

# Auditing Counterfire: Evaluating Advanced Counterargument Generation with Evidence and Style

Anonymous ACL submission

## Abstract

We conducted an audit of counterarguments generated by large language models (LLMs), focusing on their ability to generate counterarguments with evidence and style. Our inputs comprised posts from the Reddit Change-MyView dataset, enriched with evidence sourced from high-quality references, with instructions to follow a particular debating style. We evaluated the counterarguments generated from GPT-3.5 turbo, Koala, and PaLM 2 models and two of their finetuned variants (N = 32,000) for their fact integration, style adherence, argument quality and overall persuasiveness. Model evaluation indicates strong paraphrasing abilities with evidence, albeit limited word overlap, while demonstrating high style integration (0.9682 for ‘reciprocity’), showing the ability of LLM to assimilate diverse styles. Of all models, GPT-3.5 turbo showed the highest scores in argument quality evaluation, showing consistent accuracy (score >0.8). In further analyses, reciprocity-style counterarguments display higher counts in most categories, possibly indicating a more creatively persuasive use of evidence. In contrast, human-written counterarguments exhibited greater argumentative richness and diversity across categories. While GPT-3.5-written ‘justification’ arguments were judged as the highest quality, the ‘No Style’ counterarguments were considered the most persuasive and second to human-written ‘justification’ counterarguments, suggesting the need to investigate trade-offs in generation for facts and style.

## 1 Introduction

Counterargument 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 counterarguments grounded in evidence is a critical aspect of natural language processing, with

applications in fields such as 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 which 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 effective argument, the controlled generation of counter-arguments, customized to user-specified preferences of evidence and style, can further enhance the contextual effectiveness of counter-arguments. Accordingly, we introduce and compare LLMs on the first dataset involving evidence and style as key attributes for controlled counter-argument generation in the political domain. Our dataset comprises high-quality counter-arguments with human- and automatic evaluation metrics. Aside from the new dataset, the Counterfire corpus, we make two key contributions to the counter-argument generation literature:

- A new style dimension for counter-arguments to control their intertextuality and engagement quality.
- Insights on fine-grained counter-argument structure, such as phrase-level expressions of reciprocity, justification, 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 fine-tuned 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 counterarguments, 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 - yet, these approaches and the resultant outputs are yet to 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 identify 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, 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 generation 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 counterarguments 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 counterargument generation in the Reddit ChangeMyView context.

To our knowledge, no prior paper has compared three LLMs - simple and fine-tuned - 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, fine-tuning, 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 fine-tuning; 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 counterarguments.

### 3 Data Collection

This study audits zero-shot prompting and fine-tuning to collect counter-arguments to CMV (Change My View) posts. The data collection process is twofold: initially using zero-shot prompting and subsequently refining through fine-tuning 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 fine-tuned 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.

**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.

**Baseline:** The primary baseline comprises the Candela dataset by Hua et al. (2019), which is (a)

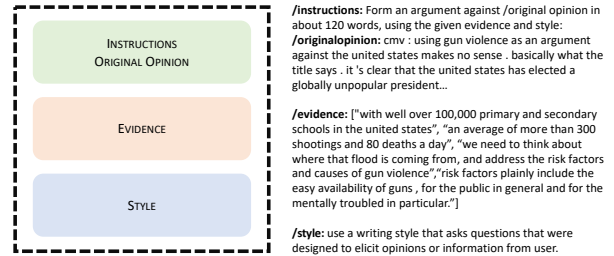


Figure 2: Example prompt for generating a reciprocal counter-argument.

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.

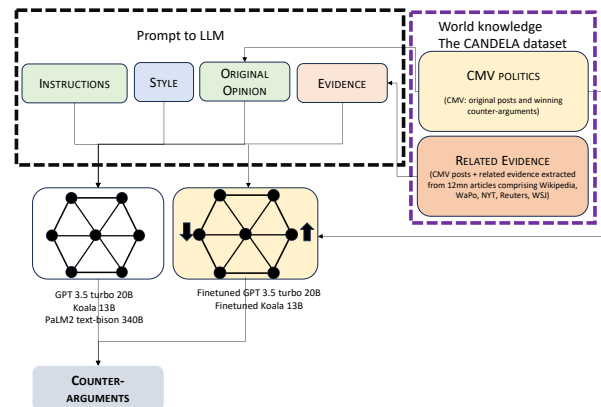


Figure 3: Experimental framework.

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 2000 rows of original posts and evidence from this dataset for further analysis.

