Improving Argument Effectiveness Across Ideologies using Instruction-tuned Large Language Models

Roxanne El Baff [1,2] **Khalid Al-Khatib** [3] **Milad Alshomary** [4]
Kai Konen [1] **Benno Stein** [2] **Henning Wachsmuth** [5]
[1] German Aerospace Center (DLR), Germany
[2] Bauhaus-Universität Weimar, Weimar, Germany [3] University of Groningen, Netherlands
[4] Columbia University, New York, NY [5] Leibniz University Hannover, Germany
roxanne.elbaff@dlr.de

Abstract

Different political ideologies (e.g., liberal and conservative Americans) hold different worldviews, which leads to opposing stances on different issues (e.g., gun control) and, thereby, fostering societal polarization. Arguments are a means of bringing the perspectives of people with different ideologies closer together, depending on how well they reach their audience. In this paper, we study how to computationally turn *ineffective* arguments into *effective* arguments for people with certain ideologies by using instruction-tuned large language models (LLMs), looking closely at style features. For development and evaluation, we collect ineffective arguments per ideology from debate.org, and we generate about 30k, which we rewrite using three LLM methods tailored to our task: zero-shot prompting, few-shot prompting, and LLM steering. Our experiments provide evidence that LLMs naturally improve argument effectiveness for liberals. Our LLM-based and human evaluation show a clear preference towards the rewritten arguments.[1]

1 Introduction

Political polarization has significantly increased over decades across countries (Boxell et al., 2022; Inglehart and Norris, 2016; Foa and Mounk, 2016). In the US, for example, it has grown dramatically not only because of differences of opinion in politics but also because of the country's cultural system (Dimock et al., 2014; Klein, 2020). According to Lakoff (2010), modern American politics revolves around conflicting worldviews. Conservatives and liberals see the world differently and often struggle to understand each other's perspectives. Such a divide leads to contentious debates among the political ideologies, either promoting agreement or pushing them apart. While some studies suggest that cross-party online interactions can increase polarization and incivility (Papacharissi, 2004), other research indicates that people moderate their views when they engage with those with different perspectives (Balietti et al., 2021; Zhang, 2019).

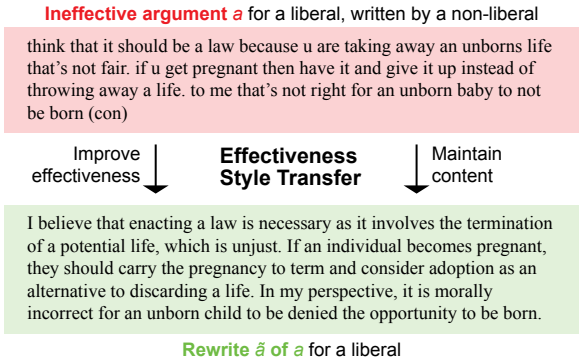


Figure 1: Example of the proposed task to rewrite an *ineffective argument*, a, into an *effective* one, $\tilde{a}$. The writer of argument a has ideology $\mathcal{I}_1$ (e.g., non-liberal), while the rewrite $\tilde{a}$ is intended to be perceived by a reader of ideology $\mathcal{I}_2$ (e.g., liberal).

To bridge ideological divides, we define a new task in this paper: to transfer ineffective arguments into effective ones. We regard an argument as *effective* if it moves the perspective of a reader with a specific ideology $\mathcal{I}_1$ (e.g., liberal) closer to that of the argument's writer holding a different ideology $\mathcal{I}_2$ (e.g., conservative). This task can be seen as an instance of text style transfer, as exemplified in Figure 1: improve the argument's effectiveness while maintaining its content.

We study this task in three phases in this paper, as overviewed in Figure 2: data curation, generation with LLM, and evaluation.

Data Curation We focus on American ideologies and create two datasets with 82k arguments for conservative and 55k for liberal readers, based on the debate dataset of Durmus and Cardie (2019). All arguments have binary labels reflecting their effectiveness as perceived by readers from an ideol-

[1]Code and link to the data are available here: https://github.com/roxanneelbaff/emnlp2024-iesta

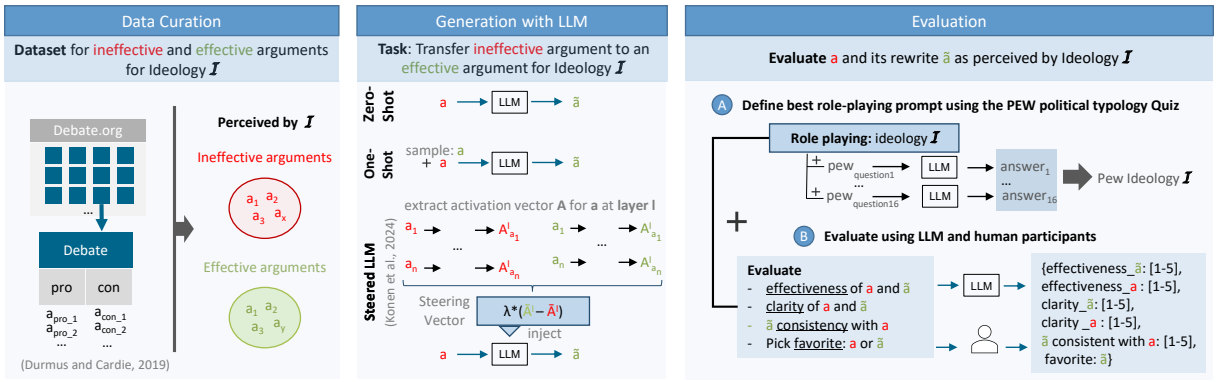


Figure 2: Overview of the approach for the task of transferring ineffective arguments into effective ones, as perceived by the ideology $\mathcal{I}$, with the three phases: *Data Curation* creates a non-parallel dataset with ineffective arguments and effective arguments for an ideology $\mathcal{I}$. *Generation with LLM* adapts three existing methods to transfer an ineffective argument (a) to a more effective one, $\tilde{a}$. *Evaluation* analyzes a and $\tilde{a}$ using the LLM-based method with role-playing for ideology $\mathcal{I}$ and the human-based evaluation with participants with ideology $\mathcal{I}$.

ogy derived from pre- and post-debate voting. The datasets are non-parallel, i.e., effective arguments are not paired with ineffective ones, making the style transfer task particularly challenging.

Generation with LLM We adapt three methods for effectiveness style transfer of arguments, leveraging the capabilities of instruction-tuned large language models with in-context learning. Each method employs a specific technique for LLM prompting and tuning: (a) *zero-shot prompting*, where we devise prompts for LLMs with various wordings that instruct the model to maintain content while changing the style to target a particular ideology; (b) *few-shot prompting*, where we add an effective argument derived from our datasets to the prompt in order to examine whether the LLM can learn from examples, and (c) *LLM steering*, where we incorporate 1 000 target examples using an activation engineering-based approach (Konen et al., 2024), a cost-effective approach fitting the non-parallel nature of the data used. In total, we generate 32 112 argument rewrites from 500 ineffective arguments per ideology, selected from the conservative and liberal datasets from the *Data Curation*-phase.[2]

Evaluation We first evaluate the effectiveness of the generated arguments $\tilde{a}$ using an automatic evaluation based on an effectiveness classifier, among others, and show a significant increase in effectiveness for $\tilde{a}$ compared to a for liberals and an opposite behavior for conservatives. To further assess the quality of $\tilde{a}$, we conduct a zero-shot LLM-based evaluation and a human evaluation for 100 argument tuples (a, $\tilde{a}$) for both ideologies, resulting in a total of 1 240 human annotations. Contrary to what our automatic evaluation suggests, both LLM and humans prefer the reformulations for both ideologies.

In conclusion, the main contributions of this paper are: (1) a new task of effectiveness style transfer, requiring transforming an ineffective argument into an effective one for a given ideology $\mathcal{I}$, (2) two ideology-based datasets for the new task, one for conservatives and one for liberals, and (3) three LLM-based method variants to tackle the task, coupled with an LLM evaluation based on role-playing.

2 Related Work

The study in this paper closely relates to research on *argument quality* and *audience-aware text generation*. Technologically, it builds on *text style transfer* and *large language models*. We discuss related work for each in the following.

Argument Quality Wachsmuth et al. (2017) present a taxonomy of argument quality with three main dimensions: *cogency* (whether an argument is reasoned well) (Gurcke et al., 2021; Stab and Gurevych, 2017), *effectiveness* (whether it persuades the target audience) (Habernal and Gurevych, 2016), and *reasonableness* (whether it contributes to an issue's resolution) (Gretz et al., 2020). We focus on the second, where we consider an argument effective if it changes a reader's stance, in line with our prior work (El Baff et al., 2018).

[2] In the following text, we denote an ineffective argument with a and its rewrite with $\tilde{a}$.

Several researchers explored argument effectiveness correlation with reader profiles (e.g., personality, political ideology) (Durmus and Cardie, 2018; Al Khatib et al., 2020; El Baff et al., 2020a). We ourselves studied the effectiveness of news editorials on American liberals and conservatives (El Baff et al., 2020b). These ideologies are also in the focus here but for the domain of debate portals and for a different task: transferring an ineffective argument to an effective one.

