

# Incivility and Rigidity: Evaluating the Risks of Fine-Tuning LLMs for Political Argumentation

Svetlana Churina<sup>1</sup> & Kokil Jaidka<sup>1,2</sup>

<sup>1</sup>Department of Communications and New Media,

<sup>2</sup>Centre for Trusted Internet & Community,

National University of Singapore

jaidka@nus.edu.sg

## Abstract

Incivility on platforms such as Twitter (now X) and Reddit complicates the development of AI systems that can support productive, rhetorically sound political argumentation. We present experiments with *GPT-3.5 Turbo* fine-tuned on two contrasting datasets of political discourse: high-incivility Twitter replies to U.S. Congress and low-incivility posts from Reddit's *r/ChangeMyView*. Our evaluation examines how data composition and prompting strategies affect the rhetorical framing and deliberative quality of model-generated arguments. Results show that Reddit-finetuned models generate safer but rhetorically rigid arguments, while cross-platform fine-tuning amplifies adversarial tone and toxicity. Prompt-based steering reduces overt toxicity (e.g., personal attacks) but cannot fully offset the influence of noisy training data. We introduce a rhetorical evaluation rubric - covering justification, reciprocity, alignment, and authority - and provide implementation guidelines for authoring, moderation, and deliberation-support systems. All code and annotation resources are available here: <https://github.com/Sveta151/CLAPTON>.

**Keywords:** politics, twitter, reddit, toxicity

**Disclaimer** — This paper contains some profanity that may be disturbing to some readers.

## 1. Introduction

Large language models (LLMs) are increasingly applied in computational social science to emulate, analyze, and generate human discourse. Their capacity for producing coherent, contextually grounded text makes them useful for studying argumentative practices in online political communication. However, using LLMs to support or simulate public deliberation raises concerns about rhetorical fidelity, bias amplification, and discourse degradation, particularly when models are fine-tuned on high-incivility data typical of social media. Two related risks are salient. First, fine-tuning on highly adversarial data, such as replies to politicians on Twitter (now X), may encode toxic or polarizing rhetorical patterns, reducing the model's ability to generalize to balanced argumentation. Second, recursive training on synthetic model outputs can compound stylistic distortions over time (Shumailov et al., 2024), degrading deliberative quality and reliability. Real-world deployments of LLM-based agents in political forums, including covert interventions where models were explicitly instructed to disregard ethical constraints, demonstrate that persuasive AI systems are already operating in the wild, often without disclosure or oversight.

This study investigates how platform-specific data shape the rhetorical and deliberative properties of AI-generated political arguments. We focus

on whether fine-tuned models preserve the reasoning and reciprocity that characterize productive political dialogue, assuming that fine-tuning transfers not only topical content but also stylistic norms, and that these tendencies can be partially offset through data curation and prompting. Our research pursues two objectives:

- **RO1:** Characterize the rhetorical and deliberative quality of political arguments produced by LLMs fine-tuned on high-incivility social media data.
- **RO2:** Evaluate mitigation strategies - such as balanced training and prompt-based steering - for improving deliberative quality in generated arguments.

## 2. Study Contributions

Political arguments are structured, stance-taking texts designed to persuade or contest viewpoints (Bender et al., 2011; Rowe, 2015). Yet not all arguments advance deliberation. We define *deliberative quality* as the degree to which an argument provides reasons, acknowledges opposition, and seeks common ground, features essential to inclusive and respectful public discourse.

Prior work has explored argument generation, stance framing, and toxicity detection in isolation. Less attention has been paid to how platform-specific norms shape the rhetorical behavior of models after fine-tuning, that is, whether training data from adversarial versus deliberative environ-

ments leaves measurable traces in the arguments models produce. Our study addresses this gap through three contributions:

- **Empirical:** We show that fine-tuned models inherit platform-specific rhetorical biases. Models trained on high-incivility data (Twitter)<sup>1</sup> produce more adversarial rhetoric, while those trained on low-incivility data (Reddit CMV) yield more civil but stylistically constrained arguments, a pattern we characterize as a bias-variance tradeoff in rhetorical behavior.
- **Methodological:** We introduce an LLM-assisted annotation pipeline for identifying *alignment* and *authority* moves in generated text, two key indicators of deliberative engagement that are underexplored in automated evaluation frameworks.
- **Practical:** We offer design guidelines for configuring models, curating training data, and crafting prompts that balance rhetorical diversity, civility, and deliberative quality across deployment contexts.

