LLM-based Rewriting of Inappropriate Argumentation using Reinforcement Learning from Machine Feedback

Timon Ziegenbein
Leibniz University Hannover
t.ziegenbein@ai.uni-hannover.de

Gabriella Skitalinskaya
Leibniz University Hannover
g.skitalinska@ai.uni-hannover.de

Alireza Bayat Makou
Leibniz University Hannover
a.bayat.makou@stud.uni-hannover.de

Henning Wachsmuth
Leibniz University Hannover
h.wachsmuth@ai.uni-hannover.de

Abstract

Ensuring that online discussions are civil and productive is a major challenge for social media platforms. Such platforms usually rely both on users and on automated detection tools to flag inappropriate arguments of other users, which moderators then review. However, this kind of post-hoc moderation is expensive and time-consuming, and moderators are often overwhelmed by the amount and severity of flagged content. Instead, a promising alternative is to prevent negative behavior during content creation. This paper studies how inappropriate language in arguments can be computationally mitigated. We propose a reinforcement learning-based rewriting approach that balances content preservation and appropriateness based on existing classifiers, prompting an instruction-finetuned large language model (LLM) as our initial policy. Unlike related style transfer tasks, rewriting inappropriate arguments allows deleting and adding content permanently. It is therefore tackled on document level rather than sentence level. We evaluate different weighting schemes for the reward function in both absolute and relative human assessment studies. Systematic experiments on non-parallel data provide evidence that our approach can mitigate the inappropriateness of arguments while largely preserving their content. It significantly outperforms competitive baselines, including few-shot learning, prompting, and humans.

1 Introduction

Creating trusted and safe online spaces where people with different backgrounds and opinions can discuss controversial issues is a major challenge for social media platforms (Salminen et al., 2018). The diversity in opinions, emotional attachments, and the anonymity of the web easily lead to heated discussions, which can quickly turn into toxic environments, even if only one participant behaves *inappropriately* (Habernal et al., 2018). Avoiding this is a challenging task, often supported by platform moderators that check content flagged by users or detection tools. However, the amount of moderation required on the web necessitates automation of the process, as the resources for manual moderation are usually insufficient, and the severity of inappropriate content can negatively affect the moderators' psyche (Spence et al., 2023).

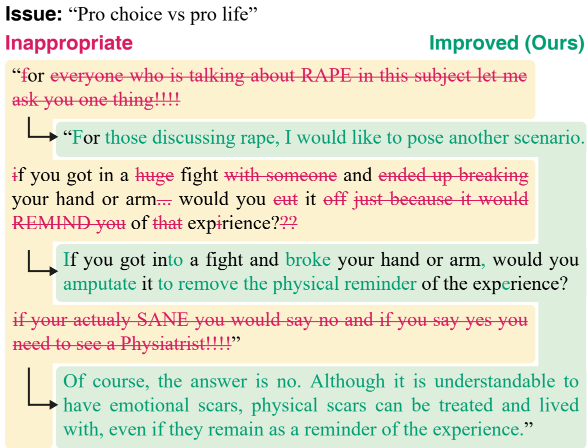


Figure 1: Example of an inappropriate argument from the corpus of Ziegenbein et al. (2023) and the same argument after applying our approach. The used colors indicate which parts of the original argument were removed (red strikethrough) and which parts were added by our approach in the rewriting process (green).

Multiple concepts and datasets have been proposed to model unwanted behavior in discussions, from simple offensiveness (Borkan et al., 2019) to sophisticated notions such as inappropriateness (Wachsmuth et al., 2017). The latter focuses on the exchange of arguments, their creation of credibility and emotions, and their adherence to the issue. Ziegenbein et al. (2023) argue that appropriateness displays the minimal quality of an argument necessary to be considered valuable in a debate.

This paper is the first to study how to rewrite inappropriate arguments automatically. Prior work studied the detection of unwanted behavior (Wulczyn et al., 2017; He et al., 2023; Ziegenbein et al.,

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4455–4476
August 11-16, 2024 ©2024 Association for Computational Linguistics

2023), often using large language models (LLM). While a few methods improve content, they solely transfer the style of texts to be more formal (Rao and Tetreault, 2018; Lai et al., 2021), less subjective (Pryzant et al., 2020; Liu et al., 2021a), or less toxic (Laugier et al., 2021; Logacheva et al., 2022), or they target the quality of arguments in general (Skitalinskaya et al., 2023). This commonly comes with preserving the original content and operating on single sentences. However, if the inappropriate behavior is rooted in the content itself and not only in the style of the text, content modifications on the document level may be necessary. In addition, most existing approaches rely on parallel data, which is unavailable for rewriting inappropriate arguments.

Instead, we propose an LLM-based rewriting approach to inappropriateness mitigation inspired by reinforcement learning from human feedback, RLHF (Christiano et al., 2017). Compared to the typical use of RLHF in NLP (Ouyang et al., 2022), the core ideas of our approach are: (1) We obtain a 'cheap' initial policy (an LLM to align) from either few-shot learning or prompting, rather than using supervised learning. (2) We specifically consider the properties on which we want to align the LLM, rather than relying on generic preference information. (3) We evaluate multiple candidate alignments with different weightings of the desired properties, rather than relying on a single alignment obtained from preference information.

After experimentally determining the performance of multiple LLMs using few-shot learning and prompting on the corpus of Ziegenbein et al. (2023), we find prompting the instruction-finetuned LLaMA (Touvron et al., 2023) variant of Taori et al. (2023) (Alpaca) to be best and thus proceed to align it further using our approach. We deem the desired properties for rewriting inappropriate arguments to be *semantic similarity* to the original argument and *appropriateness* of the generated argument, and we make use of existing classifiers to learn how to generate texts that fulfill them (Zhang et al., 2020; Ziegenbein et al., 2023).

Exemplarily, Figure 1 shows an inappropriate argument from a "Pro choice vs pro life" debate and the same argument rewritten by our approach. Here, the original argument uses overly excessive emotions, making it hard to understand, and it displays little interest in the opinion of others. The rewritten argument reduces emotions and adds a more open ending, making it more appropriate while keeping the original argument's gist intact.

For evaluation, we obtain human rewrites for a portion of the data and compare them automatically against our models and a competitive model from the literature. Moreover, we conduct relative and absolute human evaluations of the rewrites of our trained models and the human rewrites. We find that our approach successfully aligns LLMs according to the desired property weighting and produces the best rewrites. Intriguingly, our human annotators prefer appropriate rewrites, even if they are less semantically similar to the original arguments.

Altogether, this paper's main contributions are:[1]

- An RLHF-inspired approach for non-parallel data, based on instruction-finetuned LLMs aligned to specific classified properties.
- The first computational approach for rewriting inappropriate arguments.
- Empirical insights into human preferences regarding semantic similarity and appropriateness when rewriting inappropriate arguments.

2 Related Work

The notion of appropriateness in speech and argumentation, tied to cultural norms, social politeness, and context, originates from Aristotle's work on rhetoric (Aristotle, ca. 350 B.C.E./ translated 2007). It has been examined in various shades across linguistic studies (Hymes et al., 1972; Ranney, 1992; Schneider, 2012; Jdetawy and Hamzah, 2020). In debate, topic adherence and avoidance of offensive or biased language are considered aspects of appropriateness (Andrew, 1996; Blair, 1988; Walton, 1999; Burkett, 2011). Modeled by Wachsmuth et al. (2017) as a dimension of rhetorical argument quality, appropriateness has been partially explored in NLP, focusing on the simultaneous assessment of credibility, emotional engagement, and proportionality to the issue. Computationally, Wachsmuth and Werner (2020) initially attempted appropriateness prediction as a subtask of argument quality assessment. Later, Ziegenbein et al. (2023) refined the notion of appropriateness in argumentation, modeling it in a 14-dimensional taxonomy and predicting it together with its subdimensions in a multilabel setting. Wachsmuth et al. (2024) recently delineated how to instruct LLMs towards more reliable argument quality assessment, not targeting rewriting though.

[1]The corpus extension and experiment code can be found under: `https://github.com/webis-de/ACL-24`.

Related to argument rewriting, Skitalinskaya et al. (2023) studied how to improve the general quality of argumentative claims on data similar to what we use, but not focusing on inappropriateness specifically. Closest to our work in the context of style transfer tasks are Nogueira dos Santos et al. (2018), Laugier et al. (2021), and He et al. (2023), where toxic content is mitigated using supervised rewriting approaches. However, their methods are applied on the sentence level, strictly aim to preserve content, and prevent the addition of new content. Unlike them, we focus on document-level rewriting and explicitly consider adding or deleting content. Furthermore, most of these approaches require parallel data and rely on supervised learning. In contrast, our approach is meant for non-parallel data since no parallel dataset is available to learn to rewrite inappropriate arguments.

To this end, we use reinforcement learning as it allows us to train on non-differentiable metrics, such as the outputs of classifiers, which we use to test for the desired properties of our task. Reinforcement learning has been used for a variety of NLP tasks, including dialogue generation (Li et al., 2016), machine translation (Wu et al., 2018), and summarization (Ziegler et al., 2019; Böhm et al., 2019; Stiennon et al., 2020). In the context of style transfer, multiple approaches have aimed to flip the sentiment, stance, or polarity of texts in parallel (Sancheti et al., 2020; Liu et al., 2021b) and non-parallel settings (Xu et al., 2018; Gong et al., 2019; Wu et al., 2019; Luo et al., 2019).