Audience-Aware Text Generation Audience-aware text generation is tackled in different forms for different genres. Takmaz et al. (2023) improved dialogical communication by changing a speaker's utterance on the fly to get closer to the listener's domain of expertise. Also, Stewart and Mihalcea (2022) developed a framework for generating questions based on social groups. Regarding argument generation, we proposed a method to arrange argument units given a rhetorical style in early work (El Baff et al., 2019). Later, Alshomary et al. (2021) generated audience-specific claims based on their stance on a known topic, whereas Alshomary et al. (2022) modeled audiences based on their moral beliefs. Research still needs to address the task of argument improvement subject to a specific ideology, which we focus on here.

Text Style Transfer One of the key properties driving text style transfer is whether the available data is parallel. For non-parallel data, as in our case, a corpus contains texts with a style, whereas another corpus contains the other style. Methods surveyed by Jin et al. (2022), such as style-content disentanglement (John et al., 2019), and prototype editing (Sudhakar et al., 2019), are usually applied for specific style features, such as politeness (Madaan et al., 2020)), and mostly on short text. To overcome the challenges of non-parallel transfer for long text, Ziegenbein et al. (2024) recently developed a reinforcement learning-based approach to teach a large language model (LLM) how to rewrite inappropriate arguments into appropriate ones. For effectiveness, we study how to best leverage the default capabilities of LLMs. In another work, Moorjani et al. (2022) apply style infusion for the *stylistic objective* of argumentative *persuasion*. Likewise, we focus on a stylistic objective, argument *effectiveness*. While Moorjani et al. (2022) use audience preference via pairwise comparison for training, we rely on a data-driven approach via ideology-based datasets labeled by effectiveness.

Large Language Models The transformer architecture (Duan and Zhao, 2020) caused significant progress, leading to instruction-turned LLMs such as ChatGPT (Brown et al., 2020; OpenAI, 2023), and Llama2-chat (Touvron et al., 2023), which are capable of solving downstream tasks such as hate speech detection (Feng et al., 2023). We leverage these capabilities to explore their performance in transforming an ineffective argument into an effective one. Previous work reveals the ideological biases of LLMs (Feng et al., 2023). Similarly, we provide evidence that the behavior of LLMs reflects a political ideology leaning, improving arguments for liberals more seamlessly.

Moreover, recent studies measure political bias of LLMs by adopting methods from political research applied previously to humans, such as the multiple-choice quiz "The Political Compass" (Hartmann et al., 2023; Rutinowski et al., 2024). Additionally, LLMs can be steered towards specific ideologies through *role-playing*, accurately mimicking personas like a Democrat or a Republican (Motoki et al., 2024). Likewise, we steer LLMs to mimic liberal/conservative personas to evaluate an argument's effectiveness, validating success through results from a political ideology quiz.

3 Task and Data

This section introduces the proposed task and the data created for our experiments and evaluation.

3.1 From Ineffective to Effective Arguments

We define the task of transferring an ineffective argument into an effective argument as follows:

Task Given an argumentative text a written by an author with the ideology $\mathcal{I}_1$ rewrite a as

$$\widetilde{a} \;:=\; f(a), \tag{1}$$

such that f improves the *effectiveness* of a perceived by readers with ideology $\mathcal{I}_2$, $\mathcal{I}_1 \neq \mathcal{I}_2$, while preserving its content. Improvements and content preservation are operationalized as

$$E(\widetilde{a}) > E(a) \quad \text{and} \quad S(a, \widetilde{a}) \geq \tau, \tag{2}$$

where E represents an effectiveness measurement for $\mathcal{I}_2$ and S a similarity function to assess content preservation against a threshold $\tau \in [0, 1]$.

3.2 Data for Development and Evaluation

The data curation phase creates two ideology-specific datasets (see *Data Curation* in Figure 2), one for conservatives and one for liberals. Each dataset has non-parallel effective and ineffective arguments for the corresponding ideology. In the following, we explain how we derive such structure from the debate dataset of Durmus and Cardie (2019). We start by describing the dataset, and then discuss what makes an argument effective (or ineffective) for an ideology.

In Durmus and Cardie (2019)'s dataset, each debate has two debaters arguing about the topic represented in the debate's title: one supporting the topic (*pro*) and the other opposing it (*con*). A debate consists of one or more rounds, where each debater presents one argument per round. On debate.org, users can vote for each debater within one debate using assessment criteria, two of which are used in our study: "Agree before debate" and "Agree after debate". Both capture whether the voter holds the same stance as a debater, once posed before and once posed after the debate. Besides, users of this platform, debaters and voters, can reveal their ideology (e.g., liberal, conservative).[3] The dataset contains 46k debates from the online platform debate.org. We illustrate a debate's structure in Figure 3.

What Makes an Argument Effective? As depicted in Figure 3, a voter v with ideology $\mathcal{I}_1$ (e.g., a liberal) perceives an argument written by a debater d with an ideology $\mathcal{I}_2$ (different than $\mathcal{I}_1$, e.g., a conservative), as *effective*, if it meets the following two criteria:

- *Before the debate.* v disagrees with d on the topic. For example, d opposes banning non-electric cars, while v supports it. The flag "Agrees before the debate" is set to *false*.
- *After the debate.* v changes their stance and now agrees with d. The flag "Agrees after the debate" is set to *true*.

In other words, v *flips their stance* after the debate to match d where v and d have different ideologies. Thus, an effective argument decreases the gap between v ($\mathcal{I}_1$) and d ($\mathcal{I}_2$) by encouraging v to reconsider their stance, reducing ideological disparity. Ineffective arguments miss this impact (the flag "Agrees after the debate" is set to *false*).

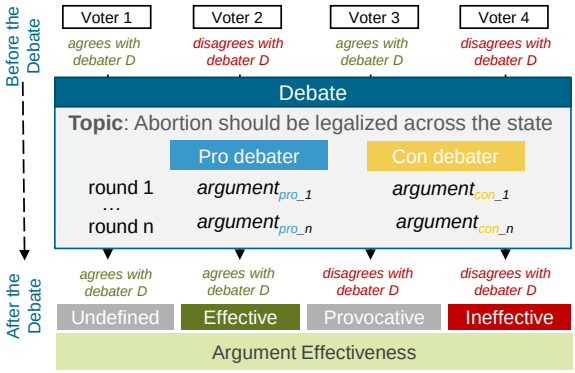


Figure 3: The debate structure and voting structure on debate.org. In each debate, one debate supports a given topic (*pro*), and the other opposes it (*con*). A debate can consist of one or multiple rounds, with one argument per round per debater. Voters have the option to indicate whether they *agree* or *disagree* with a debater's stance (pro/con) before the debate and then after the debate.

We consider two scenarios here, one with liberal voters and non-liberal debaters and the other with conservative voters and non-conservative debaters. To this end, we consider the answers to the two questions alongside the users' revealed ideologies to categorize each argument into one of four effectiveness groups (as illustrated by Figure 3). We consider the following two:[4]

- **Effective:** The voter disagrees with a debater before the debate, but agrees after the debate.
- **Ineffective:** The voter disagrees with the debater before and after the debate.

While these criteria model the overall effectiveness of the debater across all his arguments in the debate, we simplify here the assumption that each individual argument reflects this effectiveness.

Table 1 displays the difficulty of affecting readers with different ideologies. Only 2.6% of conservative readers were swayed by arguments from opposing ideologies, a trend mirrored by liberal readers (3.6%). Overall, the majority found the argument effective in $\leq 2.2\%$ of samples.

We construct two datasets, *liberal_dataset* and *conservative_dataset*, where the ideology denotes the voters' ideology.[5]

[3] There are 13 ideologies in debate.org: we picked the two prominent ones with the highest user count: conservatives, with 2,500 users, and liberals, with 1800 users.

[4] We omit the usage of *undefined* and *provocative* arguments because the former has no clear quality, and the latter is a special type of *ineffective*.

[5] The data is preprocessed to ensure its quality (Appendix A).

	Voters	Debates	At least 1 Vote		Majority Votes	
			Ineff.	Eff.	Ineff.	Eff.
Conservative	**1**	24 781	77.8%	1.9%	77.8%	1.9%
	2	6 671	92.3%	3.7%	65.7%	0.1%
	3	2 059	96.0%	4.7%	85.0%	0.3%
	All	**35 121**	**82.5%**	**2.6%**	**76.3%**	**1.4%**
Liberal	**1**	16 503	78.8%	2.9%	78.8%	2.9%
	2	3 590	92.6%	5.3%	70.3%	0.3%
	3	1 223	96.8%	5.9%	88.6%	0.7%
	All	**22 241**	**82.8%**	**3.6%**	**78.4%**	**2.2%**

Table 1: Proportions of debates having at least one effective or ineffective vote, or majority votes, given the number of *voters* (1-3 or All) per debate. Upper part: for conservative voters; lower part: for liberal ones.

	Conservative		Liberal	
	Ineffective	Effective	Ineffective	Effective
Training	57 890	1 995	25 824	1 785
Validation	15 732	599	10 552	512
Test	7 943	336	5 328	248

Table 2: Distribution of arguments over the two effectiveness labels in the training, validation, and test set for conservative voters (left) and liberal voters (right) before preprocessing.