Together, these contributions advance the study of rhetoric-aware fine-tuning and provide empirical foundations for designing AI systems that enhance, rather than erode, the quality of deliberative discourse.

### 3. Related Work

Research on AI-driven argumentation has examined prompting strategies, factual grounding, and discourse quality across a range of political and deliberative contexts. [Figueras and Aggeri \(2024\)](#) and [Lin et al. \(2023\)](#) demonstrated that LLMs can generate concise, deliberative arguments, though persuasive impact varies considerably by ideological framing and rhetorical register ([Simmons, 2023](#); [El Baff et al., 2024](#)).

A separate body of work has raised concerns about what models learn when fine-tuned on politically charged or low-quality data. Fine-tuning on adversarial discourse risks amplifying bias and misinformation ([Giarelis et al., 2024](#); [Dykes et al., 2024](#)), and recursive training on synthetic outputs can further degrade lexical diversity and factual integrity over successive generations ([Shumailov et al., 2024](#)). These findings motivate closer scrutiny of how training data quality shapes the rhetorical, and not just factual, properties of model outputs.

Efforts to quantify deliberative quality have produced indices covering civility, rationality, and reciprocity ([Behrendt et al., 2024](#)), providing a

foundation for systematic evaluation. Building on this work, we shift focus from classification to generation. Using the annotated *CLAPTON* dataset ([Jaidka, 2022b](#)), we examine how LLMs internalize and reproduce deliberative features, specifically *justification* and *reciprocity* as defined by [Steenbergen et al. \(2003\)](#), after fine-tuning on platform-specific corpora. Our analysis connects rhetorical modeling with practical design considerations for AI-mediated deliberation.

## 4. Method

We compared model outputs generated after fine-tuning on corpora from different online platforms to identify rhetorical patterns associated with incivility. Our approach rested on three assumptions. First, that rhetorical qualities such as justification and reciprocity could be reliably identified and measured using annotated discourse labels. Second, that fine-tuning on platform-specific corpora transferred not only topical content but also stylistic and argumentative patterns. Third, that prompting and data-balancing strategies could influence, though not fully override, these learned rhetorical tendencies.

### 4.1. Prompting Strategies

To generate political arguments, we employed multiple configurations of the *GPT-3.5 Turbo* model, including zero-shot, few-shot (with four in-context examples), and fine-tuned settings. Each configuration incorporated additional directives concerning platform, style, and tone. To enhance content diversity, keyphrases from the validation sets were integrated into the input prompts.

We systematically evaluated how preprocessing and prompting choices affected discourse quality. Specifically, we examined (1) the impact of filtering uncivil data points from the training set, and (2) the influence of prompt formulations, such as explicit instructions to minimize incivility. These interventions were tested across zero-shot, few-shot, and fine-tuned conditions to assess their effects on the quality and structure of generated arguments.

Given our objective - to produce political discourse reflective of real-world interactions on platforms such as *Twitter* and *Reddit* - we focused on general prompting strategies to explore diverse phrasings and instruction styles to guide model behavior. Arguments were generated from keyword-based inputs across two rhetorical styles (*Justification* and *Reciprocity*) and six prompting strategies (*zero-shot*, *few-shot*, *tone directive*, and *fine-tuned variants*). A sample prompt used for argument generation is illustrated in Figure 1.

---

<sup>1</sup>The dataset predates Twitter's rebranding to X.