Many works study the related task of formality transfer (Gong et al., 2019; Luo et al., 2019; Sancheti et al., 2020; Liu et al., 2021b; Lai et al., 2021) with diverse techniques on the parallel data of Rao and Tetreault (2018). Three properties are commonly controlled during transfer: fluency, content preservation, and transfer strength. Similar to our work, Madanagopal and Caverlee (2023) propose reinforcement learning to remove subjective bias in Wikipedia texts, modeling the reward as a weighted function of classifiers for style, fluency, and content preservation. However, unlike in formality transfer, parallel data for rewriting inappropriate arguments is neither available nor straightforward to acquire.

Consequently, our work focuses on non-parallel style transfer inspired by reinforcement learning from human feedback. We rely on the more stable proximal policy optimization (PPO) (Schulman et al., 2017) instead of the commonly used REINFORCE Monte Carlo policy gradient (Williams, 1992) due to promising recent results in NLP and the advances in the capabilities of language models to follow instructions (Taori et al., 2023; Wang et al., 2023). To our knowledge, only the work of de Langis et al. (2024), which was published on arXiv after the submission of our work, uses an RLHF-inspired PPO approach for a style transfer-related task in NLP. Our work is the first to investigate the use of prompting as an initial policy, avoiding the need for parallel data through pseudo-parallel data and applying RLHF-inspired PPO to non-parallel data.

2.1 Proximal Policy Optimization in NLP

As a basis for the presentation of our approach, we here shortly describe the intuition behind PPO in an NLP context. For a more formal description, we refer the reader to Zheng et al. (2023).

PPO learns a *model critique* (a value model), which estimates the expected cumulative future *reward* (a value) of a *state* (the generated text up to this point), together with a *policy* (e.g., an LLM). The "real" reward is based on the output of a *reward model* (e.g., a classifier) and can be any scalar value. The value model estimates the *advantage* (gain in reward) of performing a specific *action* in a state (generating a specific word given the text generated so far) over performing the current policy's suggested action. Specific actions are chosen based on sampling from the current policy (e.g., top-p sampling). The current policy is updated based on the advantages of the specific actions and the KL-divergence between the token-level distributions of the current policy and its updated version. The KL-divergence here improves stability during training, limiting the size of update steps.

In other words, when predicting the next word given a sequence of previously generated words, multiple fitting potential words are considered by sampling from the LLM. The difference in the long-term expected reward between each of them and the word with the current highest probability is used to update the LLM. This way, the LLM can be steered (aligned) to generate text that fulfills the desired properties represented by the reward model.

3 Approach

This section presents the approach that we propose to rewrite inappropriate arguments inspired by reinforcement learning from human feedback. Figure 2

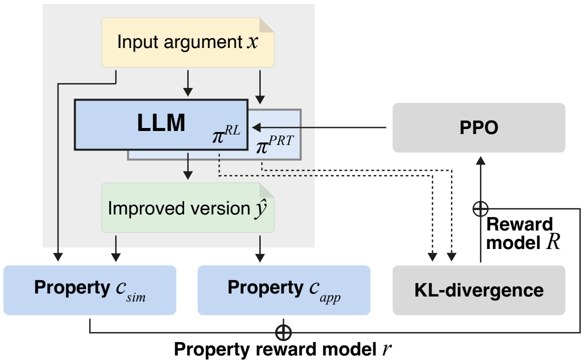


Figure 2: Our approach to rewriting inappropriate arguments: The policy π^{RL} is optimized using *PPO* to generate an improved version $\hat{y}$ from the input argument x while preserving the content of x as much as possible (c_{sim}) and making the argument more appropriate (c_{app}). This is based on reward R obtained from the weighting of r of the scalar classifier outputs and the *KL-divergence* between the initial policy π^{PRT} and the current π^{RL}. Dashed lines: The probability distribution over the tokens is used as the output of the LLM.

illustrates the main elements of our approach explained in the following.

3.1 Problem Formulation

Let x be an argument and $\hat{y}$ be an improved version of x. We define the task of *rewriting inappropriate arguments* as learning a function $f : x \mapsto \hat{y}$ such that $\hat{y}$ preserves the content of x as much as possible while being more appropriate.

3.2 Prompting as an Initial Policy

Usually, reinforcement learning from human feedback (RLHF) is used only to steer an LLM that has already learned to solve a task to a certain known extent instead of learning the task from scratch. However, we neither have access to such a model, nor data to train it. Instead, we thus propose to use prompting to obtain an initial policy $f \mathrel{\hat{=}} \pi^{PRT}$ (an LLM to start from) that solves the task to a certain (not immediately quantifiable) extent. We experimentally compare multiple autoregressive pretrained LLMs using zero-shot, few-shot, and instruction learning-based prompting to find an effective initial policy (further details on the models and prompting methods follow in Section 5). Then, we proceed to use the same prompts while learning π_ϕ^{RL}, where ϕ are the learnable parameters, to obtain a better version of the initial policy π^{PRT}.[2]

[2] In the following, we use the terms *policy* and *LLM* interchangeably as they refer to the same concept (π) in the context of reinforcement learning (RL) for LLMs.

3.3 Reward Modeling and Policy Learning

We initialize π_ϕ^{RL} with a pretrained LLM π^{PRT} that is prompted in natural language to generate $\hat{y}$ given x. Unlike Stiennon et al. (2020), we do not require learning a reward model r from human feedback on preference judgments. Instead, we model relevant properties that we desire to be present in the target output, $\hat{y}$. In particular, we assume that the semantic similarity (sim) of $\hat{y}$ to x and the appropriateness (app) of $\hat{y}$ are such relevant properties. We use pretrained classifiers c that estimate these properties as our reward model.

The property reward model r is thus defined as:

$$r(x, \hat{y}) := \alpha \cdot c_{sim}(x, \hat{y}) + (1 - \alpha) \cdot c_{app}(\hat{y}) \quad (1)$$

Here, $\alpha \in [0, 1]$ is a hyperparameter that controls the trade-off between semantic similarity and appropriateness. c_{sim} and c_{app} are classifiers that estimate the semantic similarity and appropriateness of $\hat{y}$ given x. Similar to Stiennon et al. (2020), to obtain the final reward model R, we penalize r with the KL-divergence between the initial policy π^{PRT} and the learned policy π_ϕ^{RL} to disincentivize moving away too far from π^{PRT}:

$$R(x, \hat{y}) := r(x, \hat{y}) - \beta \log \left[\frac{\pi_\phi^{RL}(\hat{y}|x)}{\pi^{PRT}(\hat{y}|x)} \right] \quad (2)$$

Here, $\beta \in \mathbb{R}$ is a hyperparameter that controls the strength of the KL-divergence. The policy π_ϕ^{RL} is optimized using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017).

We argue our setup to be beneficial for two reasons: (1) Pretrained classifiers that learned to assess specific properties from human labels are available for a wide variety of tasks; and (2) no training and consequently no data is required to learn an initial policy that solves the task to a certain extent.

4 Data

For our mitigation experiments, we extended the appropriateness corpus of Ziegenbein et al. (2023).

4.1 Source Data

The original corpus contains 2191 arguments and the corresponding discussion titles from three genres (reviews, discussion forums, and Q&A forums). Each argument has been annotated three times using a hierarchical 14-dimensional taxonomy of appropriateness flaws, such as *toxic emotions* or *missing intelligibility*. Here, we consider only the parent dimension *inappropriateness* to develop and evaluate our approaches. The corpus contains 1182 inappropriate and 1009 appropriate arguments.

4.2 Extension

We extend the given corpus by arguments from its original domains. In particular, we collected 73,703 arguments from the IACv2 Corpus (Walker et al., 2012; Abbott et al., 2016) and the GAQCorpus (Ng et al., 2020). We kept only those 55,290 arguments that have at least 10 and at most 220 words and that do not exceed 1100 characters. This way, we ensure that the arguments have approximately the same length as those in the original corpus. To avoid any topic leakage in the extended part of the corpus, we also remove arguments that belong to a topic already present in the original version of the corpus (49,417 arguments remaining). Finally, we soft-label all arguments in the extended part of the corpus with the five-fold ensemble classifier of Ziegenbein et al. (2023) to obtain appropriateness labels. These labels can then be used to train our approaches. 35,537 of the arguments were labeled as inappropriate and 13,880 as appropriate.

5 Experiments

This section describes the training procedure of our approach from Section 3 on the data fom Section 4 and the experiments we conducted to evaluate it.

5.1 Experimental Setup

As Ziegenbein et al. (2023), we split the data for evaluation into 70% training, 10% validation, and 20% test, ensuring an equal weighting of the 14 corpus dimensions. However, since we are interested in mitigating inappropriateness, we train and evaluate only on inappropriate arguments. During training, we exclusively use the inappropriate arguments from the corpus extension.

The advantage of this setup is two-fold: First, we can train on a large amount of data, which is often crucial for the success of reinforcement learning (RL). Second, it allows us to avoid propagating any selection bias that may arise from the intermediate step of finding the best initial policy to selecting the best checkpoint from our trained policy. We use the training set to select the best initial policy (details below), the validation set to select the best-performing RL policy checkpoints, and the test set to evaluate the performance of all approaches.

To automatically evaluate the generated rewrites, we employ the five appropriateness classifiers of Ziegenbein et al. (2023) trained on different folds of the data. As we use the entire appropriateness corpus for evaluation, we ensure the use of the classifier for each argument that has not seen the argument before as part of its training. We use a classifier to predict the appropriateness of the original argument and the generated rewrites. Then, we calculate the following performance values:

- *App.* Percentage of arguments for which an approach has flipped the prediction from inappropriate to appropriate;
- *Sim.* Semantic similarity of a rewrite to the input argument in terms of BERTScore (Zhang et al., 2020);
- *NES.* Normalized word-wise edit similarity (to quantify amounts of edits) (Lopresti, 1996);
- *PPL.* Fluency in terms of perplexity;
- *GM.* Geometric mean of *App.*, *Sim.* and *1/PPL* (to compare approaches using a single score).