We first split each dataset into two subsets pseudo-randomly: 60% of the debates are used to develop and evaluate approaches that *change* arguments from ineffective to effective and for style analysis. The other 40% are used to train a classifier to *assess* whether an argument is effective. For the latter, each subset is further partitioned into datasets for training (70% of all debates), validation (10%), and test (20%). We base our split on debate percentages rather than individual data points (i.e., one argument per debate) to prevent potential leakage between the splits. Statistics related to the two subsets are presented in Table 2.

4 Approach

This section presents the methods that we adapt to the task of effectiveness style transfer (Figure 2, generation), as well as our general evaluation methodology (Figure 2, evaluation).

4.1 Methods Adaptation for Style Transfer

To tackle the task of transferring an ineffective argument to an effective one for an ideology $\mathcal{I}$ (Section 3.1) using non-parallel data (as the one from Section 3.2), we adapt three existing methods, based on the idea of in-context-learning of large language models (LLMs). The three methods represent three different levels of using context (in terms of training data) when prompting an LLM: (a) zero-shot learning, (b) one-shot learning, and (c) a Steered LLM where any number of $n \geq 1$ data points can be incorporated within the context.

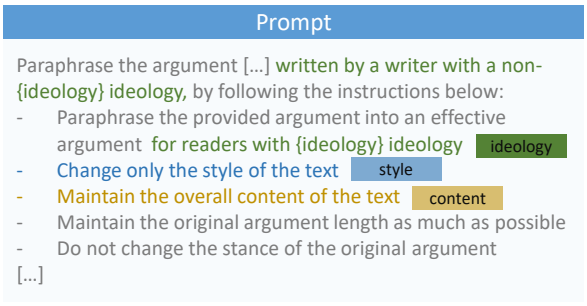


Figure 4: Prompt wording used to change an ineffective argument into an effective one. The *base* prompt (non-colored words) is used alone or together with other context information, such as the *ideology* of the reader and writer, and textual aspects, such as *style* and *content*.

Prompt Wording In all three methods, we examine a series of prompts that contain different instructions related to the task. We form prompts that encompass information about the argument's *effectiveness*, the reader's *ideology*, and text features such as *content* and *style*, as shown in Figure 4. Our prompts consider different combinations of inputs (e.g., *ideology* alone, *ideology+text features*).

Zero- and One-shot Prompting The zero-shot setting serves as a baseline to reflect the immediate effect of different prompts. For one-shot prompting, the non-parallelism of our data makes it almost impossible to provide examples that precisely reflect our task in the form of *input-output*. Instead, we develop prompts with a *sample* of what constitutes an "effective" argument by providing a random training instance of an effective argument on the same topic as the ineffective one we aim to enhance. Due to the notable length of arguments, we provide only one example per prompt.[6]

Steering the LLM Given the non-parallel nature of our data, fine-tuning the LLM specifically for our task is not a possible option. To overcome these limitations and provide the LLM with more examples, we follow the approach of Konen et al. (2024), namely to steer the underlying LLM to generate effective arguments when given a prompt. This

[6]Most common LLMs allow a context of 4 096 tokens; the argument length can exceed 2 048 tokens in our datasets.

approach allows the inclusion of several (e.g., thousands) of training instances in a cost-effective manner. In particular, activations are modified during inference to alter the model's behavior by injecting a *steering vector* in the forward pass, where this steering vector encapsulates the source (ineffective) and target (effective) classes.

More precisely, the injected steering vector is calculated as follows (Figure 2, generation > *Steered LLM*): (1) select n samples from each class, *effective* and *ineffective*, (2) for each sample, extract the activation vectors at layers 18, 19, and 20, following Konen et al. (2024), (3) for each class and layer, concatenate the extracted activation vectors A^{layer_l}, and (4) substract $A^{layer_l}_{ineffective}$ from "effective" $A^{layer_l}_{effective}$ with a coefficient λ.

Unlike Konen et al. (2024), who employed this approach on question-answering tasks for steering the answer's sentiment, we employ it in a style transfer task, using our dataset, which contains longer texts than the datasets used in the original work.

4.2 Evaluation Methodology

As defined by Equation 2, $\widetilde{a}$ must have a higher effectiveness score and preserve the meaning of a. Our evaluation consists of three phases employing different methods for E and S: (1) *Automatic evaluation*: We train a classifier for E and calculate the semantic similarity between a and $\widetilde{a}$ for S. We then pick models based on the results. (2) *LLM-based evaluation*, and (3) *human evaluation*: We conduct the same evaluation for k tuples $(a, \widetilde{a})$ to compare the values of E and S (below, we set $k = 100$). For the former, we use a zero-shot LLM approach that imitates a human-based evaluation using role-playing to impersonate an ideology, which is more cost-effective than human evaluation.

5 Automatic Evaluation

We report on our experiments by describing how we generated $\widetilde{a}$ from a. Then, we report on the automatic evaluation metrics and results.

The experiments are implemented using LangChain (Chase, 2022) and HuggingFace (Wolf et al., 2020), and the code is executed on Google Colab with an A100 GPU and 40GB of RAM.

5.1 Data

For evaluation, we input 500 randomly selected ineffective arguments from both the liberal and the conservative dataset (see Section 3) to two LLMs, one proprietary and one open-source model:

- ChatGPT (Brown et al., 2020)
- Llama-2-7b-chat (Touvron et al., 2023)[7]

Both models are employed in zero-shot and one-shot experiments.[8] For the Steered-LLM experiments, we employ Llama-2-7b-chat only due to the need for model altering. Specifically, as illustrated in Figure 4, we use five different prompts for each of the three methods (Zero-shot, One-shot, and Steered LLM). All prompts include the *base* wordings (in black) along with additional context: *base*, *content*, *ideology*, *style*, and *content+ideology+style* (complete). In total, we generate 15 500 arguments for each ideology dataset in a series of experiments.

Each experiment requires selecting a specific model type (Llama-2-7b-chat or ChatGPT), employing one of the three methods, and using specific prompt wording. With each model, method, and prompt wording, we generate 2 500 arguments.

After post-processing the data, we end up with 80.3% of the pairs $(a, \widetilde{a})$. Appendix C elaborates on error analysis, showing that Llama2 fails to answer when a is toxic, or the prompt contains *ideology*).

5.2 Metrics

We assess the arguments generated by the LLMs in terms of two dimensions: (1) an *effectiveness score* derived from our dataset, and (2) *Embeddings similarity* representing the percentage of similar ineffective-generated arguments.

Effectiveness Score For each ideology, we train one classifier to distinguish between effective and ineffective arguments, using the model *allenai/longformer-base-4096* (Beltagy et al., 2020) on the data from Section 3. The training is repeated ten times with different random seeds for ten epochs. The best macro-F_1 score achieves for both conservative and liberal classifiers is 0.6. A random classifier using the same seeds achieves a mean macro-F_1 of 0.38 (liberal) and 0.37 (conservative). The Longformer models significantly outperform these baselines at $p < .01$. The limited performance is due to the task's difficulty, as highlighted in Table 1. Similarly, El Baff et al. (2020b)

[7]The Llama-2 family was the best open-source performing model in initial experiments. Due to the limited resources, we use the 7B model. For simplicity, we refer to it as Llama2.

[8]We only use 1-shot due to the limited context for Llama2 and the long argumentative texts.

	Models	Prompt	Conservatives: Effectiveness	Conservatives: Similarity
Zero-Shot	ChatGPT	base	17% (↓0.02, ↓0.01, 0.08)†	0.91 (±0.06)
		complete	26% (↓0.0, ↑0.01, 0.07)$^{-}$	0.91 (±0.07)
	Llama2	base	19% (↓0.02, ↓0.03, 0.15)†	0.87 (±0.11)
		style	20% (↓0.02, ↓0.03, 0.15)*	0.85 (±0.10)
One-Shot	ChatGPT	base	14% (↓0.04, ↓0.02, 0.09)	0.89 (±0.10)
		complete	19% (↓0.02, ↓0.0, 0.09)†	0.89 (±0.10)
	Llama2	base	20% (↓0.01, ↓0.01, 0.16)$^{-}$	0.82(±0.11)
		complete	23% (↓0.02, ↓0.02, 0.15)*	0.81 (±0.12)
Steered-LLM	Llama2 mean, λ0.2	base	21% (↓0.02, ↓0.01, 0.15)*	0.86 (±0.10)
		style	18% (↓0.02, ↓0.02, 0.16)*	0.86 (±0.09)
	Llama2 mean, λ0.5	base	18% (↓0.01, ↓0.01, 0.15)$^{-}$	0.84 (±0.11)
		ideology	21% (↓0.02, ↓0.02, 0.16)$^{-}$	0.82 (±0.12)

