the-shelf and finetuned LLMs were prompted three times, with the original post and the evidence from the subsampled Candela dataset. The prompts calling for different stylistic variations are based on operationalization in prior work (Steenbergen et al., 2003; Jaidka, 2022). To validate that incorporating real-world evidence was effective, we also made a set of prompts without including the curated real-world evidence.

Figure 2 reports a sample prompt to generate a reciprocal counter-argument. The last part of the prompt constitutes style instructions, and Table 1 includes the style instructions used in the experiments.

We generated counter-arguments from each of the five LLMs after providing them with 2000 (prompts with the original opinion and evidence) x 3 variants for style control (N = 32,000)[1]. We used the Candela dataset (Hua et al., 2019) ~~dataset~~ for the input and evidence used in our prompts. The evidence comprises talking points retrieved from passages in a database of 20 million articles.

We prompted three LLMs and two of their finetuned variants with these inputs (finetuned using instruction tuning on the full Candela dataset) and collected the outputs. These outputs were benchmarked against Candela counter-arguments - the pre-LLM era auto-generated counter-arguments included in the Candela dataset. The Candela counter-argument was created by applying a biLSTM encoder on the retrieved evidence, followed by two decoders in series to plan and then populate the final counter-argument. The following ~~were~~ paragraphs describe the LLMs we tested~~:~~ .

3.1 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.

3.2 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.

3.3 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. The configuration parameters for the LLMs are reported in the Appendix. Figures 1 illustrate some of the outputs from GPT-3.5 turbo. Examples from the other models are included in the Appendix. The full dataset is available in the anonymous online repository.

3.4 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

[1] Candela (2k) + Koala-13B (6k) + GPT3.5-turbo (6k) + Koala finetuned (6k) + GPT3.5-turbo finetuned (6k) + PaLM2 (6k) = 32k

precision. The training loss plot and the hyper-parameter settings are reported in the Appendix. Finetuning for PaLM2 was not performed because of errors noted in the outputs discussed in Section 5.

4 Analyses

4.1 Validation tasks

We performed three validation tasks to audit the ability of LLMs to adhere to the instructed prompts: (a) Fact integration, where the incorporation of evidence is assessed, and (b) Style validation, to gauge whether the outputs reflect the expected discussion style. Finally, we performed the (c) Quality evaluation, encompassing both automatic and human assessment, to gauge the effectiveness and coherence of the generated counter-arguments. The human assessment task was launched on Amazon Mechanical Turk, and the detailed instructions are provided in the appendix.

4.2 ~~Rhetorical insights~~

~~We performed automatic content analyses to characterize and compare the generated counter-arguments for the presence of rhetorical moves related to alignment, authority, and persuasion. Alignment moves constitute the phrases used by authors to indicate agreement with each other. Authority moves are the phrases used by authors to express their credibility. The source of the phrases was the Alignment and Authority in Wikipedia Discussions (AAWD) corpus (Bender et al., 2011) with the Counterfire corpus and the original Reddit corpus.~~

~~Next, persuasive moves comprise features such as politeness, contingency, expansion, claims, and premise that have been applied to study online persuasion and to model politeness and trustworthiness in social media posts (Danescu-Niculescu-Mizil et al., 2013b; Niculae et al., 2015) .~~

4.2 ~~Argument preference analysis~~

~~In the argument preference analysis, following the design of similar user experiments reported in prior work (Goyal et al., 2022), we surveyed Amazon Mechanical Turk workers to obtain user rankings for the best-performing counter-arguments as pitted against the human-written counter-argument. As before, the survey was launched on Amazon Mechanical Turk, and the detailed instructions are provided in the appendix. The goal was to examine patterns in whether a user would favor a justification- or a reciprocity-style counter-argument. In this manner, 10,000 counter-argument rankings were collected from 1879 respondents. Further details about the ranking task are reported in the Appendix.~~

5 ~~Results~~

4.1 ~~Fact and style integration~~

4.0.1 Fact integration

For fact integration validation, we analyzed whether our prompts effectively got the LLMs to apply the provided evidence in the generated counter-arguments in the fact integration validation task. This involved comparing the similarity and absolute overlap of evidence with the outputs from the off-the-shelf LLMs, using similarity metrics, such as BERTScore (Zhang et al., 2019) and ROUGE-1 (Lin, 2004). ~~We~~

4.0.2 Style and quality assessment

For the style validation of generated counter-arguments, we examined whether the LLMs could integrate the expected style into the outputs for style integration validation. This was done with the help of crowdsourced annotations from Amazon Mechanical Turk and through finetuning OpenAI ada models on the CLAPTON dataset (Jaidka, 2022) to automatically label the presence of justification and reciprocity in the generated outputs. Results are reported in subsection 5.1. Details of the finetuning task are reported in the Appendix.