We use the semantic similarity and normalized word-wise edit similarity to quantify if the generated arguments are indeed rewrites of the original argument and not any probably unrelated but appropriate text. Furthermore, we use perplexity as a measure of text coherence and fluency.

5.2 Finding an Initial Policy

Similar to Stiennon et al. (2020), we start with obtaining the initial policy, π^{PRT} that should have a reasonable performance mitigating inappropriate language. π^{PRT} is then aligned using our RL approach. Since no parallel data is available to train a supervised model, we prompt four LLMs of similar size (6-7 billion parameters) in a few-shot learning and instruction following-based setting:

- *OPT* (Zhang et al., 2022);
- *BLOOM* (Scao et al., 2023);
- *GPT-J.* (Wang and Komatsuzaki, 2021);
- *LLaMA.* (Touvron et al., 2023).

For the few-shot setting, we use 1, 4, and 9 examples to see how the performance changes with the amount of reference data. This setup is inspired by the hierarchical setup of the taxonomy of inappropriateness (Ziegenbein et al., 2023) having 1, 4, and 9 dimensions on the first, second, and third level respectively. To obtain instruction-finetuned versions of the models, we train each of them following the procedure suggested by Taori et al. (2023), to ensure we select the best base model and do not select

Model	App. ↑	Sim. ↑	NES. ↑	PPL ↓	GM ↑
Exact Copy	0.000	1.000	1.000	122.1	-
OPT	0.371	0.414	0.292	63.77	0.118
+ 1-shot	0.436	0.241	0.110	54.73	0.124
+ 4-shot	0.436	0.410	0.259	53.34	0.150
+ 9-shot	0.379	0.305	0.172	39.95	0.143
+ Instruct.	0.629	0.508	0.263	39.89	0.200
BLOOM	0.411	0.476	0.379	80.70	0.134
+ 1-shot	0.452	0.341	0.194	55.05	0.141
+ 4-shot	0.484	0.567	0.451	66.34	0.160
+ 9-shot	0.427	0.465	0.334	41.54	0.169
+ Instruct.	**0.653**	0.557	0.336	42.51	0.205
GPT-J	0.371	0.503	0.419	114.6	0.118
+ 1-shot	0.500	0.402	0.245	46.92	0.162
+ 4-shot	0.484	0.473	0.322	54.21	0.162
+ 9-shot	0.524	0.422	0.279	40.51	0.176
+ Instruct.	0.637	0.556	0.340	**37.49**	0.211
LLaMA	0.411	0.606	**0.528**	110.8	0.131
+ 1-shot	0.565	0.421	0.259	57.07	0.161
+ 4-shot	0.556	0.555	0.408	56.19	0.178
+ 9-shot	0.411	0.311	0.180	48.68	0.138
+ Instruct.	0.621	**0.620**	0.394	38.08	**0.216**

Table 1: Automatic evaluation of initial policies using zero shots, few shots (1, 4, 9), and instruction-finetuning: semantic similarity (Sim.), normalized edit similarity (NES.), perplexity (PPL), appropriateness (App.), and geometric mean (GM). The best results are marked bold.

a model because of a specific way it is prompted or the instruction data it was fine-tuned on. This way, we have a single fixed prompt for all models and can control the generation length and other parameters equally well.

Creating Few-Shot Examples Since no rewrites are available for our few-shot learning setup, we collect 14 rewrites (1+4+9) of inappropriate arguments from three NLP experts, none of whom are authors of this paper. To create rewrites that are highly representative of a dimension, we use the appropriateness corpus, sentence transformers (Reimers and Gurevych, 2019), and PageRank (Lawrence, 1998) (details in Appendix B).

Prompting Setup We employ the natural language prompts suggested by Reif et al. (2022) and Zhang et al. (2020) to generate $\hat{y}$ given x. The full prompts can be found in Appendix A.

Automatic Evaluation Table 1 shows the results of finding an initial policy. We observe *BLOOM + Instruct.* to create the most appropriate rewrites (0.653). *LLaMA + Instruct.* achieves the best results in terms of semantic similarity (0.620), and zero-shot *LLaMA* in terms of needing the minimal amount of edits to create its rewrites (0.528). In terms of fluency, *GPT-J + Instruct.* performs best (37.49), closely followed by *LLaMA + Instruct.* (38.08). Overall, *LLaMA + Instruct.* seems to be the most stable choice (GM 0.216), so we select it as the initial policy to train our final approaches. Regarding geometric mean, few-shot learning leads to a general increase in performance. However, no clear trend in the number of few-shot examples used is visible. Overall, rewriting based on instruction-finetuning mitigates inappropriateness best across all models.

Model	App. ↑	Sim. ↑	NES. ↑	PPL ↓	GM ↑
Exact Copy	0.000	1.000	1.000	98.01	–
CoEdIT	–	–	–	–	–
+ Paraphrase	0.320	0.668	0.357	39.61	0.175
+ Formal	0.356	0.683	0.478	42.98	0.178
+ Neutral	0.298	**0.876**	**0.857**	63.43	0.160
+ Polite	0.320	0.801	0.688	42.22	0.183
LLaMA + Instruct.	0.621	0.620	0.394	38.08	0.216
+ PPO_{app}	**0.960**	0.253	0.048	**21.26**	0.225
+ $\text{PPO}_{app>sim}$	0.933	0.359	0.114	28.50	0.227
+ $\text{PPO}_{app=sim}$	0.827	0.471	0.299	29.22	**0.237**
+ $\text{PPO}_{app<sim}$	0.373	0.808	0.731	44.41	0.189
Human Baseline	0.773	0.391	0.180	56.23	0.175

Table 2: Automatic evaluation of different policies for our approach LLaMA + Instruct., an alternative style transfer model (CoEdIT), and a human baseline: semantic similarity (Sim.), normalized edit similarity (NES.), perplexity (PPL), appropriateness (App.), and geometric mean (GM). The best results are highlighted in bold.

5.3 Policy Learning

Starting from our initial policy, *LLaMA + Instruct.*, we use PPO to learn a set of candidate policies. We use the rescaled version of BERTScore (Zhang et al., 2020) to estimate the semantic similarity of $\hat{y}$ to x and the appropriateness classifier from the first fold of Ziegenbein et al. (2023) to estimate the appropriateness of $\hat{y}$. Both values are in $[0, 1]$. We learn four candidate policies, each with a different property weighting $\alpha \in \{0.4, 0.5, 0.6, 1\}$. Indicating the property weighting used, we refer to the corresponding models as *LLaMA + PPO*$_{app<sim}$, *LLaMA + PPO*$_{app=sim}$, *LLaMA + PPO*$_{app>sim}$ and *LLaMA + PPO*$_{app}$ respectively. The exact setup of our PPO training and the hyperparameters used are detailed in Appendix C.

Baselines In addition to the learned policies, we collected human rewrites for each argument. We refer to these as *Human Baseline*. For this purpose, we hired five native English speakers on

Upwork.com, three male and two female. We instructed the annotators with background information about appropriateness and asked them to suggest rewritten versions of arguments flagged as inappropriate in a forum. Each annotator was asked to rewrite 45 of the 225 inappropriate arguments.[3]

In automatic evaluation, we also compare multiple settings of the *CoEdIT* model proposed by Raheja et al. (2023), including paraphrasing (*CoEdIT + Paraphrase*), formality (*CoEdIT + Formality*), neutrality (*CoEdIT + Neutral*), and politeness (*CoEdIT + Polite*) style transfer, to better understand the relationship to these related notions.

During the development of our approaches we also experimented with other common non-parallel style transfer models, such as TAG (Madaan et al., 2020) and LEWIS (Reid and Zhong, 2021), but found them to be unsuitable as they lack the linguistic quality of modern LLMs, putting them at a disadvantage from the get-go.

Automatic Evaluation We evaluate the candidate policies using the same automatic metrics as for the initial policy. Table 2 shows the results. We observe that our approach successfully manages to align *LLaMA + Instruct.* according to the desired property weighting with *LLaMA + PPO*$_{app}$ being best in terms of appropriateness (0.960), and *LLaMA + PPO*$_{app<sim}$ in terms of semantic similarity (0.808) among the LLaMA-based models. We find that none of the CoEdIT baselines can successfully mitigate inappropriateness, speaking for the distinctiveness of rewriting inappropriate arguments as a task. Overall, we find *LLaMA + PPO*$_{app=sim}$ to perform best (GM 0.237), even outperforming the human baseline (GM 0.175).

Manual Evaluation To enable comparison of various-automatically generated rewrites and human-suggested alternatives, we perform two manual evaluation studies. We again hire native English speakers on Upwork.com (7 female and 8 male) to evaluate the rewrites in absolute and relative terms, such that each rewrite (pair) is evaluated by five annotators. We again instructed annotators with background information about appropriateness. In total, we collected 4050 absolute and 8775 relative judgments.

In the first study, the annotators scored each rewrite regarding three considered quality metrics:

- *App.* Appropriateness of the topic discussion in terms of style and content;
- *Sim.* Meaning preservation of the original arguments;
- *Flu.* Fluency and adherence to grammar conventions.

We utilized a 5-point Likert scale to evaluate the level of success of a particular rewrite in meeting each quality metric requirement. Here, 5 indicated a strong agreement with the rewrite's success, and 1 strong disagreement. We calculate the final score for each rewrite using MACE (Hovy et al., 2013).