**Generating stylized counter-arguments:** Five off-the-shelf and fine-tuned 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)  $\times$  3 variants for style control ( $N = 32,000$ )<sup>1</sup>. 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 fine-tuned variants with these inputs (fine-tuned 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 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 fine-tuning 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,

<sup>1</sup>Candela (2k) + Koala-13B (6k) + GPT3.5-turbo (6k) + Koala finetuned (6k) + GPT3.5-turbo finetuned (6k) + PaLM2 (6k) = 32k

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 Fine-tuned variants of GPT-3.5 turbo and Koala

GPT-3.5 turbo was fine-tuned using OpenAI’s Application Programming Interface (API) for three epochs. Fine-tuning for Koala-13B was done on 70,000 input and counter-argument pairs from our primary dataset using Colab Nvidia A100 GPU. 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. The training loss plot and the hyperparameter settings are reported in the Appendix. Fine-tuning 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 factual accuracy and relevance of the counter-arguments are 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 counterarguments.

## 4.2 Rhetorical insights

We performed automatic content analyses to characterize and compare the generated counterarguments 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).

## 4.3 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 to obtain user rankings for the best-performing counterarguments as pitted against the human-written counterargument. The survey was open to residents of the United States with at least a 96% approval rate (based on recent recommendations (Huang et al., 2023)) who had at least 5000 approved hits. The goal was to examine patterns in whether a user would favor an evidential or reciprocal argument style. In this manner, 10,000 counterargument rankings were collected from 1879 respondents. Further details about the ranking task are reported in the Appendix.

# 5 Results

## 5.1 Fact and style integration

For fact integration validation, we analyzed whether our prompts effectively got the LLMs to apply the provided evidence in the generated counterarguments 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). For style integration validation, we examined whether the LLMs could integrate the expected style into the outputs. This was done with the help of crowdsourced annotations from Amazon Mechanical Turk, and through fine-tuning OpenAI ada models on the CLAPTON dataset (Jaidka, 2022) to automatically label the presence of justification and reciprocity in the generated outputs. Details of the fine-tuning task are reported in the Appendix.

## 5.2 Argument quality assessment

Next, our evaluation techniques measure the content and argument quality of the counterarguments:

- Automatic content quality evaluation: ROUGE(1/2/L) and BLEU, recognized as overlap-based metrics (Lin, 2004; Papineni et al., 2002), are commonly used to measure the quality of generated counter-arguments against the target 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 argument quality evaluation: 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 Amazon Mechanical Turk. Details of the annotation task are reported in the Appendix.

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 counterargument yet yielded a low absolute overlap in the words used.

Table 2: The three variants of the style specifications added to the LLM prompt.  $\theta$  is the average annotator accuracy across true-positives and negatives (Passonneau and Carpenter, 2014)

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
Style integration		
Style	$\theta$ (Inter-annotator Accuracy)	
Reciprocity	0.9682	
Justification	0.7680	

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 second part of Table 2 demonstrates a high manual validation of the incorporation of style, with inter-annotator reliability of  $\theta = 0.9682$  for reciprocity and 0.7680 for justification, respectively.  $\theta$  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). 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 to be satisfactory as  $\theta \geq 0.65$ .

### 5.3 Evaluating argument quality

Table 3 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 the argument quality differences (but we later find that the quality scores may not reflect user preferences). Conversely, GPT-3.5 turbo counterarguments contain fewer specific details when they offer greater 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" counterargument 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 high median score in Relevance but also have a wide spread in Overall effectiveness, suggesting that they consistently perform better than Koala-13B, but with some inconsistency in how effective they are. 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 for the fine-tuned 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 fine-tuned Koala 13B are reported in the Appendix.

### 5.4 Rhetorical insights

In Table 4, we report the distribution of argument moves across the different types of counterargument variants. Alignment moves are examples of social acts that can involve agreement or refutation in argumentation. Of the exemplars of positive and negative alignment moves identified in the AAWD corpus, the Reddit counterarguments contained 12, while the GPT-3.5 turbo justification