4.1 ~~Argument quality assessment~~

~~Next, our evaluation techniques measure the content and argument quality of the counter-arguments~~ For the quality evaluation, we used automatic and human evaluation metrics:

- Automatic content quality evaluation: ~~ROUGE(1/2/L)~~ ROUGE-1 and BLEU, recognized as overlap-based metrics (Lin, 2004; Papineni et al., 2002), ~~is~~ are used to measure the quality of generated counter-arguments against the human-written counter-argument. These metrics were computed using the Pyrouge package for ROUGE and a corresponding package for BLEU. We also used the textstat package to calculate readability metrics, such as the Flesch Kincaid grade, Flesch Reading ease, the Gunning Fog index, and the Smog index. We have also added the Debater API scores (Bar-Haim et al., 2021) that score the stance of a sentence, as well as the quality (from 0 to 1).

- Manual style validation and argument quality evaluation: ~~Following~~ We evaluated the style integration of the LLMs corresponding to the prompt they were provided. Furthermore, following recent approaches for manual evaluation of argument quality (Goyal et al., 2022; Wachsmuth et al., 2017), we also crafted a human evaluation task focusing on the logic, rhetoric, and dialectic (Wachsmuth et al., 2017) of arguments with measures of Content, Grammaticality, Logic, Relevance, and Overall effectiveness. ~~The evaluation was done with the help of crowdsourced annotations from~~ 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). Results are reported in subsection 5.2 (style) and subsection 5.3 (argument quality). Further details of the annotation task and instructions are reported in the Appendix.

4.1 Rhetorical Insights

We performed automatic content analyses to characterize and compare the generated counter-arguments for 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) offers the closest approximation to other work on counter-argument analysis from previous work, as it comprises annotated phrases used in context by authors to indicate their agreement or disagreement with each other. Authority moves are the phrases used by authors to express their credibility. A phrase-level matching was performed to identify the use of the same phrase in the Counterfire corpus.

Next, persuasive moves comprise features such as politeness, contingency, expansion, claims, and premise that have been applied to study online persuasion and to model politeness and trustworthiness in social media posts (Danescu-Niculescu-Mizil et al., 2013b; Niculae et al., 2015) . Persuasive moves measured using the Python convokit toolkit also work by searching through text for the presence of various lexical features that are reflective of each of the different categories, such as politeness, which were developed and validated through supervised machine learning approaches by prior work. The results are reported in subsection 5.4.

4.2 Argument preference analysis

In the argument preference analysis, following the design of similar user experiments reported in prior work (Goyal et al., 2022), we surveyed Amazon Mechanical Turk ~~. Details of the annotation~~ workers to obtain user rankings for the best-performing counter-arguments as pitted against the human-written counter-argument. As before, the survey was launched on Amazon Mechanical Turk, and the detailed instructions are provided in the appendix. The goal was to examine patterns in whether a user would favor a justification- or a reciprocity-style counter-argument. In this manner, 10,000 counter-argument rankings were collected from 1879 respondents. The results are reported in subsection 5.5. Further details about the ranking task are reported in the Appendix.

5 Results

5.1 Fact integration

Table 2 reports the similarity between the evidence provided and the outputs generated, where the average BERTScore F1 value across the three LLMs was 0.725, and the average ROUGE-1 recall was 0.313. The findings suggest that LLMs may have been good at paraphrasing the evidence into the counter-argument yet yielded a low absolute overlap in the words used.