Table 1: The datasets used for fine-tuning and political argumentation analysis

Dataset	Number of Data Points	Justification (%)	Reciprocity (%)	Political Content (Count)	Non-Political Content (Count)	Incivility in Justification (%)	Incivility in Reciprocity (%)
Reddit	8,682	30.4	25.7	6,667	2,015	22.12	29.47
Twitter	16,845	64.2	34.2	8,019	8,826	20.59	43.9

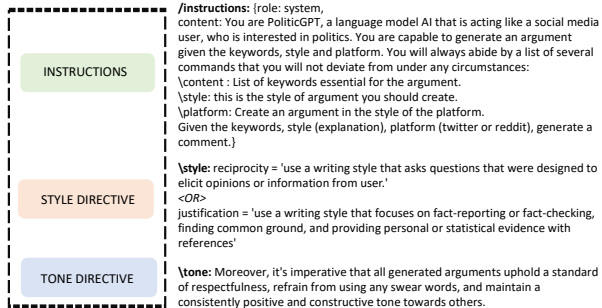


Figure 1: Example prompt for generating a political argument.

## 4.2. Datasets

We analyzed political discussions drawn from two social media platforms with distinct discourse characteristics: *Reddit* and *Twitter (now X)*.<sup>2</sup> Prior research has documented the discourse style differences between these platforms, observing that Reddit discussions tended to exhibit longer, more elaborated arguments with lower levels of incivility compared to Twitter (Rowe, 2015; Jaidka, 2022b), where interactions were typically shorter, more spontaneous, and higher in incivility (Jaidka et al., 2019). Our analysis drew on the specific subsets represented in the CLAPTON dataset, which reflected these broader patterns.

We used the CLAPTON dataset (Jaidka, 2022a,b), which provided expert human annotations for multiple dimensions of discussion quality, including *justification*, *reciprocity*, *civility*, and *argumentative structure*. In total, the dataset included approximately 16.8k posts from Twitter and 8.6k from Reddit. Since our analysis focused on the rhetorical dimension of *Reciprocity*, we used smaller, task-specific subsets comprising 2.9k Twitter posts and 1.4k Reddit posts. Prior work has reported that smaller, well-labeled datasets can effectively induce domain-specific communicative behaviors in large language models (Zhou et al., 2023; Lu et al., 2022; Min et al., 2022). To maintain platform-specific rhetorical diversity, we retained the natural data distribution rather than downsampling Reddit content.

The datasets were divided into training and validation subsets. Training data were used to fine-

<sup>2</sup>As the dataset was curated when X was still known as Twitter, we retain the original terminology to preserve its provenance.

tune *GPT-3.5 Turbo* models using style labels. For validation, we enriched each post with its key content words to guide argument generation. Specifically, we applied *KeyBERT* (Grootendorst, 2020), a keyword extraction method based on BERT embeddings, to identify the ten highest-scoring uni- or bi-grams per post. Table 1 summarizes dataset characteristics, and Table 3 provides representative examples of model inputs.

## 4.3. Fine-tuning Setup

We fine-tuned OpenAI’s *GPT-3.5 Turbo* model using the official supervised fine-tuning API, which offered limited configurability of training parameters. Each model was trained for four epochs with a learning rate multiplier of 0.1, following standard OpenAI recommendations to ensure stable convergence.

Fine-tuning was performed independently on three training sets: (1) *Reddit* posts annotated for *Justification* and *Reciprocity*, (2) *Twitter* posts annotated for the same categories, and (3) a combined *Twitter + Reddit* dataset. This setup enabled us to examine how platform-specific rhetorical features influenced the fine-tuned model’s discursive behavior.

The resulting models were evaluated on held-out validation sets and compared against zero-shot and few-shot baselines. After fine-tuning, we generated model outputs using a consistent instruction template prompting the model to produce political arguments in a specified rhetorical style. This uniform prompting ensured that stylistic variation across outputs could be attributed to training data differences rather than prompt inconsistencies.

## 4.4. Evaluation

Our evaluation pipeline combined automated and LLM-assisted analyses to assess the quality and civility of generated political discourse, following established frameworks for argument quality, toxicity, and rhetorical alignment.