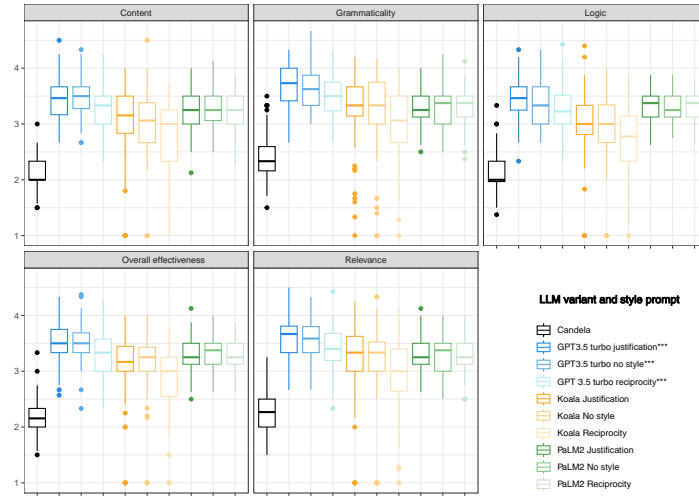


Figure 4: Results from the human evaluation on various dimensions. Candela is seen to trail GPT-3.5 turbo outputs on all aspects of content, grammar, logic, relevance, and overall effectiveness, with a Bonferroni-corrected statistical significance ( $p < 0.001$ ). GPT 3.5 turbo also outperforms Koala 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)	<b>0.33 0.33 (0.07)</b>	0.17 0.17 (0.06)	0.17 0.17 (0.06)
ROUGE-2	0.03 0.03 (0.03)	<b>0.10 0.09 (0.06)</b>	0.02 0.01 (0.02)	0.01 0.01 (0.02)
ROUGE-L	0.21 0.21 (0.06)	<b>0.29 0.29 (0.07)</b>	0.15 0.15 (0.05)	0.14 0.14 (0.04)
BLEU	0.00 0.00 (0.01)	<b>0.06 0.06 (0.06)</b>	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	<b>0.81</b>	0.75
Automatic evaluation: Style (Accuracy)				
Reciprocity	0.17	0.09	0.12	<b>0.49</b>
Justification	<b>0.42</b>	0.26	0.24	0.22
Automatic evaluation: Readability (0 to 1 scale)				
Flesch Kincaid Grade	6.40 6.00 (2.18)	<b>12.81 12.70 (2.07)</b>	12.75 12.70 (2.07)	11.79 11.60 (2.08)
Flesch Reading Ease	<b>83.10 84.00 (10.41)</b>	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)	<b>15.05 14.88 (2.23)</b>	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)	<b>14.87 14.80 (1.68)</b>	14.09 14.00 (1.72)

Table 3: 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 4: Total number of alignment moves identified in Counterfire outputs. Based on the AAWD corpus (Bender et al., 2011).

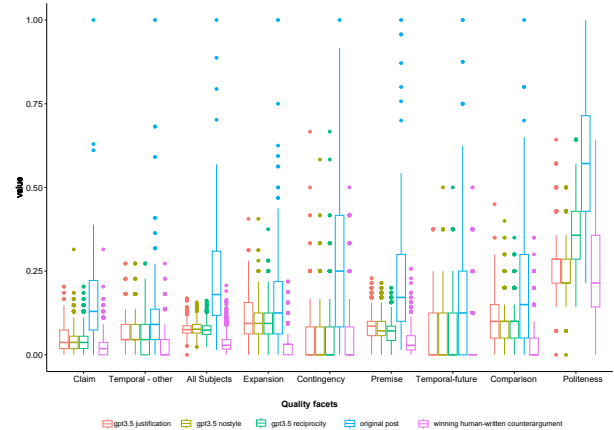


Figure 5: Results from automatic evaluation of argumentation using the discursive and politeness features of Convokit for the arguments considered in the human evaluation.

and reciprocity style counterarguments contained 2 and 4, respectively, exemplifying explicit agreement and positive alignment, such as praise thinking, and negative 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

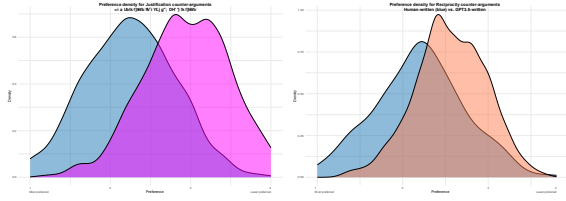


Figure 6: User preference analysis for human-written (blue) vs. GPT3.5-written counterarguments for (a) justification and (b) reciprocity.

counts among the variants, highlighting the domain differences compared to the AAWD corpus. The reciprocity-style counterargument appears to have more argument moves than the no-style and justification counterargument, perhaps because of its interpersonal nature. Finally, human-written arguments are the most argumentatively rich and diverse, with a higher number of unique moves across the different categories than the generated outputs.

Similarly, in the discursive analysis reported in Figure 5, we observe that the GPT3.5-written counterarguments are typically at par with each other concerning most of the discursive features, they significantly differ ( $p < 0.001$ ) from human-written counterarguments in covering more claims, temporal features, reference to subjects, premises, comparisons, and even politeness. Human-written counter-arguments appear more focus on fewer claims with greater specificity, to offer a more focused and less polite counterargument.