Table 2: BERTScore (default baseline rescaling) and ROUGE-1 scores to measure how much evidence is incorporated by the models

Fact Integration		
Model	**BERTscore (F1 value)**	**ROUGE-1 (Recall)**
GPT-3.5 turbo	0.7312	0.3556
Koala-13B	0.7271	0.3631
Palm-2	0.7175	0.3103

~~Next, we evaluated the style integration of the LLMs corresponding to the prompt they were provided. For the manual validation , we provided annotators on Amazon Mechanical Turk with the outputs and requested them to annotate each for Reciprocity and Justification on a five-point scale. The~~

5.2 Style integration

Results for the style validation are reported in the second part of Table 2~~demonstrates~~ , which demonstrate a high manual validation of the incorporation of style, with inter-annotator reliability of θ = 0.9682 for reciprocity and 0.7680 for justification, respectively. θ overcame 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). 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 $\theta >= 0.65$.

Table 3: The three variants of the style specifications added to the LLM prompt. θ is the average annotator accuracy across true-positives and negatives (Passonneau and Carpenter, 2014)

Style integration	
Style	**θ (Inter-annotator Accuracy)**
Reciprocity	0.9682
Justification	0.7680

5.3 Evaluating argument quality

Table 4 reports the content and style evaluation for GPT-3.5 turbo, where we have more content adherence, argument quality, and readability in text generated from prompts to LLMs. We observe that Debater API scores are very sensitive to argument quality differences (but we later find that the quality scores may not reflect user preferences). Conversely, GPT-3.5 turbo counter-arguments contain fewer specific details when they offer more stylistic variation. Similar tables for the other models are reported in the Appendix. However, we chose to keep the table for GPT-3.5 turbo here as we observe that GPT-3.5 turbo also outperforms Koala 13B and PaLM2 on all the parameters.

Figure 4 reports the human evaluation of quality. Each boxplot shows the median~~(the line within the box)~~, the interquartile range~~(IQR; the box itself)~~, and the range (whiskers). Dots outside the whiskers are outliers. The different colors and box styles represent various models and style prompts. First, the lowest scores on preference were reported for Candela. Among the GPT-3.5 turbo variants, the "no style" counter-argument had a higher median score for Grammaticality and Logic than even the "justification" and "reciprocity" styles, indicating it may produce more grammatically correct and logical content. However, it seems to have a broader spread in Overall effectiveness and Relevance, suggesting more variability in these aspects. Among the Koala-13B variants, there was a tight distribution in Content and Logic but a lower median in Overall effectiveness, indicating they may not perform as well as other models. Finally, the PaLM2 variants show a higher median score in Relevance but also have a broader spread in Overall effectiveness, suggesting that they consistently perform better than Koala-13B, but with some inconsistency in their effectiveness. In summary, GPT-3.5 turbo models outperformed all other outputs as they were perceived to be more grammatical, relevant, coherent, content-complete, and effective than others, controlling for style. The difference was statistically significant in paired t-tests over the 2000 generated counter-arguments after Bonferroni correction for multiple comparisons ($p < 0.001$). The findings with the finetuned variants are similar and are reported in the Appendix. Reporting the human quality evaluation for the best variant, i.e., GPT-3.5 turbo, Figure 4 illustrates that 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 ($p < 0.001$). The human evaluation results for Koala 13B and finetuned Koala 13B are reported in the Appendix.

5.4 Rhetorical insights

In Table 5, we report the distribution of argument moves across the different types of counter-argument variants. Alignment moves are examples of social acts involving agreement or refutation in argumentation. Of the exemplars of positive and negative alignment moves identified in the AAWD corpus, the Reddit counter-arguments contained 12.

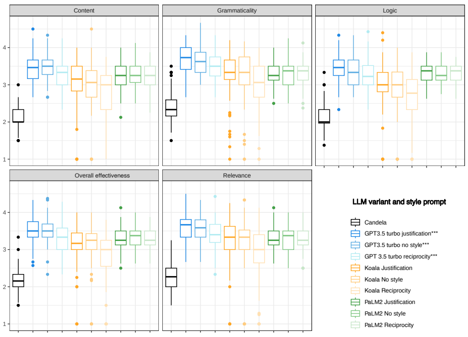

Figure 4: Results from the human evaluation on various dimensions. Candela is seen to trail GPT-3.5 turbo outputs on all aspects of content, grammar, logic, relevance, and overall effectiveness, with a Bonferroni-corrected statistical significance ($p < 0.001$). GPT 3.5 turbo also outperforms Koala 13B and PaLM2 on all the parameters. No significant differences existed between the three GPT-3.5 turbo outputs on any parameter ($p > 0.05$).