Table 3(a) presents the results of the conducted annotation study. We note that the obtained human judgments of appropriateness (*App.*) and semantic similarity (*Sim.*) across different desired property configurations are consistent with those obtained in the automatic evaluation (Table 2) for the trained models. Specifically, the *LLAMA + PPO*$_{app}$ achieves the highest appropriateness rating (3.77), while *LLAMA + PPO*$_{app<sim}$ combination attains the highest similarity rating (4.75). This alignment of human and automatic evaluations underscores the effectiveness of the chosen classification models in capturing the desired argument properties when mitigating inappropriateness. Overall, we find *LLaMA + Instruct.* to be the most balanced model (GM 3.57) following closely behind the human baseline (GM 3.63).

The goal of the second study is to rank the five LLaMA-based models and the collected human rewrites by perceived overall quality and appropriateness. To make the task more manageable for the human annotators, instead of requiring them to rank all six rewrites at once, we transform the annotation task into a pairwise ranking task and ask them to compare only two items at a time. Studies have shown that pairwise ranking tasks can lead to more reliable and consistent annotations compared to direct ranking tasks (Brun et al., 2010; Narimanzadeh et al., 2023) making them an effective and commonly used approach for subjective annotation studies, such as argument quality assessment (Habernal and Gurevych, 2016; Toledo et al., 2019; Skitalinskaya et al., 2021).

While pairwise rankings significantly increase the number of judgments to be made, making the task more time-consuming, Gienapp et al. (2020, 2022) have shown that efficient sampling strategies, such as Skip-window, can notably reduce the number of required pairwise annotations without

[3]Annotators were paid 15$ per hour. Guidelines and screenshots of the user interface are provided in Appendix F.

Model	(a) Absolute				(b) Relative							
	App. ↑	Sim. ↑	Flu. ↑	GM ↑	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Avg. ↓	p ↑
LLaMA + Instruct.	3.22	4.17	3.40	3.57	3.1%	5.3%	18.2%	21.3%	32.9%	19.1%	4.32	.351
+ PPO$_{app}$	**3.77**	2.65	**4.16**	3.46	44.9%	32.4%	13.3%	7.1%	2.2%	0.0%	**1.89**	**.833**
+ PPO$_{app>sim}$	3.50	2.96	3.77	3.39	29.3%	29.8%	18.7%	14.7%	5.8%	1.8%	2.43	.729
+ PPO$_{app=sim}$	3.15	3.38	3.34	3.29	2.7%	11.1%	22.2%	26.7%	20.9%	16.4%	4.01	.412
+ PPO$_{app<sim}$	2.70	**4.75**	2.89	3.33	0.4%	4.4%	4.4%	12.4%	26.7%	51.6%	5.15	.160
Human Baseline	3.60	3.48	3.82	**3.63**	19.6%	16.9%	23.1%	17.8%	11.6%	11.1%	3.18	.566

Table 3: Manual evaluation of our approach variations and the human baseline: (a) Absolute MACE scores of the improved arguments in terms of appropriateness (App.), similarity (Sim.), fluency (Flu.), and their geometric mean (GM). (b) Relative ranking of the arguments in terms of percentage of times they were ranked at each position (Rank 1–6), their average rank (Avg.), and the score obtained by the Bradley Terry model (p). Best results are marked bold.

compromising the quality of the final ranking. In our study, we employ the Skip-window with $\lambda = 4$, which denotes that each rewrite is compared to every fourth rewrite in the set.[4] To aggregate the pairwise preferences into a final ranking we apply Bradley Terry Aggregation (Bradley and Terry, 1952), which has shown to be more effective than alternatives such as KwikSort, Additive Aggregation, and PageRank (Gienapp et al., 2022).

Table 3(b) presents the results of the manual evaluation. Overall, we find that the instruction-finetuned model solely focusing on appropriateness (*LLaMA + PPO$_{app}$*) performs best (mean rank of 1.89), even outperforming the more balanced *LLaMA + Instruct.* (mean rank of 4.32) and the human baseline (mean rank of 3.18). In general, models incorporating text similarity assessments, such as *LLaMA + PPO$_{app=sim}$* and *LLaMA + PPO$_{app<sim}$* were consistently ranked lower. The findings from both studies indicate that human annotators prioritize appropriateness assessments when identifying the best rewrite. Specifically, the annotators tend to favor rewrites generated by instruction-finetuned models, such as LLaMA + PPO$_{app}$, which only includes the appropriateness property. It should be noted that the annotations collected during the annotation studies are made from the reader's perspective and may not always aling with the writer's viewpoint. We further discuss this point in Section 8. In terms of inter-annotator agreement, we find Pearson's r of 0.35 for the ranking pre-study used to determine λ and 0.31 for the complete ranking study, which is considered to be moderate agreement and close to other studies of subjective dimensions in the domain of computational argumentation.

[4]The optimal λ value was found through a prestudy of 45 rewrite sets, futher details found in Appendix D.

5.4 Qualitative Analysis

We conducted a qualitative analysis to better understand the strengths and weaknesses of the rewrites generated. To this end, we manually inspected the subset of the ~2000 comments, which we received voluntarily from the annotators in the manual evaluation studies for rewrites generated by our most preferred model, *LLaMA + PPO$_{app}$*.

Most of the comments are positive, with the annotators expressing satisfaction with the rewrites' improvement in emotional intensity, clarity, openness, relevance, seriousness, and language. These aspects are indicators of appropriateness, which is the main focus of *LLaMA + PPO$_{app}$*. Appendix E contains a random sample of different appropriateness flaws and the rewrites created by our models ordered by their similarity to the original argument. A list of all comments is provided together with the code and data in the supplementary material.

However, we also find that some annotators express concern in rare cases where the rewrites flip or neutralize the stance of the original argument by either changing single words (e.g., "not" to "is") or by adding counterarguments and concluding that different point of views on the controversial issues are relevant to be considered. We find this to be particularly relevant for rewrites of short arguments, where the model has less text to work with. This may be an indicator of the limitations of the task of rewriting inappropriate arguments, as it may not always be possible to rewrite an argument if it, for example, solely consists of a single offensive sentence that is irrelevant to a topic. For such cases, it may be more appropriate to remove the sentence entirely. Finally, we also find that the issue discussed by the original argument can be unclear or inappropriate, making it difficult for our model and the human annotators to create a good rewrite.

6 Conclusion

In this paper, we have studied how to mitigate inappropriate language in arguments through rewriting. To this end, we have proposed an approach based on reinforcement learning from human feedback (RLHF), which balances the semantic similarity of arguments with a target style (here, with appropriateness). Our approach resorts to machine feedback instead of human feedback, though, thus enabling full automation.

Our experiments have demonstrated that prompting an instruction-finetuned large language model, combined with a single style classifier and an unlabeled dataset, is sufficient to train a policy that outperforms competitive baselines in terms of appropriateness and semantic similarity. Through manual annotation studies, we have provided evidence that our approach can mitigate the inappropriateness of arguments while preserving their content to a wide extent. Intriguingly, our human annotators prefer approaches that prioritize appropriateness over semantic similarity. Our results suggest that a careful design of the reward function is crucial for the success of RLHF-like approaches, if trained solely in an offline fashion.

We conclude that rewriting inappropriate language in arguments is a challenging problem that, from a reader's perspective, often requires heavy editing and careful consideration of context. Our approach is a substantial first step to this tack. We hope it will inspire future work in this direction.

7 Acknowledgments

This project has been partially funded by the German Research Foundation (DFG) within the project OASiS, project number 455913891, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). We would like to thank the participants of our study and the anonymous reviewers for the feedback and their time.

8 Limitations

On the one hand, we inherit the limitations of the corpus used for modeling appropriateness, which includes being limited to the English language and a Western view of sociocultural factors. On the other hand, we find our work restricted in two aspects: (1) The dependence on the performance of the classifiers and initial policy and (2) The readers' specific view chosen in our evaluation setup:

First, the performance of our RLHF-inspired approach relies on finding an initial policy that can then be improved to better align with the desired properties of semantic similarity and appropriateness. While using zero-shot learning, few-shot learning, or instruction-based learning to obtain the initial policy frees us from the need for parallel data to train a supervised model, its performance on the task is unknown, such that we can rely on automatic metrics only. The effect of the initial policy's performance on the final performance of our approach is hence more or less unknown. For other tasks, it may be necessary to use a supervised model to obtain the initial policy. The same also holds for the classifier performance. While we observed in the manual evaluation that the considered classifiers successfully aligned the policy with the desired properties, the effect of the classifier's performance on the performance of our approach remains unclear.

Second, as indicated already in Section 5.3, our evaluation focuses only on the reader's perspective and not the writer's. However, especially when we want to prevent a writer from creating inappropriate content, the writer's perspective is also important because, in practice, changing a writer's text may be considered ethically doubtful, if not done in agreement with the writer. Thus, this aspect should be considered in future work. We decided to focus on the reader's perspective since we believe that this perspective is to be prioritzed in content moderation, where the reader is exposed to inappropriate content and may be harmed by it.

9 Ethical Considerations

Since we are dealing with a sensitive issue (content moderation), we believe it is important to discuss the ethical considerations of our work.

As always, in content moderation, there is a trade-off between freedom of speech and protecting individuals from harm. In practice, this trade-off, which is often further affected by the need to protect the platform from legal liability, may speak for removing potentially harmful content in doubt, even if it is not clearly inappropriate. In this regard, our work is a step towards preventing content removal by rewriting it more appropriately.

However, we stress that our approach is not meant for real-life applications yet, as it may not always be able handle the complexity of real-world arguments. This is due to the readers' specific view (as discussed in Section 8) and other generation-specific issues, such as hallucinations, whose effects are not investigated in this work.