### 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 of justification-style counterarguments written by other humans more preferable to those written by GPT3.5, and this preference is statistically significant ( $p < 0.001$ ). The low preference for Justification raises red flags. On examining the outputs, we speculate that the reason may be the essay-style structure of the arguments devoid of any interpersonal engagement. Taken together with findings from Figure 5, the findings suggest that the highly-focused, specific, and less polite human counterarguments are somehow more persuasive than GPT3.5-generated counterarguments to hu-

mans, thereby offering food for thought in how accurately stylized text may still fall short of human expectations. We wonder if we are observing a tradeoff between fact integration and style while generating counterarguments.

## 6 Discussion and Conclusion

In this study, we have addressed the need for more research on style in political arguments and its relationship with persuasion and offered a new dataset for related research into fine-tuning, 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 counterarguments 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 preferred to LLM outputs, human-generated counterarguments tend to 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 counterarguments 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. Therefore, we didn't fine-tune it.

In future work, we are interested in developing dynamic models that accommodate a conversation partner's stylistic choices in generating a finely tailored counterargument 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 an annotation of the generated counter-arguments to ensure veracity and to pre-empt the selection or curation of irrelevant facts in the list of evidence ([Mendes et al., 2023](#)). Fine-tuning 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.

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 important. 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 fine-tune or domain-transfer pre-trained models to new contexts and populations. Furthermore, the data and message vocabulary is biased toward the topics popular in the subreddit and may not reflect contemporary events or even facts.

Additionally, GPT models have certain biases, and the hallucination problem can not be fully solved even when we provide external evidence. It is possible that the GPT-3.5 turbo was already trained on the CMV dataset. We will explore and fine-tune 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 an audit purpose, it would be potentially dangerous to fine-tune 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 fine-tuned 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 hard 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](#)). 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