Metric	**Candela**	**GPT 3.5 turbo No style**	**GPT 3.5 turbo Justification**	**GPT 3.5 turbo Reciprocity**
Automatic evaluation: Content (F1 scores)				
ROUGE-1	0.24 0.24 (0.07)	**0.33 0.33 (0.07)**	0.17 0.17 (0.06)	0.17 0.17 (0.06)
ROUGE-2	0.03 0.03 (0.03)	**0.10 0.09 (0.06)**	0.02 0.01 (0.02)	0.01 0.01 (0.02)
ROUGE-L	0.21 0.21 (0.06)	**0.29 0.29 (0.07)**	0.15 0.15 (0.05)	0.14 0.14 (0.04)
BLEU	0.00 0.00 (0.01)	**0.06 0.06 (0.06)**	0.00 0.00 (0.01)	0.00 0.00 (0.01)
Automatic evaluation: Style (Debater API)				
Evidence support (Pro; Con; Neutral)	0.99; 0.00;0.00	0.99; 0.00; 0.00	0.99; 0.00; 0.00	0.62; 0.07; 0.30
Argument Quality	0.54	0.74	**0.81**	0.75
Automatic evaluation: Style (Accuracy)				
Reciprocity	0.17	0.09	0.12	**0.49**
Justification	**0.42**	0.26	0.24	0.22
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	**12.81 12.70 (2.07)**	12.75 12.70 (2.07)	11.79 11.60 (2.08)
Flesch Reading Ease	**83.10 84.00 (10.41)**	40.94 41.70 (11.31)	41.78 41.90 (10.62)	46.23 45.76 (11.37)
Gunning Fog	8.85 8.57 (2.05)	**15.05 14.88 (2.23)**	15.03 14.88 (2.23)	13.93 13.87 (2.17)
Smog Index	8.53 8.30 (2.39)	14.85 14.80 (1.89)	**14.87 14.80 (1.68)**	14.09 14.00 (1.72)

Table 4: Evaluation of the counter-arguments generated by GPT 3.5 reported as the [mean median (standard deviation)]. We observe greater content coverage and readability in text generated from prompts to LLMs; on the other hand, GPT 3.5 counter-arguments contain fewer specific details when they offer greater stylistic variation.

Move type	**Human-written Reddit counterargument**	**GPT3.5-turbo** No style	**GPT3.5-turbo** Reciprocity	**GPT3.5-turbo** Justification
Alignment moves				
Positive	12	0	2	4
Negative	12	0	6	4
Authority moves				
Experiential	10	0	6	0
External	10	0	2	4
Forum	10	0	4	4
Social expectations	8	0	2	0

∗ Positive types: 'other + explicit agreement', 'praise thanking + positive reference + explicit agreement', 'positive types'
∗ ∗ Negative types : 'negative types', 'doubting + explicit disagreement + dismissing'

Table 5: Total number of alignment moves identified in Counterfire outputs. Based on the AAWD corpus (Bender et al., 2011).

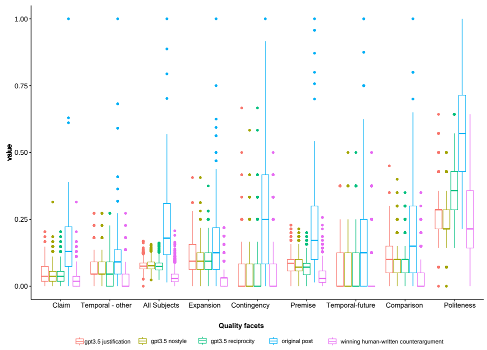

Figure 5: Results from the automatic evaluation of argumentation using the discursive and politeness features of Convokit for the arguments considered in the human evaluation.