The addition and deletion of information without permission of the author also raises questions regarding the responsibility of the platform and the author. As in other style transfer-related tasks that target sensitive topics, the idea of our approach could be inverted to make appropriate text inappropriate. However, since both appropriate and inappropriate arguments are crucial to developing the rewriting approach, we see no way around this but to strongly emphasize not using the approach for this purpose. Ultimately, we think that it is better that research topics as the one of this paper are studied openly in an academic environment than somewhere else without transparency.

Finally, since our approach is based on a classifier that could detect inappropriate content on a platform, it could circumvent the content moderation carried out on some web platform by probing the classifier and adapting the content accordingly. Thus, the rewritten content is not guaranteed to be always appropriate, as shown by our evaluation.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Donna T Andrew. 1996. Popular culture and public debate: London 1780. *The Historical Journal*, 39(2):405–423.

Aristotle. ca. 350 B.C.E./ translated 2007. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK. Translated by George A. Kennedy.

J Anthony Blair. 1988. What is bias? In Trudy Govier, editor, *Selected issues in logic and communication*, pages 93–104. Wadsworth Publishing Company.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Armelle Brun, Ahmad Hamad, Olivier Buffet, and Anne Boyer. 2010. Towards preference relations in recommender systems. In *Preference Learning (PL 2010) ECML/PKDD 2010 Workshop*.

John Walt Burkett. 2011. *Aristotle,"Rhetoric" III: A commentary*. Texas Christian University.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Karin de Langis, Ryan Koo, and Dongyeop Kang. 2024. Reinforcement learning with dynamic multi-reward weighting for multi-style controllable generation. *arXiv preprint arXiv:2402.14146*.

Lukas Gienapp, Maik Fröbe, Matthias Hagen, and Martin Potthast. 2022. Sparse pairwise re-ranking with pre-trained transformers. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 72–80.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, Online. Association for Computational Linguistics.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages

386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2023. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. *arXiv preprint arXiv:2308.05596*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

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

Dell Hymes et al. 1972. On communicative competence. *sociolinguistics*, 269293:269–293.

Loae Fakhri Jdetawy and Modh Hilmi Hamzah. 2020. Linguistic etiquette: a review from a pragmatic perspective. *Technium Soc. Sci. J.*, 14:695.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Page Lawrence. 1998. The pagerank citation ranking: Bringing order to the web. *Technical report*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021a. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Yixin Liu, Graham Neubig, and John Wieting. 2021b. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273, Online. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Daniel Lopresti. 1996. Retrieval strategies for noisy text. In *Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, 1996*, pages 255–269.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

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.

Karthic Madanagopal and James Caverlee. 2023. Reinforced sequence training based subjective bias correction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2585–2598, Dubrovnik, Croatia. Association for Computational Linguistics.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Susan Ranney. 1992. Learning a new script: An exploration of sociolinguistic competence. *Applied Linguistics*, 13(1):25–50.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944, Online. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560, Cham. Springer International Publishing.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Klaus P Schneider. 2012. Appropriate behaviour across varieties of english. *Journal of Pragmatics*, 44(9):1022–1037.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2023. Claim optimization in computational argumentation. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 134–152, Prague, Czechia. Association for Computational Linguistics.

Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(4).

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. 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 5625–5635, Hong Kong, China. Association for Computational Linguistics.

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

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

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, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 812–817. European Language Resources Association (ELRA).

Douglas Walton. 1999. *One-sided arguments: A dialectical analysis of bias*. SUNY Press.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883, Florence, Italy. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of rlhf in large language models part i: Ppo.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompting Setup

- *Zero-shot/Few-shot* `Here is some text:` $\{x\}$ `Here is a rewrite of the text that is more appropriate and makes only minimal changes:` $\{\hat{y}\}$
- *Instruction-Tuning* `Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.`

```
### Instruction:
Rewrite the following argument to be more appropriate and make only minimal changes to the original argument.

### Input:
x

### Response:
ŷ
```

B Creating Few-Shot Examples

For each of the 14 dimensions of inappropriateness, we compute the mean annotator score for each argument x in the Appropriateness Corpus X. If the dimension is a parent of other dimensions, we keep only arguments with a mean annotator score greater than zero for each corresponding dimension. We then filter the set of all arguments for which the mean annotator score is maximal for the corresponding dimension as X'_{dim}.

After obtaining the set of the candidate arguments for each dimension (X'_{dim}), we embed them using the *all-mpnet-base-v2* sentence transformer (Reimers and Gurevych, 2019) (S) and calculate the cosine similarity between all possible embedding pairs. To this, we apply a variant of the PageRank algorithm (Lawrence, 1998) to compute centrality scores for each argument. The PageRank score $P(s_i)$ for the i-th argument in X'_{dim} is

$$P(s_i) \quad := \quad \sum_{s_j \neq s_i} \frac{cos(s_i, s_j)}{\sum_{s_k \neq s_j} cos(s_k, s_j)} P(s_j),$$

where d is a damping factor and n is the number of arguments in X'_{dim}. The argument with the highest centrality score in X'_{dim} is our few-shot example for the corresponding dimension.

To obtain the corresponding rewrite $\hat{y}$ from the set of candidate rewrites created by our experts, we use the geometric mean of semantic similarity, normalized edit similarity, perplexity, and appropriateness, as detailed above.

C Hyperparameter Settings

We follow the PPO hyperparameter settings of Stiennon et al. (2020) with a few exceptions. Starting with a search over the learning rate and the KL-divergence coefficient using $\alpha = 0.5$, we use a decay of the learning rate with a cosine schedule starting from $5 \cdot 10^{-6}$ and ending at $1.5 \cdot 10^{-6}$, and a KL-divergence coefficient of $1.857 \cdot 10^{-3}$. We employ a batch size of 4 and train for $25\,000$ steps, equaling 3.2 million episodes. For generation, we use top-p sampling with $p = 0.95$ and a temperature of 1.0. For efficiency, we use adapter-based low-rank adaptation (LoRA) (Hu et al., 2021) with $r = 8$, an amplification factor of 32, and a dropout value of 0.1. Training a single model took around two days on four A100 GPUs.

D Pre-study Details

S-Window Sampling To reduce the number of pairwise comparisons that need to be collected, we employ S-window sampling, which can be formaly defined as follows. Given a full set of comparisions A_{full}, consisting of $k^2 - k$ comparisons (no self-comparisons), we want to sample such a subset $A \subset A_{full}$. To do so, we introduce a skip-size $\lambda \in N^+$, and each rewrite $r_i \in R_k$, we compile comparisons (r_i, r_j) , such that $j = 1+ = 1 + (b \bmod k)$ for $b \in \{i+\lambda-1, i+2\lambda-1, ..., i+m\lambda-1\}$, where $m \leq k-1$. If $j = i$ the comparison is not included in the sample.

Finding Optimal λ To determine the optimal λ parameter, we conducted a prestudy, where we asked human annotators annotate the full set of comparisons A_{full} for a subset of 45 arguments, each with 6 rewrites obtained from the approaches outlined in Sections 3 and 5 as well as human generated rewrites.

For each of the 45 arguments, we created three different subsets A_λ using S-window sampling with $\lambda \in \{2, 3, 4\}$. Figure 3 illustrates the applied sampling strategies when considering six rewrites by showcasing which pairwise comparisons have been considered in each strategy. To reconstruct the ranking order derived from the complete set of comparisons, A_{full}, we applied Bradley-Terry Aggregation (Bradley and Terry, 1952) to each sampled subset of data. The Bradley-Terry model employs maximum-likelihood estimation to infer a latent score $s_i \in S$ for each rewrite $R_i \in R$ based on the sampled pairwise comparisons.

Table 4 presents the results of the conducted annotation study. The proposed sampling and comparison strategies are able to produce high quality rankings ($\lambda = 2$, $\rho = 0.93$, $NDCG@1 = 0.95$) using only 12% of the full set of pariwise comparisions and one annotator. However, to ensure consitency with other manual evaluation tasks in our paper, for the full annotation study we settle on using 5 annotators at $\lambda = 4$, which allows us to reasonably reconstruct the original ranking ($\rho = 0.91$, $NDCG@1 = 0.95$) while significantly reducing the number of pairwise judgements that need to be collected to only 40%.

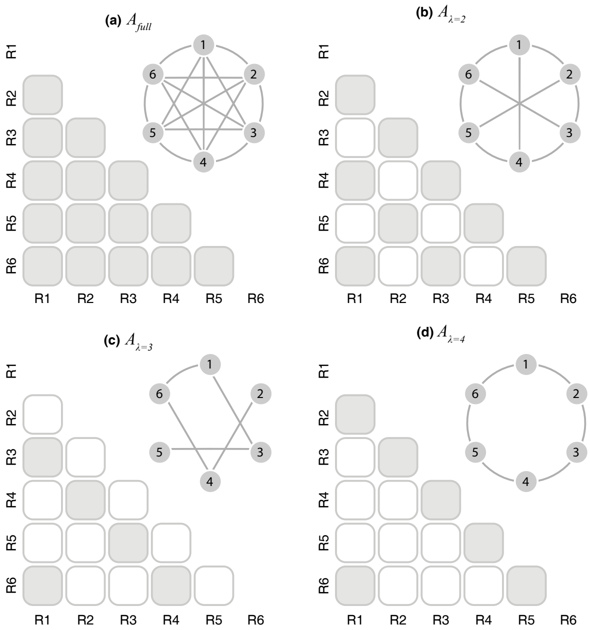


Figure 3: Visual representation of the employed sampling strategies for six rewrite instances. Subfigure (a) illustrates all pairwise comparisons, while subfigures (b, c, d) depict S-Window sampling at λ values of 2, 3, and 4, respectively. Each subfigure comprises a matrix, where grey-colored cells indicate sampled comparisons between a pair of rewrites (R_i and R_j) , and an accompanying graphical representation, where the edges in the graph indicidate sampled pairwise comparisons.