711	<i>Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021</i> , pages 1816–1827.	
712		
713	Tim Althoff, Cristian Danescu-Niculescu-Mizil, and	
714	Dan Jurafsky. 2014. How to ask for a favor: A case	
715	study on the success of altruistic requests. In <i>Pro-</i>	
716	<i>ceedings of the International AAAI Conference on</i>	
717	<i>Web and Social Media</i> , volume 8, pages 12–21.	
718	Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz,	
719	and Noam Slonim. 2021. Project debater apis: De-	
720	composing the ai grand challenge. In <i>Conference on</i>	
721	<i>Empirical Methods in Natural Language Processing</i> .	
722	Aviv Ben-Haim and Oren Tsur. 2021. <a href="#">Open-</a>	
723	<a href="#">mindedness and style coordination in argumentative</a>	
724	<a href="#">discussions</a> . In <i>Proceedings of the 16th Conference</i>	
725	<i>of the European Chapter of the Association for Com-</i>	
726	<i>putational Linguistics: Main Volume</i> , pages 1876–	
727	1886, Online. Association for Computational Lin-	
728	guistics.	
729	Emily M. Bender, Jonathan T. Morgan, Meghan Oxley,	
730	Mark Zachry, Brian Hutchinson, Alex Marin, Bin	
731	Zhang, and Mari Ostendorf. 2011. <a href="#">Annotating so-</a>	
732	<a href="#">cial acts: Authority claims and alignment moves in</a>	
733	<a href="#">Wikipedia talk pages</a> . In <i>Proceedings of the Work-</i>	
734	<i>shop on Language in Social Media (LSM 2011)</i> ,	
735	pages 48–57, Portland, Oregon. Association for Com-	
736	putational Linguistics.	
737	Yonatan Bilu, Daniel Hershcovich, and Noam Slonim.	
738	2015. Automatic claim negation: Why, how and	
739	when. In <i>Proceedings of the 2nd Workshop on Argu-</i>	
740	<i>mentation Mining</i> , pages 84–93.	
741	Tom Brown, Benjamin Mann, Nick Ryder, Melanie	
742	Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind	
743	Neelakantan, Pranav Shyam, Girish Sastry, Amanda	
744	Askell, et al. 2020. Language models are few-shot	
745	learners. <i>Advances in neural information processing</i>	
746	<i>systems</i> , 33:1877–1901.	
747	Kushal Chawla, Balaji Vasan Srinivasan, and Niyati	
748	Chhaya. 2019. Generating formality-tuned sum-	
749	maries using input-dependent rewards. In <i>Proceed-</i>	
750	<i>ings of the 23rd Conference on Computational Natu-</i>	
751	<i>ral Language Learning (CoNLL)</i> , pages 833–842.	
752	Jiaao Chen and Diyi Yang. 2023. Controllable con-	
753	versation generation with conversation structures via	
754	diffusion models. In <i>Findings of the Association for</i>	
755	<i>Computational Linguistics: ACL 2023</i> , pages 7238–	
756	7251.	
757	Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal	
758	Chanda, and Jaya Singh. 2018. Frustrated, polite, or	
759	formal: Quantifying feelings and tone in email. In	
760	<i>Proceedings of the Second Workshop on Computa-</i>	
761	<i>tional Modeling of People’s Opinions, Personality,</i>	
762	<i>and Emotions in Social Media</i> , pages 76–86.	
763	Aaron Daniel Cohen, Adam Roberts, Alejandra Molina,	
764	Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben	
765	Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-	
766	Arcas, Chung-ching Chang, et al. 2022. Lamda: Lan-	
767	guage models for dialog applications.	
	Kevyn Collins-Thompson. 2014. Computational as-	768
	essment of text readability: A survey of current and	769
	future research. <i>ITL-International Journal of Applied</i>	770
	<i>Linguistics</i> , 165(2):97–135.	771
	Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan	772
	Jurafsky, Jure Leskovec, and Christopher Potts.	773
	2013a. <a href="#">A computational approach to politeness with</a>	774
	<a href="#">application to social factors</a> . pages 250–259.	775
	Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan	776
	Jurafsky, Jure Leskovec, and Christopher Potts.	777
	2013b. A computational approach to politeness with	778
	application to social factors. In <i>Proceedings of the</i>	779
	<i>51st Annual Meeting of the Association for Compu-</i>	780
	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages	781
	250–259, Sofia, Bulgaria. Association for Computa-	782
	tional Linguistics.	783
	Aida Mostafazadeh Davani, Mark Díaz, and Vinodku-	784
	mar Prabhakaran. 2022. Dealing with disagreements:	785
	Looking beyond the majority vote in subjective an-	786
	notations. <i>Transactions of the Association for Com-</i>	787
	<i>putational Linguistics</i> , 10:92–110.	788
	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and	789
	Luke Zettlemoyer. 2023. <a href="#">Qlora: Efficient finetuning</a>	790
	<a href="#">of quantized llms</a> .	791
	Yixi Ding, Yanxia Qin, Qian Liu, and Min-Yen Kan.	792
	2023. Cocoscisum: A scientific summarization	793
	toolkit with compositional controllability. In <i>Pro-</i>	794
	<i>ceedings of the 2023 Conference on Empirical Meth-</i>	795
	<i>ods in Natural Language Processing: System Demon-</i>	796
	<i>strations</i> , pages 518–526.	797
	Roxanne El Baff, Henning Wachsmuth, Khalid	798
	Al Khatib, and Benno Stein. 2020. <a href="#">Analyzing the Per-</a>	799
	<a href="#">suasive Effect of Style in News Editorial Argumenta-</a>	800
	<a href="#">tion</a> . In <i>Proceedings of the 58th Annual Meeting of</i>	801
	<i>the Association for Computational Linguistics</i> , pages	802
	3154–3160, Online. Association for Computational	803
	Linguistics.	804
	Marc Esteve Del Valle, Rimmert Sijsma, and Hanne	805
	Stegeman. 2018. Social media and the public sphere	806
	in the Dutch parliamentary Twitter network: A space	807
	for political deliberation? Hamburg, Germany.	808
	ECPR General Conference.	809
	Dennis Friess and Christiane Eilders. 2015. <a href="#">A system-</a>	810
	<a href="#">atic review of online deliberation research</a> . <i>Policy &amp;</i>	811
	<i>Internet</i> , 7(3):319–339.	812
	Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wal-	813
	lace, Pieter Abbeel, Sergey Levine, and Dawn Song.	814
	2023. <a href="#">Koala: A dialogue model for academic re-</a>	815
	<a href="#">search</a> . Blog post.	816
	Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022.	817
	News summarization and evaluation in the era of	818
	gpt-3. <i>arXiv preprint arXiv:2209.12356</i> .	819
	Christopher Hidey and Kathleen McKeown. 2019.	820
	Fixed that for you: Generating contrastive claims	821
	with semantic edits. In <i>Proceedings of the 2019</i>	822