In contrast, the GPT-3.5 turbo justification and reciprocity style counter-arguments contained 2 and 4, respectively, exemplifying explicit agreement and positive alignment, such as praise thinking, and opposing alignment, such as criticizing or doubting. On the other hand, authority moves are markers of social expectations, credentials, experiential claims, forum claims, and external claims. Certain moves in the AAWD corpus, such as 'credentials' and 'experiential,' had no counts or low counts among the variants, highlighting the domain differences compared to the AAWD corpus. The reciprocity-style counter-argument appears to have more argument moves than the no-style and justification counter-argument, perhaps because of its interpersonal nature. Finally, human-written arguments are the most argumentatively rich and diverse, with more unique moves across the different categories than the generated outputs. Similarly, in the discursive analysis reported in Figure 5, we observe

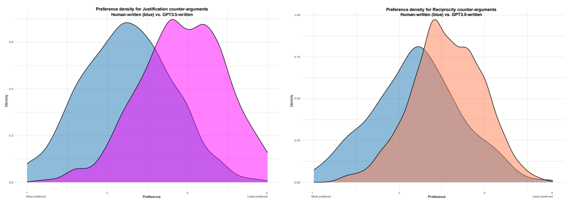

Figure 6: User preference analysis for human-written (blue) vs. GPT3.5-written counter-arguments for (a) justification and (b) reciprocity.

that the GPT3.5-written counter-arguments are typically at par with each other concerning most of the discursive features; 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 to offer a more focused and less polite counter-argument.

5.5 Argument 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. The data illustrates that in a comparison of 2000 original posts and counter-arguments sourced from ChangeMyView and the Counterfire corpus, humans still find the 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) written by other humans more preferable to those written by GPT3.5 (Means 2.93 and 2.56 respectively). This preference is statistically significant (Welch Two Sample t-Test, $p < 0.001$).

The low preference for Justification raises red flags. ~~On~~ In justification-style counterarguments, the LLMs are directed to focus on intertextuality through citing the evidence provided. On the other hand, reciprocity counterarguments were directed to focus on interactivity with the reader. On examining the outputs, we speculate that ~~the reason may be the~~ these directives resulted in an essay-style ~~structure of the arguments~~ counterargument devoid of any interpersonal engagement. Taken together with findings from Figure 5, the findings suggest that the highly-focused, specific, and less polite human counter-arguments are somehow more persuasive than GPT3.5-generated counter-arguments to humans, thereby offering food for thought in how accurately stylized text may still fall short of human expectations. The findings suggest a tradeoff between fact integration and style while generating counter-arguments.

6 Discussion and Conclusion

This study addresses the need for more research on style in political arguments and its relationship with persuasion. We also offered a new dataset for related research into finetuning, prompt structures, prompt lengths, and other novel techniques for domains that have not been extensively studied. Addressing the most pressing issues of factuality and interactive dialogic exchange currently at the forefront of LLM research (Ziems et al., 2023), we created the Counterfire corpus, focusing mainly on incorporating justification and reciprocity in the counter-arguments.

The findings underscore significant implications for generating and analyzing counter-arguments using language models. 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. ~~While overwhelmingly~~ Yet, the potential of future work lies in the preference analysis findings that highlight the limitations of directed prompt adherence compared to complex human rhetoric. While preferred to LLM outputs, human-generated counter-arguments also show more complexity and variety in argumentative tactics. 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. ~~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.~~

~~In future work, we will 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 pre-tuning 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.~~

7 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 pre-tuning 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.

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 Zenodo.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

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.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. pages 8289–8311.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? arXiv preprint arXiv:2305.03514.

8 Appendix

8.1 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.

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.