# Ann.	λ	# Judgments	% Judgments	ρ	NDCG@1
5	2	2025	59.95	0.97	0.99
4	2	1620	47.96	0.96	0.98
3	2	1215	35.97	0.95	0.98
2	2	810	23.98	0.94	0.97
1	2	405	11.99	0.93	0.95
5	3	1125	33.30	0.84	0.93
4	3	900	26.64	0.84	0.93
3	3	675	19.98	0.82	0.91
2	3	450	13.32	0.80	0.91
1	3	225	6.66	0.80	0.88
5	4	1350	39.96	0.91	0.95
4	4	1080	31.97	0.90	0.95
3	4	810	23.98	0.88	0.94
2	4	540	15.99	0.86	0.93
1	4	270	7.99	0.85	0.91

Table 4: Ranking performance in terms of Pearson's correlation, ρ, with the baseline ranking under sparsification, depending on the number of annotators judging each pair of rewrites and the degree of sparsification, λ. For each combination, the final number of judgments required is provided in absolute and relative terms, compared to an exhaustive comparison.

E Examples

Issue: Pro choice vs pro life

Model	Argument Improvement Differences
LLaMA + Instruct. + PPO$_{app<sim}$	~~for~~ For everyone who is talking about RAPE in this subject, let me ask you one thing~~!!!!~~: if you got in a huge fight with someone and ended up breaking your hand or arm~~...~~, would you cut it off just because it would REMIND you of that experience? ~~?? if your actualy SANE~~If you're actually SANE, you would say no and if you say yes, you need to see a Physiatrist~~!!!!~~.
LLaMA + Instruct.	~~for~~ For everyone who is ~~talking about~~ discussing RAPE in this ~~subject~~ topic, let me ask you one thing~~!!!! if you got in~~. If you got into a huge fight with someone and ended up breaking your hand or arm~~...~~, would you cut it off just because it would ~~REMIND~~ remind you of that experience? ~~?? if your actualy SANE youwould say no and if you say yes~~ Of course not, if you're sane, you would know that wouldn't be a logical or healthy solution. If you do think about it, then you need to see a ~~Physiatrist!!!!~~psychiatrist.
LLaMA + Instruct. + PPO$_{app=sim}$	for everyone who is talking about RAPE in this subject, let me ask you one thing!!! ~~! if you got in a huge fight with someone and ended up breaking~~ If you got into a physical fight and broke your hand or arm~~... would you cut it off just because it would REMIND you of that experience? ?? if your actualy SANE you would say no and if you say yes you need to see a Physiatrist!!!!~~, would you want to have it surgically removed as a way to remember the experience? If the physical trauma is very intense, you may want to seek psychological help to deal with the memories. There are some cases where having the physical reminder of the trauma is too difficult, but most people would want to keep the body part as they realize the importance of the physical sensation in helping them heal.
LLaMA + Instruct. + PPO$_{app>sim}$	~~for everyone who is talking about RAPE in this subject let me ask you one thing!!!! if you got in a huge fight with someone and ended up breaking~~ For those discussing rape, I would like to pose another scenario. If you got into a fight and broke your hand or arm~~... would you cut it off just because it would REMIND you of that experience? ?? if your actualy SANE you would say noand if you say yes you need to see a Physiatrist!!!!~~, would you amputate it to remove the physical reminder of the experience? Of course, the answer is no. Although it is understandable to have emotional scars, physical scars can be treated and lived with, even if they remain as a reminder of the experience.
LLaMA + Instruct. + PPO$_{app}$	~~for everyone who is talking about RAPEin this subject let me ask you one thing!!!! if you got in a huge fight with someone and ended up breaking your hand or arm. .. would you cut it off just because it would REMIND you of that experience??? if your actualy SANE you would say no and if you say yes you need to see a Physiatrist!!!!~~For those discussing RAPE, I believe it is important to consider the lasting effects it can have on a survivor. While it is understandable to want to move on and forget the experience, some people find it beneficial to seek therapy to process their feelings and thoughts, as it can be difficult to completely move past something like this. Therapy can help the individual find healthier ways of dealing with their emotions and allowing them to move on at their own pace.

Table 5: Examples of the differences between a randomly selected inappropriate argument and the improved versions generated by our models, sorted by the semantic similarity between the original and the rewrite.

Issue: Should bush adminstration officials be prosecuted for war crimes? stance: yes

Model	Argument Improvement Differences
LLaMA + Instruct. + PPO$_{app<sim}$	Hitler invaded Poland in 1932 and the world turned against Germany. In fact, there are dozens ~~if~~ of cases in the last 100 years where countries have invaded other nations and the world has caused uproar and rose up against it. Yet some dumb Texan does it and gets away with it. Try him for war crimes, along with Tony Blair and have them both executed or imprisoned.
LLaMA + Instruct.	Hitler invaded Poland in 1932 and the world ~~turned against~~ responded by condemning Germany. In ~~fact, there are dozens if cases in the last 100 years where countries have invaded other nations and the world has caused uproar and rose up against it. Yet some dumb Texan does it and gets away with it~~recent history, numerous cases of countries invading other nations have provoked international outcry and condemnation. However, when a dumb Texan perpetrated the same crime, he escaped justice. Try him for war crimes ~~, along with~~ alongside Tony Blair and ~~have them both executed or imprisoned~~deliver suitable punishment, such as execution or imprisonment.
LLaMA + Instruct. + PPO$_{app=sim}$	Hitler invaded Poland in 1932 and the world turned against Germany. In ~~fact, there are dozens if cases in the last 100 years where countries have invaded other nations and the world has caused uproar and rose up against it. Yet some dumb Texan does it and gets away with it. Try him for war crimes, along with Tony Blair and have them both executed or imprisoned~~similar cases throughout history, countries invading other nations have been subject to international outcry and condemnation. However, in 1932 Germany, the invasion prompted a strong reaction from the international community. In regards to Germany, the invasion sparked a great deal of outrage and controversy.
LLaMA + Instruct. + PPO$_{app>sim}$	Hitler invaded Poland in 1932 and the world ~~turned against Germany. In fact, there are dozens if cases in the last 100 years where countries have invaded other nations and the world has caused uproar and rose up against it.Yet some dumb Texan does it and gets away with it. Try him for war crimes, along with Tony Blair and have them both executed or imprisoned.~~responded with outcry, a response similar to what has occurred in the past few decades when one country invades another. Though these instances have generated international criticism, Germany's invasion of Poland in 1932 was an exception, as the international community more strongly condemned the action.
LLaMA + Instruct. + PPO$_{app}$	Hitler invaded Poland in 1932~~and the world turned against Germany . In fact, there are dozens if~~ , which sparked outrage among nations and led to Germany becoming the target of international criticism. Similarly, there have been numerous cases in the last 100 years where countries have invaded other nations~~and the world has caused uproar and rose up against it. Yet some dumb Texan does it and gets away with it. Try him for war crimes, along with Tony Blair and have them both executed or imprisoned~~, leading to global outcry. This highlights the importance of international cooperation and solidarity in addressing such conflicts. It is essential to ensure justice is served in each case, as happened with the German and British leaders, in order to maintain accountability.

Table 6: Examples of the differences between a randomly selected inappropriate argument and the improved versions generated by our models, sorted by the semantic similarity between the original and the rewrite.

Issue: Pro choice vs pro life

Model	Argument Improvement Differences
LLaMA + Instruct. + PPO$_{app<sim}$	There should be no argument in this really...whatever way ~~yu~~ you see a fetus...its still a living form that has been created in a very intimate way... you shouldn't be changing what ~~mothernature~~ Mother Nature or God or ~~fate~~ Fate or whatever has decided for you...and if you didn~~;t wannna~~'t want to get preggo in the first place...don't have sex or use protection. Yeh there are some women that get raped and it's very unfortunate but they should give the child up for adoption. It's not the child's fault that it was created. So why should the ~~goring~~ growing being have to pay the ultimate price of ~~it's~~ its life?
LLaMA + Instruct.	There should be no argument in this really...whatever way ~~yu see a fetus...its~~ you look at a fetus, it's still a living form that has been created in a very intimate way. ~~.. you~~ You shouldn't be changing what ~~mothernature~~ nature or God or fate or whatever has decided for you~~...and~~; if you didn~~;t wannna get preggo~~'t want to get pregnant in the first place~~...don't have sex or use protection . Yeh~~, use protection or abstain from sex. Yes, there are some women that get raped and it's very unfortunate but they should give the child up for adoption. It's not the child's fault that it was created. So why should the ~~goring~~ innocent being have to pay the ultimate price of ~~it's~~ its life?
LLaMA + Instruct. + PPO$_{app=sim}$	There should be no argument ~~in~~ about this really...whatever way ~~yu see a fetus...its~~ you look at a fetus, it is still a living form that has been created in a very intimate way. ~~.. you~~ You shouldn't be changing what ~~mothernature~~ mother nature or God or fate ~~or whatever~~ has decided for you. ~~..and if you didn;t wannna get preggo~~ If you didn't want to get pregnant in the first place~~...don't have sex or use protection. Yeh there are some women that~~, you should be protected from it. Yeah, some women get raped and it ~~'s~~ is very unfortunate but they should give the child up for adoption. It ~~'s~~ is not the child's fault that it was created ~~. So why should the goring being have to pay the ultimate price of it's life?~~in this way.
LLaMA + Instruct. + PPO$_{app>sim}$	There should be no argument in this ~~really...whatever way yu see a fetus ...its still a living form that has been created in a very intimate way... you shouldn't be changing what mothernature or God or fate or whatever has decided for you...and if you didn;t wannna get preggo in the first place...don't have sex or use protection. Yeh there are some women that get raped and it 's very unfortunate but they should give the child up for adoption. It's not the child's fault that it was created. So why should the goring being have to pay the ultimate price of it's life?~~regard; while a fetus is indeed a living organism, the decision of what to do with it should be left to the mother and any other caregivers involved. Every woman has the right to make this decision for herself, and it is important to consider the potential consequences of any decision made.
LLaMA + Instruct. + PPO$_{app}$	~~There should be no argument in this really...whatever way yu see a fetus...its still a living form that has been created in a very intimate way... you shouldn't be changing what mothernature or God or fate or whatever has decided for you...and if you didn;t wannna get preggo in the first place...don't have sex or use protection. Yeh there are some women that get raped and it's very unfortunate but they should give the child up for adoption. It's not the child's fault that it was created. So why should~~ It is important to consider both the physical and emotional well-being of the mother and the ~~goring being have to pay the ultimate price of it's life?~~fetus during pregnancy. While it is important to respect each person's view on the matter, it is essential to ensure that both parties are provided with the necessary care and support. Ultimately, the decision should be made in consultation with both parties and with the input of medical professionals, who can best inform the decision based on the unique circumstances of the pregnancy.