823	<i>Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 1756–1767.	
824		
825		
826		
827	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean, Lu Wang, and Weizhu Chen. 2021. <a href="#">Lora: Low-rank adaptation of large language models</a> .	
828		
829		
830		
831	Xinyu Hua, Zhe Hu, and Lu Wang. 2019. <a href="#">Argument generation with retrieval, planning, and realization</a> .	
832	pages 2661–2672.	
833		
834	Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into mturk annotation practices for nlp. In <i>The 2023 Conference on Empirical Methods in Natural Language Processing</i> .	
835		
836		
837		
838	Kokil Jaidka. 2022. Talking politics: Building and validating data-driven lexica to measure political discussion quality. <i>Computational Communication Research</i> , 4(2):486–527.	
839		
840		
841		
842	Kokil Jaidka, Hansin Ahuja, and Lynnette Ng. 2023. It takes two to negotiate: Modeling social exchange in online multiplayer games. <i>arXiv preprint arXiv:2311.08666</i> .	
843		
844		
845		
846	Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. <a href="#">Knowledge-enhanced evidence retrieval for counterargument generation</a> .	
847	pages 3074–3094.	
848		
849		
850	Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. Controlled text generation with hidden representation transformations. <i>Image</i> .	
851		
852		
853		
854	Chin-Yew Lin. 2004. <a href="#">ROUGE: A package for automatic evaluation of summaries</a> . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	
855		
856		
857		
858	Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. <a href="#">Argument strength is in the eye of the beholder: Audience effects in persuasion</a> .	
859	pages 742–753.	
860		
861		
862	Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. <a href="#">Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments</a> . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.	
863		
864		
865		
866		
867		
868		
869	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. <i>arXiv preprint arXiv:2112.09332</i> .	
870		
871		
872		
873		
874		
	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 <i>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)</i> , pages 1650–1659.	875
		876
		877
		878
		879
		880
		881
		882
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. <a href="#">Bleu: a method for automatic evaluation of machine translation</a> . In <i>Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics</i> , pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.	883
		884
		885
		886
		887
		888
		889
	Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. <i>Transactions of the Association for Computational Linguistics</i> , 2:311–326.	890
		891
		892
		893
	Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In <i>Proceedings of the 2008 conference on empirical methods in natural language processing</i> , pages 186–195.	894
		895
		896
		897
		898
	Eike Mark Rinke. 2015. Mediated deliberation. <i>The International Encyclopedia of Political Communication</i> .	899
		900
		901
	Ian Rowe. 2015. Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. <i>Journal of broadcasting &amp; electronic media</i> , 59(4):539–555.	902
		903
		904
		905
	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. <i>arXiv preprint arXiv:2208.03188</i> .	906
		907
		908
		909
		910
		911
	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. <i>Nature</i> , 591(7850):379–384.	912
		913
		914
		915
		916
	Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. <a href="#">Measuring political deliberation: A discourse quality index</a> . <i>Comparative European Politics</i> , 1(1):21–48.	917
		918
		919
		920
	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. <a href="#">Llama: Open and efficient foundation language models</a> .	921
		922
		923
		924
		925
		926
	Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in	927
		928
		929
		930



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*.

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*.

## Appendix

## 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 fine-tuning hyperparameters:

- per\_device\_train\_batch = 1
- learning rate = 0.0002
- optimizer = Paged Adam 8bit optimizer

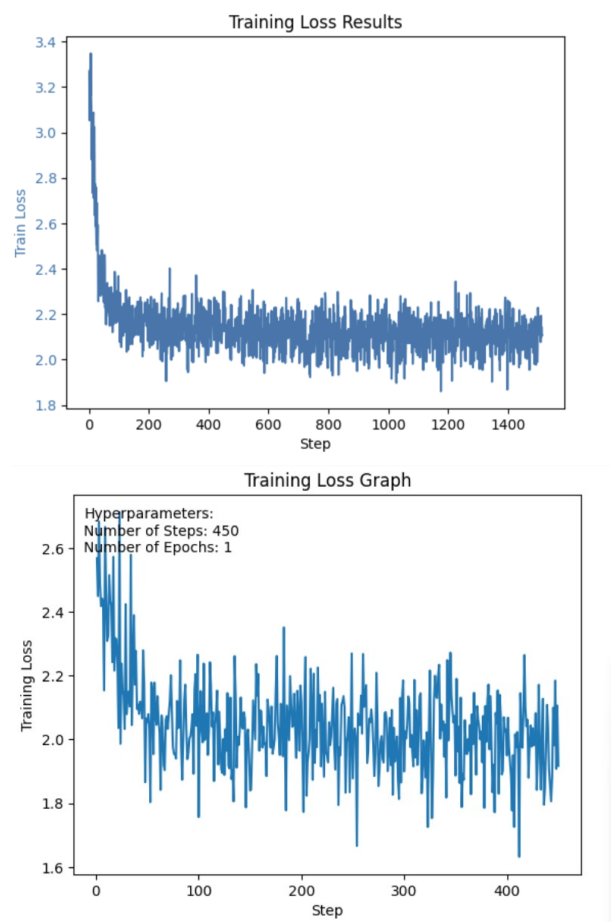


Figure 7: Fine-tuning training loss plots for (a) GPT3.5-turbo and (b) Koala

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

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 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. 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.