8.2 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

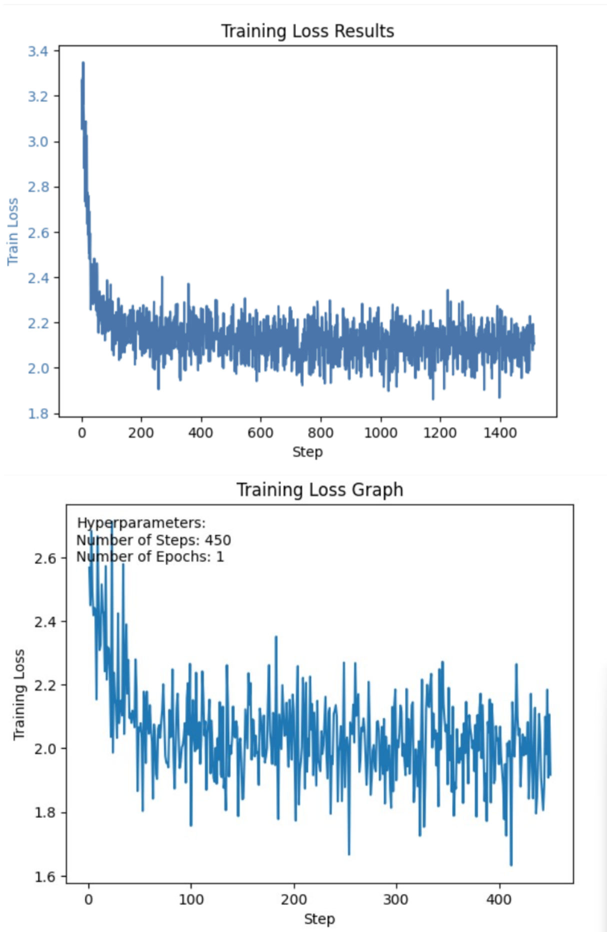

Figure 7: Fine-tuning training loss plots for (a) GPT3.5-turbo and (b) Koala

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.

8.2.1 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

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.

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.

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.

8.2.2 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.

8.2.3 ~~Instructions for user preference analysis~~

8.3 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.

8.4 Additional results

8.4.1 Automatic evaluation

Table 6 reports the automatic scores for content and quality for Koala 13B-generated counter-arguments. Table 7 reports the automatic scores for content and quality for Koala 13B-generated counter-arguments.

For finetuned Koala 13B, Table 8 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[2].

Table 9 reports the automatic scores for content and quality for Koala 13B-generated counter-arguments.

8.4.2 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.

8.4.3 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 ratings were crowdsourced through Amazon Mechanical Turk~~. The inter-annotator agreement statistics are reported in Table 10 and indicate that the annotation quality is reliable (θ >0.65).

9 Error analysis

9.1 Inspection of human evaluation scores

The examples in Table 11 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 effectiveness of the argument to stay on point.

9.2 Inspection of ROUGE-L F1 scores

In Table 12, 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

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

Table 6: Evaluation of the counter-arguments generated by GPT-3.5 turbo fine-tuned reported as the [mean median (standard deviation)].

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

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 8: 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	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 9: Evaluation of the counter-arguments generated by PaLM 2 reported as the [mean median (standard deviation)].

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)

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 mili-

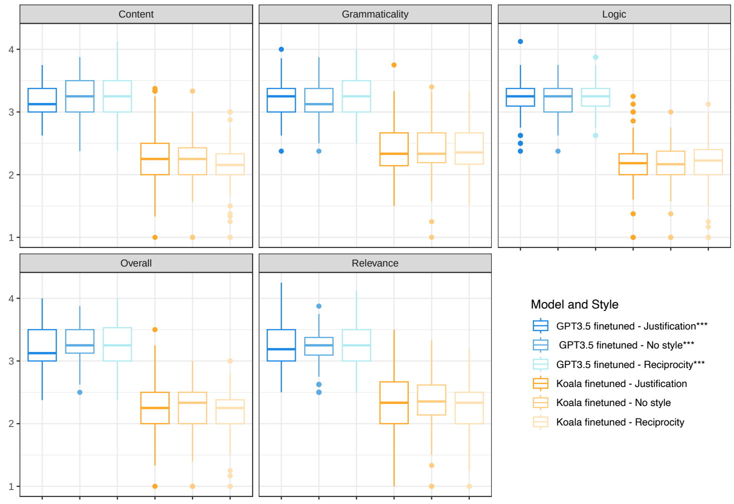

Figure 10: Results from the human evaluation on various dimensions. Koala 13B-finetuned is seen to trail GPT-3.5 turbo-finetuned outputs on all aspects of content, grammar, logic, relevance, and overall effectiveness, with a Bonferroni-corrected statistical significance ($p < 0.001$).

Table 10: Inter-annotator reliability statistics. θ is the average annotator accuracy across true-positives and negatives (Passonneau and Carpenter, 2014).

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

tary 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 counter-arguments, 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.

Table 11: Error analysis for human evaluation scores.

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 12: Error analysis for ROUGE-L F1 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