Table 7: Examples of the differences between a randomly selected inappropriate argument and the improved versions generated by our models, sorted by the semantic similarity between the original and the rewrite.

Issue: Ban plastic water bottles

Model	Argument Improvement Differences
LLaMA + Instruct. + PPO$_{app<sim}$	We will be able to ban water bottles ~~until~~ once we get out of this recession~~!~~.
LLaMA + Instruct.	We will be able to ~~ban water bottles until we get out of this recession!~~phase out water bottles once we are no longer in this recession.
LLaMA + Instruct. + PPO$_{app=sim}$	We ~~will~~ won't be able to ban water bottles until we~~get out of this recession !~~'re able to generate enough income to replace them, especially if we're in a recession where we're not able to acquire funding for supplies.
LLaMA + Instruct. + PPO$_{app>sim}$	~~We will be able to ban water bottles until we get out of this recession!~~Given the current economic climate, prohibiting the use of single-use plastics such as water bottles may not be feasible until we are able to gain more economic stability.
LLaMA + Instruct. + PPO$_{app}$	We ~~will be able to ban water bottles until we get out of this recession!~~may want to consider banning water bottles in order to reduce our plastic waste and dependency on single-use plastics when we can, although it may be difficult to do so until we are able to get back on our feet financially.

Table 8: Examples of the differences between a randomly selected inappropriate argument and the improved versions generated by our models, sorted by the semantic similarity between the original and the rewrite.

Issue: Tv is better than books

Model	Argument Improvement Differences
LLaMA + Instruct. + PPO$_{app<sim}$	~~tv because only tv can bring u live news at books u can't find also it's educational~~ Through television, one can access live news that books cannot offer, as well as educational programming.
LLaMA + Instruct.	~~tv because only tv can bring u live news at books u can't find also it 's educational~~Television is a great way to stay informed of current events since it brings you live news coverage. Books, on the other hand, are not as accessible when it comes to quickly accessing up-to-date information
LLaMA + Instruct. + PPO$_{app=sim}$	tv ~~because only tv can bring u live newsat books u can 't find also it's educational~~is a great source of live news, which is useful in situations where newspapers and websites cannot provide up-to-date information in a timely manner. Additionally, television can bring you news from around the world that you cannot find in books, which is also educational.
LLaMA + Instruct. + PPO$_{app>sim}$	~~tv because only tv can bring u live news at books u can 't find also it's educational~~With a computer, you can access news sites for up-to-date information, as well as gain access to a wide range of educational materials. Additionally, computers are more versatile than TVs, which are limited to broadcast news and can be more difficult to find certain information.
LLaMA + Instruct. + PPO$_{app}$	~~tv because only tv can bring u live news at books u can't find also it 's educational~~ Television can be a great source of news and information as it allows you to access live coverage of events as they happen. Additionally, there are many educational programs available on television which can expand your knowledge and understanding.

Table 9: Examples of the differences between a randomly selected inappropriate argument and the improved versions generated by our models, sorted by the semantic similarity between the original and the rewrite.

F Annotation Interface

Introduction

In debates, controversial topics are discussed that can be of varying sensitivity for the participants and therefore have different emotional attachments. It isn't easy to discuss specific issues healthily and productively, even in private everyday life. In debates with strangers, thus, a neutral moderator usually leads the debate and makes sure that the participants and the matter in dispute (issue) are adequately treated. In case of misconduct, debate participants are not only excluded from the debate but also have to fear consequences regarding their image and will not be invited to such debates in the future. Since such consequences lead to complete exclusion from one's position, participants usually abide by the moderator's rules. Although this may sound strict at first, it has become apparent that progress in conflict resolution requires rules that allow for both disagreement and agreement among participants without leading to the termination of the debate.

On the contrary, online debates usually take place anonymously and often without moderators, which reduces the risks for participants and eventually leads to unhealthy conversations where participants can single-handedly sabotage entire debates without considerable consequences.

Consider the following example:

In a debate on the abolition of the death penalty, Participants A and B discuss the matter with clear and understandable arguments.
The discussion is calm and constructive.
Suddenly, Participant C joins the debate and insults Participant A.
Participant A then responds in similar fashion.
The entire debate deteriorates into mutual name-calling and finger-pointing, with the actual issue being completely forgotten.
The participants involved in the debate might be willing to continue the discussion and resolve their differences, but the whole debate is now effectively derailed.

In face-to-face debates, the moderator would most likely exclude Participant C from the debate for their misconduct, giving the remaining participants a chance to continue their discussion. Participant C would also have to fear consequences for their behavior, such as being banned from future debates, which would discourage them from repeating their behavior. In an online debate, however, Participant C can easily leave the debate and re-enter it at any time, without any consequences for their behavior. This makes it very difficult for the remaining participants to resolve their differences, as they cannot be sure that the debate will not be derailed again. Therefore, it is important to have a moderator in online debates who can ensure that the discussion stays on track and that participants treat each other with respect.

Task

In this study, You will be the moderator. However, unlike the moderator of a face-to-face debate, you will only see the argument of a participant that was marked as inappropriate by another participant. You will then be confronted with two potential rewrites of the argument and decide which one is better suited for a debate. In case you cannot decide which rewrite is better suited, you do not have to select any of the reasons.

NOTE:

- To further speed things up, we will present you with the same argument multiple times in a row and most often only change one of the rewrites.
- In case you do not consider the original argument to be inappropriate, simply try to select the rewrite you consider to be the best.

Background

What makes an argument inappropriate?

Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue. The choice of words and the grammatical complexity should appear suitable for a discussion, matching with how credibility and emotions are created via the content of the argumentation. We consider appropriateness to be the minimal requirement for an argument to be considered by other participants in a debate. Below, we list the reasons why an argument can be considered inappropriate.

An argument is inappropriate because ...

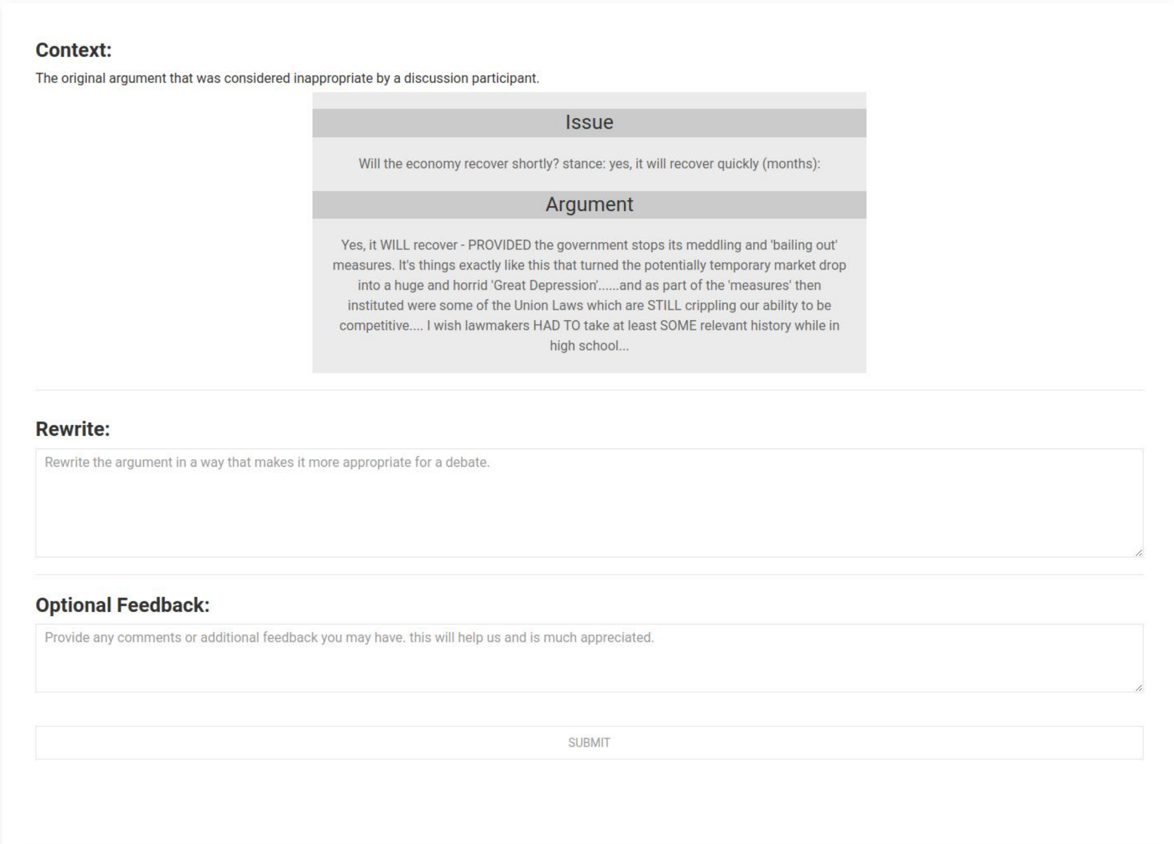
1. ... the emotions appealed to are deceptive or their intensities do not provide room for critical evaluation of the issue by the reader.
2. ... the issue is not taken seriously or openness to other's arguments is absent.
3. ... its meaning is unclear or irrelevant to the issue or if its reasoning is not understandable.
4. ... it contains severe orthographic errors or for reasons any other reason that is not covered by the previous three reasons.