**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.

### 8.2.1 Instructions for style validation

The following was 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, may be 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

**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.

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, or
- \* 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

1053	outcome. Is this your genuine sentiment? Did someone influence your opinion?	medical and food stamps from the government. It's essential to stick to the facts and avoid spreading misinformation.	1104
1054			1105
1055			1106
1056	* The tax bill seems to require more than just minor adjustments. It appears to need a complete overhaul. Why not just reject it?	* Records show that you received \$6,986,620 from the NRA. This presents a clear conflict of interest when it comes to enacting common-sense gun laws.	1107
1057			1108
1058			1109
1059			1110
1060	* It's evident that Trey Gowdy speaks assertively, but when will we see him take decisive actions to match his words?		1111
1061			1112
1062		– NO: This comment does not offer a justification.	1113
1063			1114
1064	* 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?		1115
1065			1116
1066			1117
1067			1118
1068			1119
1069			1120
1070	* 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?		1121
1071			1122
1072			1123
1073			1124
1074			1125
1075	– NO: This comment does not ask a genuine question or asks rhetorical questions.		1126
1076			1127
1077			1128
1078	• <b>JUSTIFICATION Description:</b>		1129
1079	– YES: Personal: Whether this comment contains personal feelings or experiences, or		1130
1080			1131
1081	* Corporate Democrats, be aware that we're watching closely. You're on notice.		1132
1082			1133
1083	* Senator [name] from the Republican party stated, "We all recognize that [name] is not up to the mark."		1134
1084			1135
1085	* It seems like [name] has been given a blank check. Their credibility is questionable at this point.		1136
1086			1137
1087	* It's essential to stay informed and make our voices heard. If our representatives don't shape up, we'll vote them out.		1138
1088			1139
1089			1140
1090			1141
1091			1142
1092	– YES: Fact-based: Whether this comment contains facts, links, or evidence from other sources, or		1143
1093			1144
1094	* It's worth noting that previous administrations, like Obama's and Clinton's, allocated funds to foreign countries.		1145
1095			1146
1096			1147
1097			1148
1098			1149
1099			1150
1100			1151
1101			
1102			
1103			

### 8.2.2 Instructions for quality evaluation

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.

1152	• Flawless: The argument is extremely rich	The following were the step-by-step instruc-	1200
1153	and detailed, with complex arguments and a	tions:	1201
1154	wealth of supporting evidence.	These are counter-arguments posted in response	1202
1155	<b>Logic and reasoning:</b>	to an "Original Post" within a Reddit community	1203
1156	• Poor: The argument is illogical and poorly	called ChangeMyView.	1204
1157	reasoned.	Each counter-argument is an attempt to persuade	1205
1158	• Fair: The argument is somewhat illogical and	people against the viewpoint presented in the	1206
1159	poorly reasoned.	Original Post.	1207
1160	• Good: The argument is neither well nor poorly	Your task is to evaluate and order these counter-	1208
1161	reasoned, and has some logical flaws.	arguments based on their persuasiveness.	1209
1162	• Excellent: The argument is quite logical and	According to your preference, please state whether	1210
1163	well-reasoned.	you agree with the opinion in the original post.	1211
1164	• Flawless: The argument is very logical and	Next, at least once for this batch of HITs, please	1212
1165	flawlessly reasoned.	share your age and gender. These questions are	1213
1166	<b>Overall effectiveness:</b>	optional.	1214
1167	• Poor: The argument is very weak and fails to	Finally, according to your preference, please rank	1215
1168	convince me.	the arguments, with the most persuasive argument	1216
1169	• Fair: The argument is somewhat weak and	as #1.	1217
1170	unconvincing.		1218
1171	• Good: The argument is neither strong nor	<b>8.3 Additional results</b>	1219
1172	weak, and is somewhat convincing.	<b>8.3.1 Automatic evaluation</b>	1220
1173	• Excellent: The argument is quite strong and	Table 5 reports the automatic scores for con-	1221
1174	convincing.	tent and quality for Koala 13B-generated counter-	1222
1175	• Flawless: The argument is very strong and	arguments. Table 6 reports the automatic scores	1223
1176	completely convincing.	for content and quality for Koala 13B-generated	1224
1177	<b>8.2.3 Instructions for user preference analysis</b>	counter-arguments.	1225
1178	The original post was presented to each survey re-	For fine-tuned Koala 13B, Table 7 reflects the	1226
1179	spondent, followed by four counterarguments: the	content and style evaluation. In general, we ob-	1227
1180	human-written argument from the Candela dataset,	serve that the content and style scores fare poorer	1228
1181	and three variants from the GPT3.5-turbo. The me-	than GPT-3.5 turbo. Koala outputs had less content	1229
1182	dian age was 34.5 years. 691 (36.7%) were female,	overlap and were less readable than those generated	1230
1183	and 854 (45.4%) were male, while 74 (3.9%) iden-	through GPT-3.5 turbo. Koala and Loala fine-tuned	1231
1184	tified as non-binary or third gender. The remaining	outputs were also less grammatical, relevant, co-	1232
1185	respondents did not share their age nor gender.	herent, and less preferred overall as compared to	1233
1186	The following was the description of the task:	the counter-arguments generated through GPT-3.5	1234
1187	In this job, you will be presented with various	turbo. The total output and the results for Koala	1235
1188	counter-arguments posted in the ChangeMyView	13B are reported in the Appendix and the supple-	1236
1189	subreddit. In ChangeMyView, users present a view-	mentary materials <sup>2</sup> .	1237
1190	point, and others respond with counter-arguments	Table 8 reports the automatic scores for con-	1238
1191	to challenge or change the original viewpoint. Your	tent and quality for Koala 13B-generated counter-	1239
1192	role is to read these counter-arguments and assess	arguments.	1240
1193	their effectiveness in persuading against the Ori-	<b>8.3.2 Human evaluation</b>	1241
1194	ginal Post. Consider the logic, evidence, and clarity	<b>Evaluation of argument quality</b>	1242
1195	of each argument in your evaluation. Each HIT	Figure 10 reports the human evaluation scores for	1243
1196	will take approximately 2-3 minutes, depending on	the fine-tuned models, where they are seen to fol-	1244
1197	the length and complexity of the arguments. Pay	low a similar pattern to the off-the-shelf models.	1245
1198	attention to the strength of the reasoning and the		1246
1199	use of evidence in each counter-argument.		