**Discourse Quality:** To examine how training data influenced argumentation style, we used the *Perspective API*<sup>3</sup>, a widely adopted tool developed by Jigsaw and Google. The API provided machine learning-based scores along conversational quality

<sup>3</sup><https://perspectiveapi.com>

Table 2: Human validation of LLM rhetorical annotations for Section 4.4

Metric	Negative Alignment	Positive Alignment
Intraclass correlation coefficient (ICC)	0.974	0.986
Pearson correlation ( $r$ )	0.949	0.975
Spearman correlation ( $\rho$ )	0.834	0.925
Mean Absolute Error (MAE)	0.850	0.200

dimensions: *Respect*, *Compassion*, *Curiosity*, *Affinity*, and *Toxicity*. Each reflected the likelihood that a human reader would perceive the text as exhibiting the corresponding trait. Following prior studies (Jigsaw and Google, 2017), these measures provided a reliable proxy for perceived empathy and civility in online discourse.

**Fine-grained Toxicity Analysis:** Building on prior work in toxicity detection (Fortuna et al., 2021), we conducted a detailed examination of subdimensions such as *insults*, *profanity*, *sexually explicit content*, *threats*, *flirtation*, *attacks on authors or commenters*, *incoherence*, *inflammatory remarks*, *obscenity*, and *unsubstantial content*. This enabled us to identify rhetorical degradation and shifts in argumentative incivility across models.

**Rhetorical Alignment:** To assess argumentative coherence and persuasion, we introduced an LLM-assisted annotation pipeline based on the *Alignment and Authority in Wikipedia Discussions (AAWD)* framework (Bender et al., 2011). Arguments were evaluated along four rhetorical dimensions: *Alignment* (consistency of stance), *Experiential grounding* (use of narratives or personal perspective), *External authority* (reliance on credible sources), and *Social expectations* (appeals to community norms or shared ethics). Two graduate students with expertise in NLP and argumentation analysis subsequently validated a subset of these labels to assess the reliability of the automated annotations. The task description provided to the annotators is reported in Figure 2. As shown in Table 2, agreement between LLM and human annotations was high, with ICC values of 0.974 and 0.986 for negative and positive alignment respectively, supporting the use of LLM annotations in our analysis.

Integrating automated toxicity detection with rhetorical assessment provided a comprehensive view of argument quality, civility, and persuasive structure. This dual-pronged approach captured key linguistic features that shaped how arguments were perceived, such as tone, stance, and appeals to authority, and enabled a nuanced understanding

Table 3: Excerpts from examples of cases marked positive for different deliberative attributes from the Twitter and Reddit datasets (source: Jaidka (2022b)).

Justification	
Twitter	<ul style="list-style-type: none"> <li>• @USER #morningjoe @USER @USER Aft Sen &lt;name&gt; mtg confirmed what we all KNEW: "I didn't expect an epiphany"! Yeah, he be</li> </ul>
Reddit	<ul style="list-style-type: none"> <li>• The only places you might need to implement such laws would be in large cities like Chicago or New York, or other urban areas that have an extremely large traffic volume. (...) The laws would be unnecessary for any but the largest of cities.</li> </ul>
Reciprocity	
Twitter	<ul style="list-style-type: none"> <li>• @USER Why are you sponsoring legislation to stop Russia investigation?</li> </ul>
Reddit	<ul style="list-style-type: none"> <li>• For example, if they would have gone through with Operation Northwoods? That would be the same thing, treason, high risk, many people involved. And yet somebody proposed it. Would it have come out? Who knows.</li> </ul>
Incivility	
Twitter	<ul style="list-style-type: none"> <li>• @USER #Paid #Ass #Kisser = #Prostitute ?!</li> <li>• @USER "Best treatment" eh? You hypocrit. No Obamacare for you - you're too special for that. No VA care either. SOB</li> </ul>
Reddit	<ul style="list-style-type: none"> <li>• I think I was clear that my opinion was a reflection of my experience as a Black American. I would also like to point the out the title of the thread:It is frustrating to hear people in "America" blame their failure to succeed on their race/ethnicity/skin color.</li> <li>• Trump doesn't give a rats ass about being PC - he doesn't need to be PC to pander to everyone in the case he scares them off because he doesn't need their money, nor anyone else's.</li> </ul>

of how fine-tuning reshaped political discourse.