	Models	Prompt	Liberals: Effectiveness	Liberals: Similarity
Zero-Shot	ChatGPT	base	46% (↑0.04, ↑0.04, 0.05)	0.91 (±0.05)
		complete	49% (↑0.04, ↑0.04, 0.05)	0.91 (±0.07)
	Llama2	base	44% (↑0.03, ↑0.03, 0.1)	0.83 (±0.11)
		content	48% (↑0.03, ↑0.03, 0.1)	0.86 (±0.10)
One-Shot	Llama2	base	42% (↑0.03, ↑0.03, 0.06)	0.88 (±0.11)
		complete	46% (↑0.04, ↑0.03, 0.06)	0.89 (±0.10)
	ChatGPT	base	52% (↑0.04, ↑0.05, 0.1)	0.83 (±0.11)
		style	54% (↑0.04, ↑0.04, 0.1)	0.83 (±0.10)
Steered-LLM	Llama2 mean, λ0.2	base	46% (↑0.03, ↑0.03, 0.1)	0.85 (±0.10)
		content	49% (↑0.03, ↑0.03, 0.1)	0.85 (±0.10)
	Llama2 mean, λ0.5	base	45% (↑0.03, ↑0.03, 0.09)	0.84 (±0.12)
		ideology	44% (↑0.03, ↑0.03, 0.1)	0.82 (±0.13)

Table 3: Evaluation of the LLM-generated *effective* arguments for *conservatives* (left) and for *liberals* (right). *Effectiveness* is the percentage of *effective* (> 0.5) arguments for *base* prompt and for the *best* performing prompt (or second best, in case *base* is the best), along with the trend (↑ increase, or decrease ↑) mean difference between the original and generated text, mean median and standard deviation. Similarity shows the median cosine similarity between a and $\widetilde{a}$. All scores are significantly different at $p < .0001$ except for: $^{-}$ indicates no significance, * at $p < .05$, † at $p < .01$, ‡ at $p < .001$.

reported low performance (< 0.45 macro-F_1) for models classifying news editorial's effectiveness.

We report on the *percentage of texts with a score > 0.5*. To address the classifier's insufficient performance, we apply it to the original text and the generated text, and we report on the significance between the scores as shown in Table 3. We report on the percentage of $E(\widetilde{a}) > 0.5$, as well as (in parenthesis) the mean, median, and standard deviation of the change in score for $E(\widetilde{a}) - E(a)$. An ↑ indicates an improvement, and a ↓ a decline.

Embedding-based Cosine Similarity To measure whether $\widetilde{a}$ maintains the content of a, we calculate their embeddings' cosine similarity and report the median and standard deviation. We choose one of the top-ranked (Muennighoff et al., 2022) embedding models supporting long context, *BAAI/bge-m3* (Chen et al., 2024).

5.3 Results

Table 3 shows the evaluation scores for each ideology, conservatives (left) and liberals (Right).

Conservatives All implemented methods using Llama2 and ChatGPT generate less effective arguments. The negative mean difference (↓ .01–.02) indicates a slight decrease in the effectiveness of $\widetilde{a}$ compared to a. ChatGPT models show more consistency (.07–.09) than Llama2 models (.15–.16). ChatGPT (zero-shot) has the highest percentage of $E(\widetilde{a}) > .50$ (26%). All models perform reasonably well regarding similarity, with ChatGPT (zero-shot) delivering the highest content preservation (.91).

Liberals All implemented methods using Llama2 and ChatGPT generate significantly more effective arguments than the original ones at $p < .0001$, showing stability across prompt and data point variations. The positive mean difference (↑ .03–.04) indicates a slight but steady increase in the effectiveness of $\widetilde{a}$ compared to a. ChatGPT models, with standard deviation range .05–.06 seem to be more consistent than Llama2 models (09–.10). Llama2 with one-shot has the highest percentage of $E(\widetilde{a}) > .50$ (54%). All models perform reasonably well regarding similarity (similar to *conservatives*). LLMs' ability to improve arguments for a liberal audience aligns with prior findings of their left-leaning tendencies (Santurkar et al., 2023).

The impact of the low-performing classifier is mitigated by subsequent LLM- and human-based evaluation.

6 LLM-Based and Human Evaluation

We conduct a zero-shot LLM-based and human-based evaluation to assess the *effectiveness*, *clarity*, *consistency*, and *preference* between a (original ar-

gument), $\widetilde{a}_{chatgpt}$, and $\widetilde{a}_{llama2}$ for arguments generated respectively from ChatGPT-based and Llama2-based models. For liberals, we chose zero-shot ChatGPT (complete-prompt) and one-shot Llama-2-7b-chat (style-prompt) since the effectiveness significantly improved (§ 5.3). For conservatives, in turn, we chose two models randomly due to the overall degradation in effectiveness (Table 3): zero-shot ChatGPT (ideology-prompt) and Steered LLM ($\lambda = 0.2$) (style-prompt).

We randomly selected 100 arguments (50 per ideology), a, with their generated effective counterparts, $\widetilde{a}$. Each triplet contained the original argument and the two rewrites, where each annotator reported an *effectiveness* score (1: fully ineffective – 5: fully effective), a *clarity* score (1–5), *consistency* with the original argument (1–5), and a preferred argument among the three (For more details, check Appendix D for LLM-based evaluation, and Appendix E for human-based evaluation).

6.1 LLM-Based Evaluation

We use zero-shot LLM prompting as an evaluator, employing role-playing to impersonate a liberal and a conservative annotator, where we prepend an ideology-impersonation prompt to the evaluation prompt ("From now on you are `{ideology}`"). We use the two models, GPT4 and Mixtral8x7B, with temperature .7 for variability.

Ideology-Impersonation. Before conducting the annotations, we perform experiments to align the LLM with a specific ideology using *ideology-impersonation* and the PEW political typology quiz[9], which contains 16 questions defining American ideologies from far conservative to liberal. We follow two steps: (1) **Initial ideology**: we determine the default ideology for each model by prompting the LLM with the PEW questions: GPT4 aligns with *Outsider Left*, and Mixtral7x8 with *Established Liberal*. (2) **Ideology impersonation:** We select *politically active* ideologies, namely, *Progressive Left* and *Faith and Flag Conservative* for liberal and conservative ideologies, respectively. We then prepend the impersonation prompt template inspired by Kong et al. (2023) ("From now on you are `{ideology}`") to each PEW question (elaborated in Appendix D).

Each quiz was repeated 30 times for stability. Results of the PEW test (Appendix D.1) show that both models consistently match the impersonation ideology across all quizzes. We use these templates to create liberal and conservative annotators by prepending the matching impersonation ideology to the prompt.

[9] https://www.pewresearch.org/ (accessed May 2024)

			Effectiveness		Clarity	Consis.	Pref.
	Argument		Mean	%HE	Mean	Mean	%
A. LLM-Based Evaluation							
Conservative	$\widetilde{a}_{Llama2}$	M	3.35 ±0.8	53%	4,60 ±0.7	4.07 ±1.3	45%
		G	3.52 ±1.0	67%	4.18 ±0.7	3.55 ±1.3	45%
	$\widetilde{a}_{ChatGPT}$	M	3.41 ±0.8	53%	4.51 ±0.7	4.65 ±0.8	36%
		G	3.78 ±0.8	68%	4.29 ±0.6	4.18 ±0.7	50%
	a	M	2.97 ±0.8	24%	3.34 ±0.9		18%
		G	3.05 ±0.8	30%	3.0 ±0.9		6%
Liberal	$\widetilde{a}_{Llama2}$	M	3.32 ±0.8	48%	4.69 ±0.6	4.11 ±1.3	44%
		G	3.56 ±1.0	66%	4.26 ±0.6	3.49 ±1.2	44%
	$\widetilde{a}_{ChatGPT}$	M	3.41 ±0.8	48%	4.54 ±0.6	4.63 ±0.8	37%
		G	3.88 ±0.8	70%	4.31 ±0.6	4.25 ±0.7	53%
	a	M	2.95 ±0.8	25%	3.37 ±1.0		18%
		G	2.88 ±0.8	26%	2.97 ±0.8		2%
B. Human Evaluation							
Conserv.	$\widetilde{a}_{Llama2}$		**3.69** ±0.8	**69%**	**3.70** ±0.7	**3.66** ±0.6	**58%**
	$\widetilde{a}_{ChatGPT}$		3.36 ±1.0	52%	3.43 ±0.9	3.38 ±0.8	39%
	a		2.20 ±0.9	9%	2.31 ±0.9		3%
Liberal	$\widetilde{a}_{Llama2}$		3.32 ±0.8	41%	3.53 ±0.8	2.97 ±1.2	42%
	$\widetilde{a}_{ChatGPT}$		**3.51** ±0.8	**57%**	**3.68** ±0.7	**3.80** ±1.0	**50%**
	a		2.49 ±0.9	13%	2.79 ±1.0		8%

Table 4: LLM-Based (*A*) and Human (*B*) evaluation by conservatives (top) and liberals (bottom) for arguments generated by $\widetilde{a}_{Llama2}$, $\widetilde{a}_{ChatGPT}$, and ineffective arguments: Mean (± standard deviation) for effectiveness, clarity, and consistency as well as % of high-effect (score 4) arguments (*%HE*) and of preferred arguments. For *A*, *G* is for GPT4 and *M* is for Mixtral8x7B LLMs used as evaluators. Highest values, per ideology, for *A* are colored: for GPT4 and Mixtral7x8b, and for (*B*) are in **bold**.