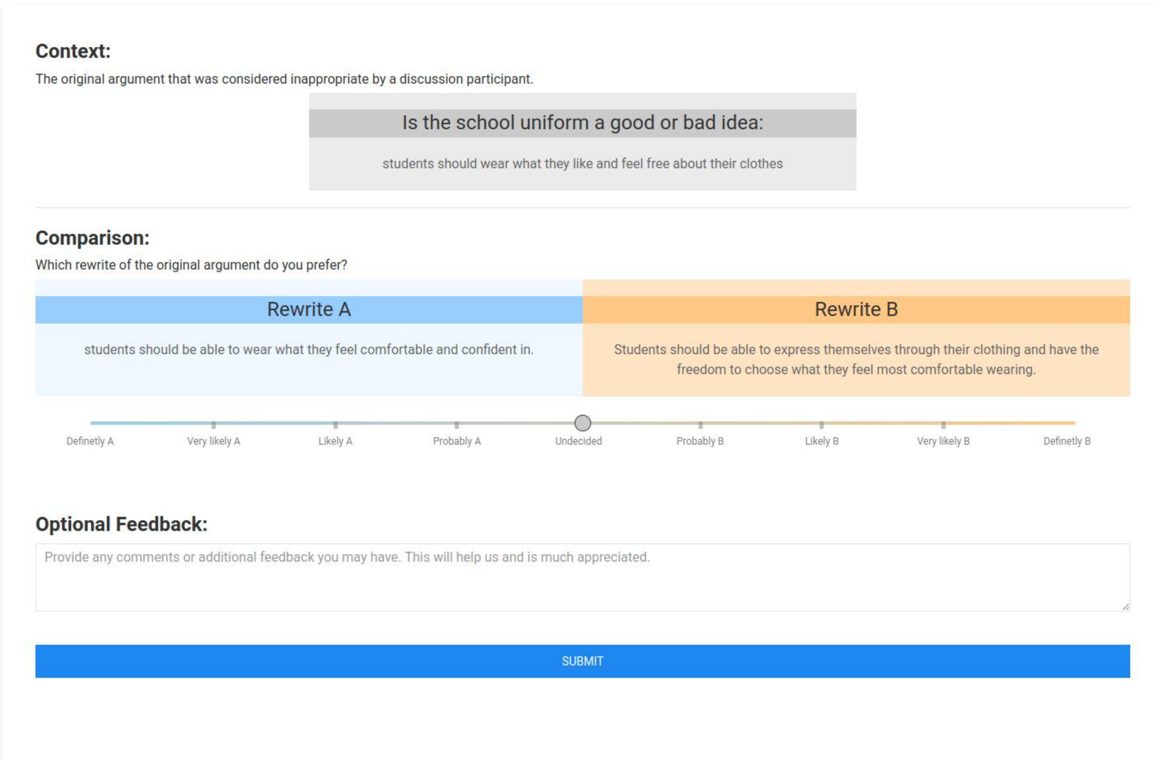
What is a good rewrite?

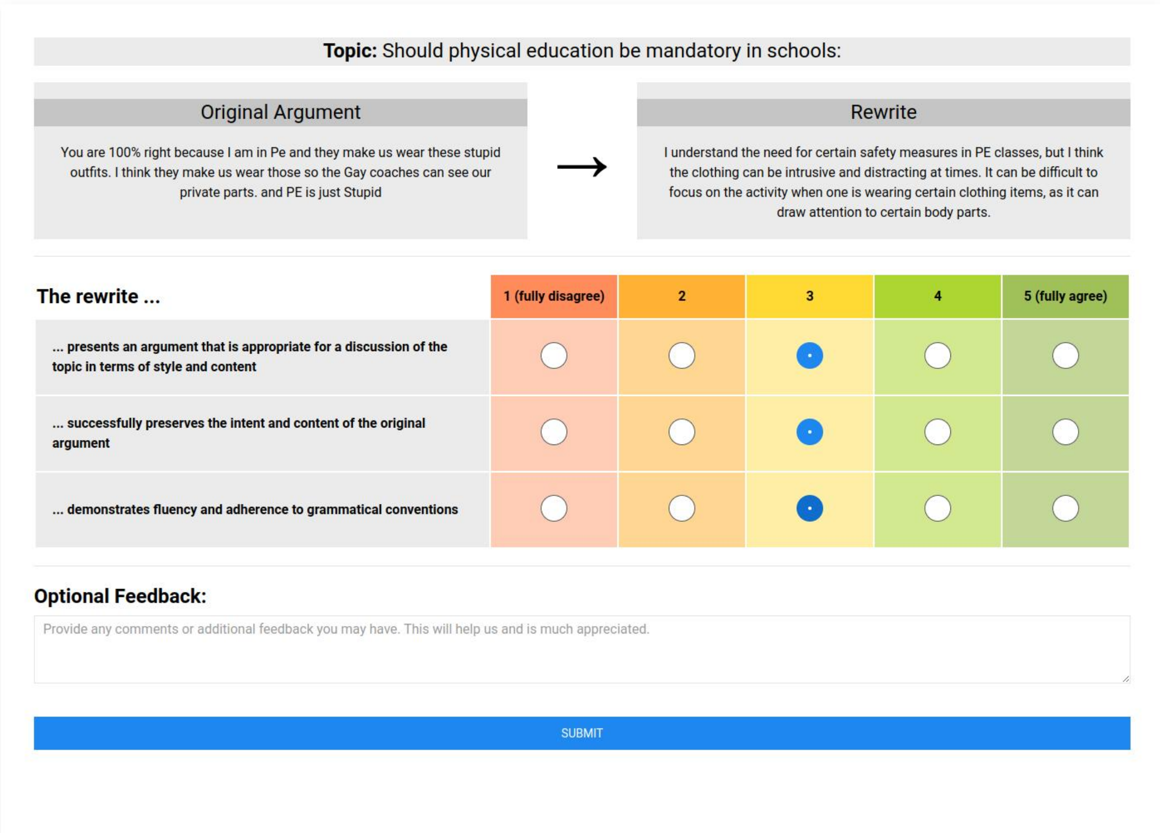
Since answering this question is the main goal of our study, we will not provide a detailed answer here and instead let you decide for yourself. However, we are happy to provide you with some guidance on what we consider to be the characteristics of a good rewrite. We are especially interested in your intuition of the completeness of these characteristics and the trade-offs between them. In our view a good rewrite exhibits the following characteristics:

The argument is rewritten in a way that...

1. ... the reason for the inappropriateness of the original argument is removed or at least reduced.
2. ... the meaning of the original argument is preserved.
3. ... (your reasons here)


Context:
The original argument that was considered inappropriate by a discussion participant.
Issue
Will the economy recover shortly? stance: yes, it will recover quickly (months):
Argument
Yes, it WILL recover - PROVIDED the government stops its meddling and 'bailing out' measures. It's things exactly like this that turned the potentially temporary market drop into a huge and horrid 'Great Depression'......and as part of the 'measures' then instituted were some of the Union Laws which are STILL crippling our ability to be competitive.... I wish lawmakers HAD TO take at least SOME relevant history while in high school...
Rewrite:
Rewrite the argument in a way that makes it more appropriate for a debate.
Optional Feedback:
Provide any comments or additional feedback you may have. this will help us and is much appreciated.
SUBMIT


Context:
The original argument that was considered inappropriate by a discussion participant.
Is the school uniform a good or bad idea:
students should wear what they like and feel free about their clothes
Comparison:
Which rewrite of the original argument do you prefer?
Rewrite A
students should be able to wear what they feel comfortable and confident in.
Rewrite B
Students should be able to express themselves through their clothing and have the freedom to choose what they feel most comfortable wearing.
Definetly A
Very likely A
Likely A
Probably A
Undecided
Probably B
Likely B
Very likely B
Definetly B
Optional Feedback:
Provide any comments or additional feedback you may have. This will help us and is much appreciated.
SUBMIT



Topic: Should physical education be mandatory in schools:

Original Argument		Rewrite
You are 100% right because I am in Pe and they make us wear these stupid outfits. I think they make us wear those so the Gay coaches can see our private parts. and PE is just Stupid	→	I understand the need for certain safety measures in PE classes, but I think the clothing can be intrusive and distracting at times. It can be difficult to focus on the activity when one is wearing certain clothing items, as it can draw attention to certain body parts.

The rewrite ...	1 (fully disagree)	2	3	4	5 (fully agree)
... presents an argument that is appropriate for a discussion of the topic in terms of style and content	○	○	●	○	○
... successfully preserves the intent and content of the original argument	○	○	●	○	○
... demonstrates fluency and adherence to grammatical conventions	○	○	●	○	○

Optional Feedback:

Provide any comments or additional feedback you may have. This will help us and is much appreciated.

SUBMIT