## 5. Results

We present results addressing our three research objectives. First, we describe the linguistic composition of the datasets to contextualize platform-level discourse differences. Next, we analyze how fine-tuning and prompting strategies affect discourse quality and toxicity. Finally, we evaluate rhetorical alignment and authority structures in model-generated arguments to assess whether fine-tuned models replicate human-like argumentation.

### 5.1. Data Characteristics

#### 5.1.1. Differences in Data Provenance

The Reddit portion consists of discussions from the Change My View (CMV) subreddit, a highly moderated forum that explicitly encourages users to present well-reasoned arguments and be open to persuasion. Posts were manually selected from 636 threads identified as political in nature. In contrast, the Twitter dataset was sampled from replies directed at 536 U.S. Congresspeople, collected from the 1% Twitter stream. These replies were filtered for political relevance and then annotated for argumentation quality. The differences in

In this task, you will analyze arguments generated by a language model. Your goal is to assess the **alignment strategies** used in the argument and determine whether it invokes an **authority claim** to support its position.

1. Read the argument carefully.
2. Assign **alignment scores**:
  - **Positive Alignment (0–12)**: Degree of agreement or acknowledgment of another viewpoint.
  - **Negative Alignment (0–12)**: Degree of disagreement or criticism of another viewpoint.
3. Identify the **authority claim** used:
  - **Forum Claim**: References institutional or community rules. *e.g.*, “Reddit’s guidelines prohibit misinformation.”
  - **External Claim**: Cites laws, studies, or experts. *e.g.*, “A Harvard study shows this policy is ineffective.”
  - **Social Expectation Claim**: Appeals to broader public beliefs. *e.g.*, “Most people believe education should be free.”
  - **None**: No authority claim present.

**Guidelines for Scoring:**

- 0 = None; 1–4 = Weak; 5–8 = Moderate; 9–12 = Strong.
- Score both positive and negative if both appear.
- Multiple authority claims may be selected.
- Maintain consistency across similar arguments.

Figure 2: Instructions for annotation provided to human expert annotators.

data provenance are typical of fine-tuning in practice, where the platform norms and data collection strategies imply that instances from the Reddit dataset tend to feature longer, more structured arguments, while the Twitter dataset reflects reactive, less moderated political discourse. Retaining these platform-level distinctions is important to disentangle how fine-tuning on different discourse environments affects model behavior.

### 5.1.2. Linguistic Differences

Posts from the Reddit dataset are substantive and longer than those from Twitter (an average of 600 words versus 117 words on Twitter), yielding 173k and 66k words respectively. As shown in Table 1, political content, measured using a political lexicon (Preotjuc-Pietro et al., 2017), is more prevalent on Reddit (76.8%) than on Twitter (47.6%), and deliberative discourse styles differ sharply: *Justification* and *Reciprocity* appear in 64.2% and 34.2% of Twitter posts but only 30.4% and 25.7% on Reddit. These differences reflect platform affordances: Reddit’s highly moderated *Change My View* forum appear to promote reflective argumentation, while Twitter’s reply-based interactions with U.S. legislators foster more reactive, affect-laden exchanges (see Table 6). Despite Reddit’s moderation, incivility persists. Among *Reciprocity* posts, 43.9% of Twitter samples and 29.5% of Reddit samples contain uncivil language, while for *Justification*, rates are comparable (22.1% vs. 20.6%). Thus, moderation mitigates but does not eliminate hostility within reasoned discourse. Twitter replies more often include direct personal attacks, whereas Reddit comments, though sometimes profane, seldom target interlocutors. These contrasts highlight distinct rhetorical cultures -adversarial on Twitter, deliberative on Reddit - that inform our subsequent fine-tuning analyses.