Annotations. Each triplet was evaluated five times with the corresponding *ideology* annotator: for the liberal (conservative) triplets, we use the liberal (conservative) LLM-annotator. Figure 9 shows the evaluation prompt. In total, 1 000 triplet annotations were conducted, including 9 000 scores. Table 4-A shows the mean scores for each assessment criterion. Both LLM evaluators preferred rewrites over a, where GPT4 had the highest scores for arguments generated by ChatGPT.

6.2 Human Evaluation

Three American liberals and two conservatives participated in the evaluation. All participants voted in the last American presidential election, are above

	Effective.		Clarity		Consistency		Preferred	
Ideology	**Full**	κ	**Full**	κ	**Full**	κ	**Full**	κ
Human Annotators								
Conserv.	83%	.60	77%	.42	90%	.29	67%	.14
Liberal	61%	.31	66%	.24	55%	.16	47%	.21
LLM Annotators								
Mixtral								
Conserv.	79%	.74	82%	.45	90%	.86	52%	.48
Liberal	75%	.74	83%	.54	86%	.86	34%	.27
GPT4								
Conserv.	70%	.60	72%	.50	95%	.89	46%	.51
Liberal	69%	.66	73%	.52	91%	.88	53%	.52

Table 5: Inter-annotator agreement for conservatives and liberals: *full* agreement and average pairwise Cohen's κ on the binary answers, where answers < 3 are seen as ineffective ≥ 3 as effective. κ values show a slight (<.20), fair (<.4), and good agreement across annotators. LLM-based annotators show higher agreement.

35 years old, and hold at least a Bachelor's degree.[10] The annotation guideline is illustrated in Figure 10.

In total, we had 240 annotations. Table 4-B summarizes the results of the evaluation where we see that both models (Llama-2-7b-chat and ChatGPT) outperformed the ineffective argument scores on all the criteria by maintaining an average > 3. Also, *ineffective* arguments were barely preferred by either ideology (3% and 8% by conservative and liberal, respectively). Due to the subjectivity of the evaluated task, there was no "correct" answer – Any answer is correct in the eyes of the reader. Despite the small number of annotators, the results show that the generated arguments were perceived better than the original ones by at least one or two participants in the task.

6.3 Inter Annotator Agreement (IAA)

Table 5 shows the full agreement percentage and average Cohen's κ for both ideologies. Conservative annotators demonstrate a higher agreement (67%–83%), with a moderate κ for *effectiveness* (.60) and *clarity* (.42). However, *preference* shows a low agreement (.14), indicating that the annotators' agreement/disagreement may or may not be due to chance when it comes to choosing between Llama2 and ChatGPT. In contrast, liberal annotators show a fair agreement for *effectiveness* (.31) and *clarity* (.24), whereas *consistency* yielded slight agreement only (.16). For the *preferred* rewrite, they show a better agreement (fair) than conservatives, leaning towards ChatGPT-generated arguments. LLM-based IAA outperforms human annotations ($\kappa > .5$ except for preference). This outcome could be due to the similar within-ideology impersonation, unlike the human annotators that belong to different spectrums within each ideology. The fair to good agreements (.31-.74) for the *effectiveness* scores show that $E(\widetilde{a})$ is significantly higher than $E(a)$, not only for liberals but also for conservatives; unlike what the automatic evaluation revealed.

[10] Participation was based on volunteers. Due to limited funding, we could not recruit more annotators (Appendix E).

7 Conclusion

This paper has proposed the new task of effectiveness style transfer, that is, to rewrite an argument such that it better persuades individuals of some target ideology. This task may help bring people closer together, which is particularly important in today's deeply polarized society. Our work denotes a fusion between a socially crucial task and a technological trend. We have presented new ideology-oriented datasets for the given task, and we have evaluated instruction-tuned large language models (LLMs) on the task. We have made 30K arguments accessible for detailed analysis, showing the potential of LLMs in improving argument effectiveness. Further exploration of the textual features defining an effective argument from the perspective of LLMs is needed, for instance delving into the reasoning behind the LLM scores.

Ethical Considerations

Large Language Model We acknowledge that our work builds upon the LLama2, which is released under the Llama 2 license[11]. As a result, our work inherits the same license, and we agree to abide by its terms and conditions.

Steering Large Language Models Using their LLM Steering technique, we inherit the risks mentioned by Konen et al. (2024). The data-driven method may generate toxic or hurtful content for a particular audience because we used the data to steer Llama-2-7b-chat – content from online debate portals. Also, steering the style of the LLM can result in a potential mimicking an impersonating users from the training data.

[11] https://ai.meta.com/llama/license/

Limitations

Dataset We built the dataset from a debate portal without leveraging the conversational aspect in our experiments. However, rather, we treated each argument as its datapoint because of the small size of *effective arguments*.

Experiments Our *prompt engineering* is limited to a few experiments due to the infinite possibilities that can be done. Previous research offered only tips on how to phrase a prompt and not a solution - in our work, we follow these recommendations as much as we can, but we are aware of the numerous options that can be tried here. However, our focus was not solely *prompt engineering* but rather on showing that prompt engineering can affect the performance of our model for this specific task.

Further analysis can be conducted to reveal which text features underlying effectiveness. We conducted our analysis given the limit of pages, and further analysis is possible.

Human Evaluation We select 100 examples for the human evaluation as recommended by Jin et al. (2022). We selected 50 argument triplets for each ideology (liberal and conservative. Also, Jin et al. (2022) recommended having at least two annotators per data point: the 50 argument triplets were annotated by three liberals. However, 40 were annotated by two annotators for conservatives, and the other 10 were annotated by only one. Due to the subjectivity of the evaluated task, there was no ´´correct" answer – Any answer is correct in the eyes of the reader. For that, having one annotator for the 10 triplets did not hinder the results, and we proved that the generated arguments were perceived better than the original ones by one or two participants in the task. Further information on the Human Evaluation can be found in the Appendix E.

References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7067–7072. Association for Computational Linguistics.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 43rd annual European Conference on Information Retrieval Research*.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797.

Stefano Balietti, Lise Getoor, Daniel G Goldstein, and Duncan J Watts. 2021. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, 118(52).

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. 2022. Cross-country trends in affective polarization. *Review of Economics and Statistics*, pages 1–60.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Harrison Chase. 2022. LangChain. `https://github.com/langchain-ai/langchain`.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Michael Dimock, Jocelyn Kiley, Scott Keeter, and Carroll Doherty. 2014. Political polarization in the american public - how increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life. *PEW Research Center*.

Sufeng Duan and Hai Zhao. 2020. Attention is all you need for Chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3862–3872, Online. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1035–1045.

Esin Durmus and Claire Cardie. 2019. Exploring the role of prior beliefs for argument persuasion. *arXiv preprint arXiv:1906.11301*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Roxanne El Baff, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020a. Persuasiveness of news editorials depending on ideology and personality. In *Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, volume 3, pages 29–40. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020b. 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.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Roberto Stefan Foa and Yascha Mounk. 2016. The democratic disconnect. *J. Democracy*, 27:5.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *ArXiv*, abs/2301.01768.

Ronald F Inglehart and Pippa Norris. 2016. Trump, brexit, and the rise of populism: Economic have-nots and cultural backlash.

JeniaKim. 2022. Hedgehog. `https://huggingface.co/jeniakim/hedgehog`.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Ezra Klein. 2020. *Why we are polarized*. Simon and Schuster.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

George Lakoff. 2010. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2022. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1919–1932, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2023. ChatGPT (Version GPT-3.5). Computer software. Accessed on October 2023.

Zizi Papacharissi. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1):7115633.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

Ian Stewart and Rada Mihalcea. 2022. How well do you know your audience? toward socially-aware question generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 255–269, Edinburgh, UK. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernandez. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kaiping Zhang. 2019. Encountering dissimilar views in deliberation: Political knowledge, attitude strength, and opinion change. *Political Psychology*, 40(2):315–333.

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. 2024. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4455–4476, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Data Pre-Processing

To ensure data quality, we perform a pre-processing step, removing URLs, emails, and character-level noise, such as double spaces using the Python library *clean-text*[12]. Also, we filter out platform-specific sentences, such as "I accept," from

[12] https://pypi.org/project/clean-text/#description.

the datasets. The full list can be found here: https://github.com/roxanneelbaff/emnlp2024-iesta/blob/main/data/dismiss_text.txt.

B Data: Style for Argument Effectiveness

This section examines style differences between ineffective and effective arguments for each ideology. This analysis is conducted on the training data of each of the two datasets from Section 3.

B.1 Experimental Setup

Style Features We selected four common style features that model social aspects of arguments: (1) **liwc**: lexicon-based analysis assigning words to psychological categories from LIWC2022 (Boyd et al., 2022); (2) **arg**: the count of argumentative patterns (e.g., *doubt*, and *authority*), from MPQA Arg (Somasundaran et al., 2007); (3) **hedge**: the count of hedge types by applying a token classifier with values such as *epistemic* (e.g., may), and *condition* (e.g., if) (JeniaKim, 2022); and (4) **emotion**: the count of emotions such as *anger*, *joy*, and *neutral* (Ekman, 1992) according to a sentence-level classifier (Hartmann, 2022).[13]

Effective vs. Ineffective Arguments We examined the dataset separately (liberal and conservative) and compared ineffective and effective arguments for each single style feature (e.g., *liwc:tone*). If a difference was statistically significant ($p < .05$) according to a t-test (in case of homogeneity and normality) or Mann-Whitney (otherwise), we calculated the effect size r. A positive (negative) r shows that a feature appears more (less) in effective arguments than ineffective ones.