<sup>2</sup>[https://anonymous.4open.science/r/Style\\_control-2018/](https://anonymous.4open.science/r/Style_control-2018/)

Table 5: 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: 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 6: 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: 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 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	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	<b>0.21 0.21 (0.06)</b>	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: 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)	<b>6.89 6.50 (3.93)</b>
Flesch Reading Ease	<b>83.10 84.00 (10.41)</b>	74.07 75.61 (17.32)	73.75 75.40 (19.47)	74.20 75.76 (18.02)
Gunning Fog	<b>8.85 8.57 (2.05)</b>	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)	<b>9.10 9.30 (3.21)</b>	9.06 9.30 (3.24)

Table 8: 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: 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)

## 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 9 and indicate that the annotation quality is reliable ( $\theta > 0.65$ ).



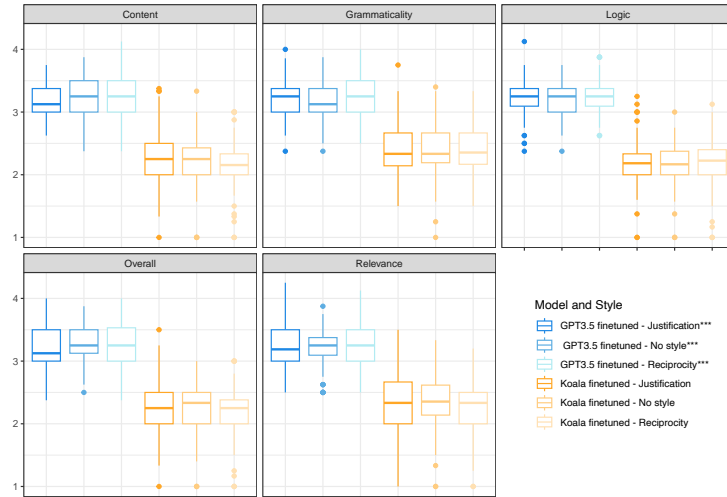


Figure 10: Results from the human evaluation on various dimensions. Koala 13B-finetuned is seen to trail GPT-3.5 turbo-finetuned outputs on all aspects of content, grammar, logic, relevance, and overall effectiveness, with a Bonferroni-corrected statistical significance ( $p < 0.001$ ).

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

Human annotation of argument quality	
	$\theta$ (Inter-annotator accuracy $\theta$ )
<b>Content</b>	0.8395
<b>Relevance</b>	0.8859
<b>Grammaticality</b>	0.8831
<b>Logic</b>	0.8891
<b>Overall effectiveness</b>	0.8951