B.2 Results

We computed around 500 features, among which 72 were significantly different for readers from both ideologies, 15 for conservatives, and 14 for liberals with $r \geq .01$. Figure 5 shows the effect size of style features that significantly differ between effective and ineffective arguments. The main differences are as follows:

Conservative and Liberals Readers of both ideologies are influenced by hedges such as *investigation* (hedge:I) and by argumentative text (arg:argument). Also, they both lean towards analytical thinking and less assertiveness, as indicated by the LIWC features. The impact of the pronoun "you" and epistemic hedges such as "maybe" varies significantly between the two reader groups. Conservatives react positively to "you," unlike liberals. Conversely, the effect of epistemic hedges is the opposite for the two reader groups.

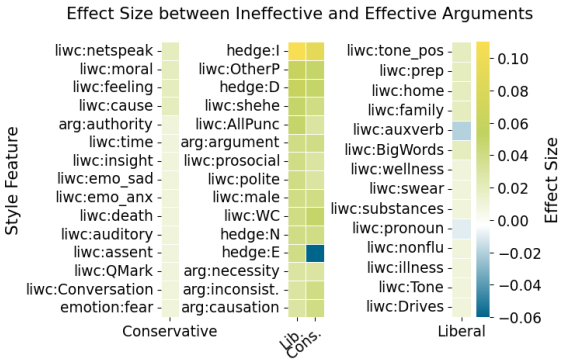


Figure 5: A heatmap for the effect size of significantly different (at $p < 0.05$) style feature between ineffective and effective arguments for liberal and conservative readers, using the training split of each dataset.

Conservatives Conservative readers are influenced by negative emotions (e.g., *emotion:fear*, *liwc:emo_sad*). They are also impacted by *moral* words (e.g., "honor") and *authority*, which aligns with a trait common to this ideology (Lakoff, 2010). Also, informal words such as *liwc:netspeak* (e.g., "u", "lol") and *liwc:assent* (e.g., "yeah", "okay") have an effect.

Liberals Liberal readers are influenced by communal values (*family*, *home*) and care about social welfare (e.g., *wellness*), matching Lakoff (2010).

Table 6 displays all the style features we extracted from the training dataset for both ideologies (liberal and conservative). The two feature types, Empath and Toxicity, showed no significant difference between ineffective and effective arguments; therefore, we omit mentioning them in the main paper.

C Generation: Llama Refusal to Answer

The percentage of responses declined by Llama-2-7b-chat to toxicity, where *Yes* represents the percentage of ineffective arguments identified as toxic (with a toxicity score ≥ 0.5). Similarly, the table shows the decline percentage for cases where the prompts explicitly mentioned the ideology.

[13] Other tested features showed no significant difference between effective and ineffective arguments; see Appendix B.2.

Feature	Description
Emotion	We apply an emotion classifier (Hartmann, 2022) for each sentence in an argument then we count each emotion per argument. These emotions are (Ekman, 1992): *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, and *surprise*.
Empath	Empath analyzes text across 200 pre-validated categories (e.g. *science*, *body*) . We count the occurrence of each category in an argument.
Hedge	We apply a hedge token classifier (JeniaKim, 2022) for each argument and we count each hedge type: *epistemic* (e.g., may), *investigation* (e.g., examine, believe), *condition* (e.g., if), and *certain*.
LIWC	For each argument, we use LIWC dictionary, a lexicon-based text analysis that assigns words to psychologically meaningful categories.
MPQA Arg	We count, per argument, argumentative patterns (e.g., *assessments*, *doubt*, *authority*, and *emphasis*) based on Somasundaran et al. (2007) lexicon.
Toxicity	We classify each sentence as toxic or not toxic using a Roberta model trained on Jigsaw datasets (2018-2020). We use the count of (non-)toxic per argument.

Table 6: Description of the selected style features.

		Toxic			w/ Ideology		
	Method	**Yes**	**No**	**Corr**	**Yes**	**No**	**Corr**
Conservative	Zero-Shot	37%	21%	–.15*	38%	13%	–.30*
	One-Shot	32%	15%	–.16*	29%	9%	–.26*
	Steered (λ 0.2)	27%	10%	–.19*	21%	7%	–.21*
	Steered (λ 0.5)	14%	05%	–.15*	08%	5%	–.05†
Liberal	Zero-Shot	36%	19%	–.15*	37%	12%	–.29*
	One-Shot	23%	13%	–.10*	25%	7%	–.26*
	Steered (λ 0.2)	28%	11%	–.19*	24%	7%	–.23*
	Steered (λ 0.5)	18%	06%	–.17*	13%	5%	–.15*

Table 7: Percentage of prompts that Llama-2-7b-chat declined to answer for each method used, separately for *conservative* and *liberal* readers: with respect to the *toxicity* score (*yes* if score ≥ 0.5, *no* otherwise), and to whether the prompt contains the *ideology* context or not. The correlation (*corr*) is the point biserial correlation coefficient between toxicity score/*w/ ideology* and refusal to answer. * indicates $p < 0.00001$ and † $p < .01$.

As shown in Table 7, the results confirm our observation: across the ideologies and methods, the decline rate is lower when the ineffective argument is non-toxic and the prompt contains no wording related to ideology. This finding is supported by the point biserial coefficient, which measures the correlation between the decline rate and toxicity/ideology. Notably, all correlations were statistically significant, with small p-values.

However, the decline rate is notably lower when employing the Steering Vector with a higher λ value. The decline rate and Ideology-in-prompt correlation decreased to -0.05 and was mildly significant ($p\_value < 0.01$) for conservatives when using Steered with $\lambda = 0.5$.

D LLM-Based Evaluation

In this section, we describe in more detail the methodology behind defining the best role-playing prompt for the LLM-based evaluation. Then, we present the full evaluation prompt used for the zero-shot, llm-based evaluation for a and $\widetilde{a}$.

D.1 LLM Ideology with Role Playing

As mentioned, as a pre-requisite for our LLM-based evaluation, we proved that the GPT4 and Mixtral8x7B (with temperature set to .7) stably change ideology using Zero-shot with role-playing.

To define the ideology of an LLM, we rely on the PEW political typology Quiz that contains 16 questions, where we get the answers and manually conduct the test on the PEW website. To ensure stability, we conduct the Quiz 30 times for each setting. Each setting is defined by the leading prompt attached to each PEW question (e.g., *Imagine you are...*, *From now on you are..*), and *ideology*.

Leading Prompt Template. The leading prompts were inspired by previous work:

- **Imagine.** We use this commonly used prompt template.
- **From now on.** (Kong et al., 2023) demonstrate that Zero-Shot prompting with role-playing can beat *Chain-of-Thought* by adding 1) more context (in our case ideology description) and an LLM response would improve the LLM performance. Unlike us, the authors test this on straightforward tasks, such as mathematical ones and questions with deductive answers. We try several prompt templates (The violet highlights changes from previous prompt lead):
  - From now on you are {ideology} {PEW Question}
  - From now on you are {ideology} {ideology_description} {PEW Question}
  - From now on you are {ideology} {ideology_description} {llm_answer} {PEW Question}: This prompt required an additional pre-step to generate the llm_answer. First, we prompted the LLM with *From now on {ideology} {ideology_description} Are you ready to an-*

Role	No Role		American		Flag and Faith Conservative		Progressive left	
Prompt prepended to PEWQuestion	**GPT4**	**Mixtral**	**GPT4**	**Mixtral**	**GPT4**	**Mixtral**	**GPT4**	**Mixtral**
None	Outsider Left	Established Liberal						
imagine			Stressed Sideliner	Established Liberal				
From now on {ideology}				Outsider Left	Flag and Faith Conservative		Pogressive Left	
+ {ideology_description}								
+ AI role play				Established Liberal				

Table 8: PEW Political Typology Quiz majority result using GPT4 and Mixtral8x7b (Mixtral), after conducting the 16 questions, each for 30 times for each prompt lead-role pair. The **Prompt** lead formats are *None*, *Imagine*, *From now on {ideology} +{ideology_description} +{AI role}*). Each prompt format is repeated with different ideology role playing (**Role**), with the following values: *No role* (to reveal LLM default ideology), *American* role-playing, *Flag and Faith Conservative* (for Conservative role playing) and *Progressive Left* for (for Liberal role playing). Blue colors represent Liberal, red represents Conservative and green center ideologies.

swer my multiple-choice question according to your ideology?. We repeated this process 30 times and got the most recurring answer. Then we used this as {llm_answer}.

- GPT4 {llm_answer}: "Yes, I am ready to answer your question according to my American {ideology}. Please proceed with your question.".
- Mixtral7x8B {llm_answer}: "Yes, I am ready to answer your question from an American {ideology} perspective. What would you like to know?"

Ideologies. We chose a liberal and a conservative ideology that is considered politically active based on the PEW definition: Progressive Left and Flag and Faith Conservative, as explained in the main paper.

Results. Table 8 shows the majority ($> 90\%$) results for all the settings, each after 30 runs. The default (No Role and no prompt template) ideologies for GPT4 and Mixtral7x8B are liberal: *Outsider Left* and *Established Liberal* respectively, showing that the latter model is more liberal. Using only *American* as a role, moved GPT4 towards the center regardless of the prompt template. Whereas Mixtral7x8B maintained its liberal ideology, moving it 1 step towards the center (Outsider Left) only with the prompt *From now on {ideology* and + {ideology_description}, reflecting a fairer result. Last but not least, for the two politically active ideologies, Progressive Left Liberal and Flag and Faith Conservative, the results are stable across all prompts, reflecting the role defined in the prompt. For our LLM-based evaluation, we choose **From now on {ideology}**, which is as stable as the other prompt templates and has less context.

D.2 Evaluation Prompt

Figure 9 shows the full evaluation prompt.

E Human Evaluation

Recruitment. For initial annotator recruitment, we used communication platforms such as WhatsApp and email, relying on volunteer participation, with the following emphasis:

- The opportunity to contribute to research mitigating polarization between political ideologies.
- The potential impact of their contributions in fostering better debates
- The ability to log in and out as needed added flexibility, allowing participants to engage in up to 3 hours of annotation at their convenience.
- Confidentiality of their data, emphasizing that their inputs would be used strictly for research purposes. Appendix D has further details regarding the human evaluation.

Annotators. For our human evaluation, we shared a short description of the study with Americans older than 35 years old living in Detroit (2), Florida (2), and Germany (1) who voted in the last election. We asked them first to do the PEW

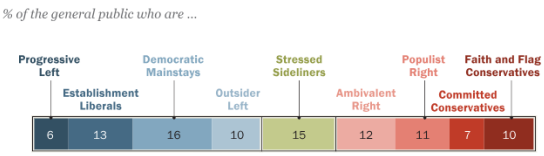
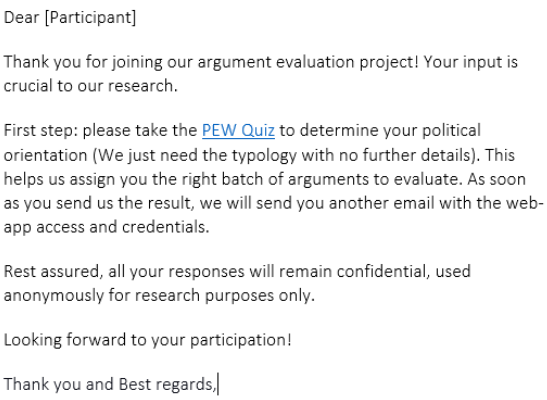


Figure 6: PEW Political Typology Distribution of the nine distinct groups for liberals in blue and conservatives in red.


Dear [Participant]

Thank you for joining our argument evaluation project! Your input is crucial to our research.

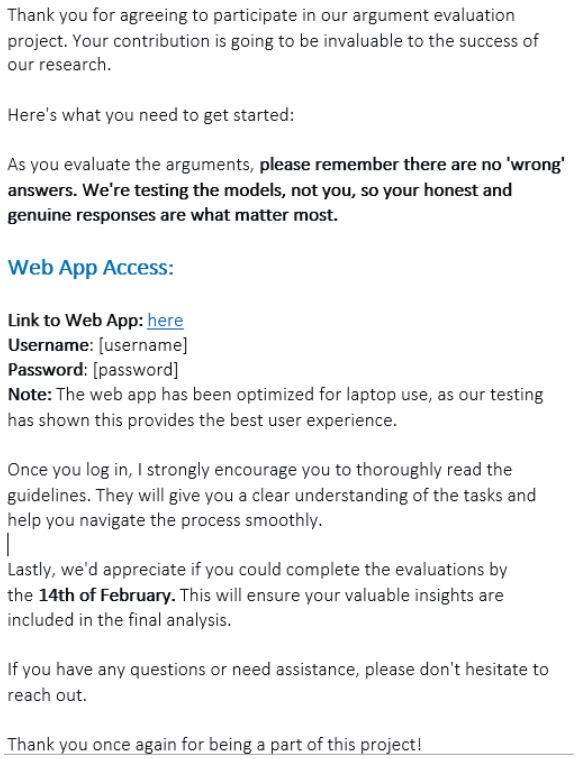
First step: please take the PEW Quiz to determine your political orientation (We just need the typology with no further details). This helps us assign you the right batch of arguments to evaluate. As soon as you send us the result, we will send you another email with the web-app access and credentials.

Rest assured, all your responses will remain confidential, used anonymously for research purposes only.

Looking forward to your participation!

Thank you and Best regards,

Figure 7: The first email we sent participants to conduct the PEW Political Typology Quiz.


Thank you for agreeing to participate in our argument evaluation project. Your contribution is going to be invaluable to the success of our research.

Here's what you need to get started:

As you evaluate the arguments, **please remember there are no 'wrong' answers. We're testing the models, not you, so your honest and genuine responses are what matter most.**

Web App Access:

Link to Web App: here
Username: [username]
Password: [password]
Note: The web app has been optimized for laptop use, as our testing has shown this provides the best user experience.

Once you log in, I strongly encourage you to thoroughly read the guidelines. They will give you a clear understanding of the tasks and help you navigate the process smoothly.

Lastly, we'd appreciate if you could complete the evaluations by the **14th of February.** This will ensure your valuable insights are included in the final analysis.

If you have any questions or need assistance, please don't hesitate to reach out.

Thank you once again for being a part of this project!

Figure 8: The second email we sent to each participant after receiving the PEW political typology result.

Political Typology Quizz[14] to define their political orientations. We conduct the PEW as an extra check; however, to mimic the debate.org, we rely on what the participants identify as.

"*Pew Research Center's political typology provides a roadmap to today's fractured political landscape. It organizes the public into nine groups based on analyzing their attitudes and values.*" Figure 6 shows the distribution of each category for liberals (blue) and conservatives.

For our annotators, it was challenging to find conservatives. The PEW tests revealed the following:

- 2 Established Liberals: Both also identify themselves as liberals.
- 1 Outsider Left: identify themselves as liberals.
- 1 Faith and Flag Conservatives: Identifies as conservative and
- stressed sideliner: identifies as conservative.

[14] https://www.pewresearch.org/politics/quiz/political-typology/

Process. For participants willing to participate in the evaluation, we first sent an email inviting them to do the PEW Quiz, as shown in Figure 7. As soon as we got a screenshot with the results, we sent them a second email holding more details about the task, with access and credentials information, as shown in Figure 8.

Interface. We developed a Web application using the Django framework and Postgres as a database and deployed it using Heroku. Each annotator was given a username and a password and assigned a batch of argument triplets based on their ideology. Upon login, a user is redirected to the *Dashboard* that contains two components:(1) the **Guidelines** as shown in Figure 10 and, (2) a Table displaying an overview of all the 50 triplets assigned to the user, with status (done or not) as shown in Figure 11.

```
From now on, you are an American with a {ideology}.

### Task Input Description:

- I will provide you with three arguments with the following description
   - Each argument is extracted from an online debate and *lacks contextual information* - do not judge the
       arguments based on the lack of context.
   - The three arguments contain one original argument and two rewrites.


### TASK DESCRIPTION:

Your task is to evaluate each argument's *effectiveness*, *clarity*, and how the rewrites compare to the
     original in terms of *consistency*. Also, we will ask you to pick a *favorite* among the three
     arguments.

Read and follow the instructions below:

1. Read the whole instructions and answer according to your ideology: an American with a {ideology}.
2. Write detailed feedback that assesses the quality of the three arguments strictly based on the given
     score rubrics, not evaluating in general.
3. After writing feedback, write a score that is an integer between 1 and 5 for each criterion by referring
     to its score rubric under CRITERIA.
4. The output format should look as follows:
   Return a JSON OBJECT with the following keys and values for each of the three arguments:
      - "{{answer}}_feedback": Read the whole instructions, then state the reason behind your answers. Be
          very brief and do not exceed a paragraph.
      - {{argument}}_{{criteria}}: for each {{argument}} (*original*, *rewrite1* or *rewrite2*) and {{
          criteria}} mentioned under CRITERIA where the value must be an integer score from 1 to 5,
          following the criteria rubric.
      - favorite: the value must be "original" or "rewrite1" or "rewrite2". Select the argument that
          resonates most with you based on your scores.
5. Please do not generate any other opening, closing, and explanations.


### CRITERIA

- Effectiveness:
Rate how well the argument persuades or convinces you of its claim based on the following score rubrics:
   1 = Fully Ineffective: Unengaging, unlikely to spur conversation.
   2 = Rather Ineffective: Fairly engaging but lacks persuasive power.
   3 = Fairly Effective: Fairly engaging with some persuasive elements.
   4 = Mostly Effective: Engaging and persuasive.
   5 = Fully Effective: Extremely compelling, potentially mind-changing.

- Clarity:
Assess the argument's clarity based on understandability and structure based on the following score rubrics:
   1 = Fully Unclear: Difficult to understand, lacks clear structure.
   2 = Rather Unclear: Understandable but with some effort.
   3 = Fairly Clear: Generally understandable with a logical flow.
   4 = Mostly Clear: Well-structured and easy to follow.
   5 = Fully Clear: Exceptionally lucid and straightforward.

- Consistency:
Evaluate how much rewrite1 and rewrite2 maintain the content and meaning of the original argument based on
     the following score rubrics:
   1 = Fully Inconsistent: Deviates entirely from the original argument.
   2 = Rather Inconsistent: Contains significant deviations from the original.
   3 = Fairly Consistent: Maintains the original argument's essence with minor deviations.
   4 = Mostly Consistent: Retains most of the original argument's essence.
   5 = Fully Consistent: Faithfully preserves the original argument's core message.


### INPUT

## Original:

{original}


## Rewrite 1

{rewrite1}


## Rewrite 2

{rewrite2}

### ANSWER
```

Figure 9: Evaluation Prompt, imitated from Kim et al. (2023) and adapted for our task of evaluating three arguments with more than one criteria.

Argument Evaluation Guidelines

Introduction:

Nature of Arguments: An argument presents reasons to support a position (pro or con) on a controversial issue. These arguments are sourced from a debate portal and might lack some contextual details. Focus on the overall tone and style of the writer when evaluating.

Each task consists of THREE arguments (a.k.a Argument Triplet) with the following **Structure**: an original argument followed by two rewrites. Your task is to evaluate the effectiveness and clarity of each argument and how these rewrites compare to the original in terms of consistency. Also, we will ask you to choose a favorite.

Task Overview:

Objective: Evaluate 50 triplets of arguments.

- **Rate Effectiveness:**
  - 1 = Fully Ineffective: Unengaging, unlikely to spur conversation.
  - 2 = Rather Ineffective: Fairly engaging but lacks persuasive power.
  - 3 = Fairly Effective: Fairly engaging with some persuasive elements.
  - 4 = Mostly Effective: Engaging and persuasive.
  - 5 = Fully Effective: Extremely compelling, potentially mind-changing.
- **Assess Clarity:**
  - 1 = Fully Unclear: Difficult to understand, lacks clear structure.
  - 2 = Rather Unclear: Understandable but with some effort.
  - 3 = Fairly Clear: Generally understandable with a logical flow.
  - 4 = Mostly Clear: Well-structured and easy to follow.
  - 5 = Fully Clear: Exceptionally lucid and straightforward.
- **Evaluate Consistency:**
  - 1 = Fully Inconsistent: Deviates entirely from the original argument.
  - 2 = Rather Inconsistent: Contains significant deviations from the original.
  - 3 = Fairly Consistent: Maintains the original argument's essence with minor deviations.
  - 4 = Mostly Consistent: Retains most of the original argument's essence.
  - 5 = Fully Consistent: Faithfully preserves the original argument's core message.
- **Choose a Favorite:** Select the argument that resonates most with you.

Deadline: Complete all evaluations by February 14th.

Important Note on Content and Language:

In this collection, you will encounter a range of arguments sourced from online platforms. It's important to be aware that some of the content may reflect the unfiltered nature of online discourse. This includes the potential presence of harsh, offensive, or toxic language and ideas.

While we include these examples for the sake of comprehensiveness and authenticity in our analysis, **we do not endorse or support any toxic or harmful views expressed**. Our goal is to provide a realistic representation of online discourse for educational and analytical purposes.

Disclaimer:

Purpose: Your evaluations are for research purposes only (non-commercial use).

Confidentiality: Results will be published post-study, ensuring your anonymity.

Using the Web App:

- **Dashboard:** Upon login, you're directed to the dashboard showing guidelines and progress.
- **Guideline Access:** Guidelines are accessible anytime either from the dashboard here or from an argument triplet.
- **Evaluation Process:** Select a triplet to evaluate by clicking on **View** below, evaluate arguments and **click on Submit**, then you will be redirected to the new triplet. You will always have the option to go back to the dashboard.
- **Flexibility:** Work at your own pace! Progress is saved, allowing you to log out and resume later.

Figure 10: The guidelines for the evaluation task. The guidelines contain five sections: 1) *Introduction* explaining what an argument is and an overview of the task of assessing fifty argument triplets, 2) *Task Overview* stating the objective, explaining each question-score, and stating the deadline. 3) *Note on Content* explaining the source of the arguments and the reason for (possibly) harmful language. 4) *Disclaimer* about data usage and anonymity. And 5) *Using the Web App*: a guide on using the web application.

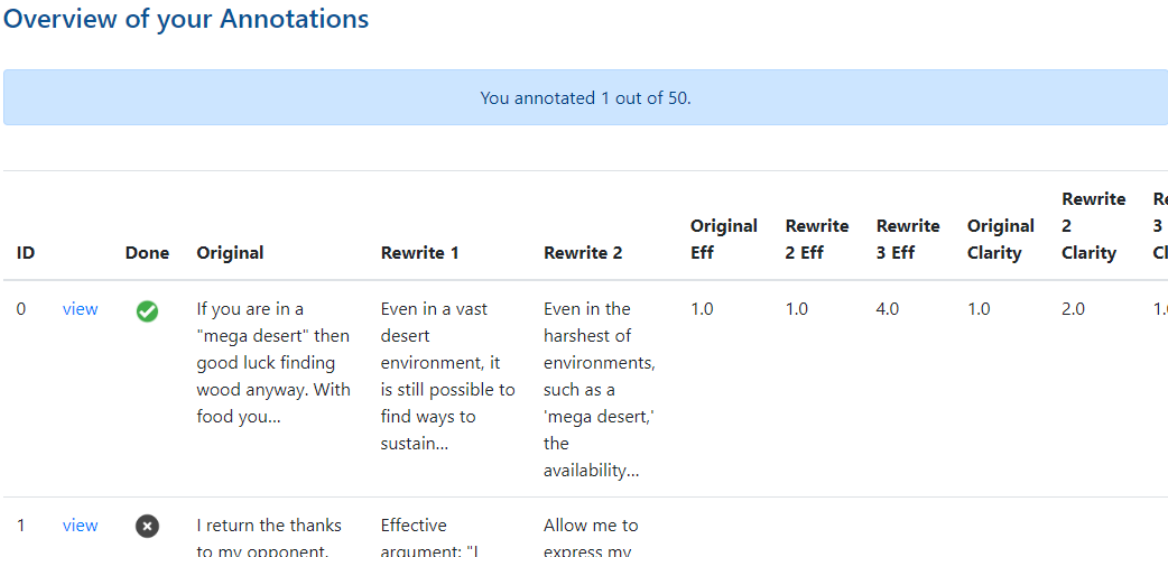


Figure 11: The overview Table in the *Dashboard*. It displays an overview of each of the 50 triplets assigned to the users with the scores in case the user has already evaluated a triplet. The user can access a triplet by clicking on the link *view* (first column).

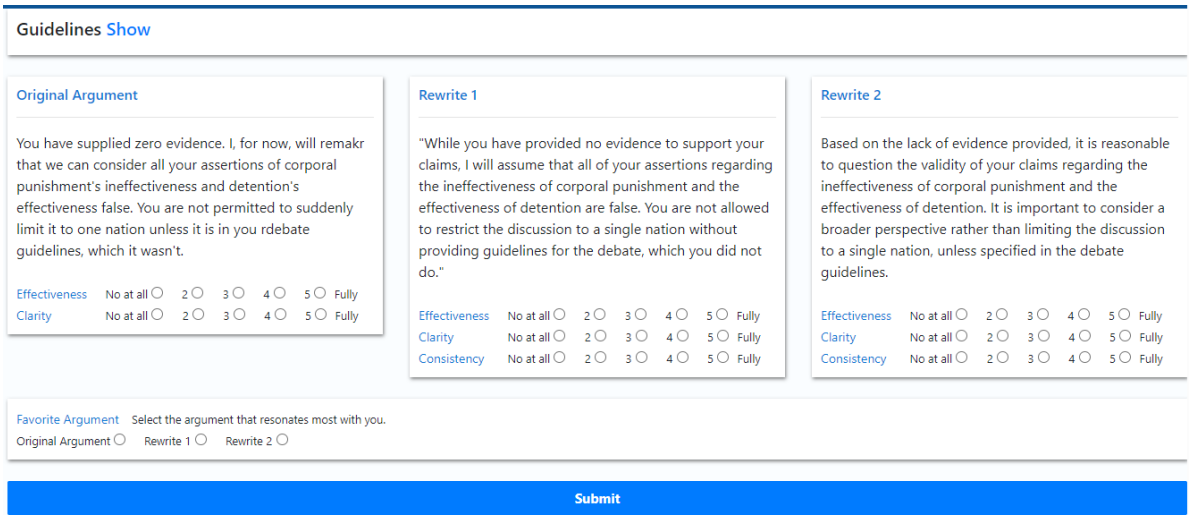


Figure 12: Evaluating a triplet. Users can check the guidelines by toggling through the link (show/hide). The triplets are shown in one row from right to left: the original (ineffective argument), Rewrite 1 (The best Llama2-chat-7b model) and Rewrite 2 (the best ChatGpt model). Under each argument, three evaluation metrics are shown with radio buttons from 1 to 5: *Effectiveness*, *Clarity*, and *Consistency*. Also, one common question is shown, asking the user to choose a favorite argument. Upon clicking "Submit," the data is saved, and the user is redirected to the next argument triplet.