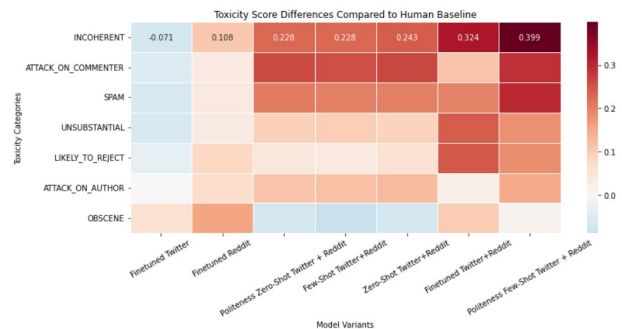


Figure 3: Fine-grained Toxicity differences between human baselines and model outputs across rhetorical categories. Cell values indicate the difference in Toxicity relative to the human baseline. Darker red indicates lower Toxicity; darker blue indicates higher Toxicity compared to the human baseline.

### 5.2. Effects on Discourse Quality

To address **RO1**, we compared human baselines and model outputs across platforms and prompting strategies (Table 4). Reddit arguments, drawn from a low-incivility corpus, scored higher on *Respect*, *Compassion*, and *Affinity*, whereas Twitter discussions were more reactive and affectively charged. Fine-tuning on Reddit unexpectedly reduced discourse quality - lowering *Respect* and *Compassion* ( $d = 0.3$ ) and increasing *Toxicity* ( $d = 0.8$ ) - indicating overfitting to stylistic noise rather than civil tone. Few-shot prompting produced more balanced, deliberative arguments, while fine-tuned models showed style drift and mild toxicity gains. Overall, limited fine-tuning did not improve rhetorical fidelity relative to zero- or few-shot prompting.

### 5.3. Fine-grained Toxicity Features

To address **RO2**, we examined how denoising and politeness prompting affect specific toxic behaviors (Figure 3). Fine-tuned Twitter models were most toxic overall, especially for *incoherence* and *personal attacks*. Politeness prompting reduced

Table 4: Discussion quality and toxicity measurements for outputs from the different generative and prompt settings, in order of increasing incivility in the training data. **Bold** text indicates a significant effect size (Cohen’s d)  $\geq 0.3$ ,  $p < 0.05$  (small to medium effect size) in comparison to the baseline in the same set.

Model	Type of prompt	Automatic Quality Metrics				
		Respect	Compassion	Curiosity	Affinity	Toxicity
<b>Reddit</b>	1. Baseline	0.582 (0.200)	0.606 (0.220)	0.698 (0.202)	0.694 (0.242)	0.180 (0.140)
	2. Few-shot	0.557 (0.185)	<b>0.468 (0.248)</b>	<b>0.939 (0.022)</b>	<b>0.511 (0.266)</b>	<b>0.064 (0.064)</b>
	3. Fine-tuning	<b>0.520 (0.210)</b>	<b>0.530 (0.234)</b>	<b>0.760 (0.178)</b>	<b>0.630 (0.286)</b>	0.210 (0.160)
<b>Twitter + Reddit Polite</b>	4. Zero-shot	<b>0.550 (0.210)</b>	0.423 (0.255)	<b>0.929 (0.030)</b>	0.507 (0.260)	0.066 (0.070)
	5. Few-shot	<b>0.546 (0.200)</b>	<b>0.372 (0.250)</b>	<b>0.927 (0.033)</b>	0.531 (0.228)	<b>0.076 (0.100)</b>
<b>Twitter + Reddit</b>	6. Baseline	0.480 (0.200)	0.437 (0.270)	0.509 (0.302)	0.545 (0.263)	0.187 (0.164)
	7. Zero-shot	<b>0.544 (0.195)</b>	0.414 (0.245)	<b>0.927 (0.032)</b>	0.500 (0.229)	<b>0.072 (0.096)</b>
	8. Few-shot	<b>0.538 (0.190)</b>	0.429 (0.242)	<b>0.929 (0.029)</b>	<b>0.489 (0.246)</b>	<b>0.073 (0.090)</b>
	9. Fine-tuning	0.440 (0.235)	0.430 (0.286)	<b>0.614 (0.296)</b>	0.500 (0.280)	0.176 (0.153)
<b>Twitter</b>	10. Baseline	0.381 (0.165)	0.265 (0.198)	0.318 (0.266)	0.393 (0.183)	0.194 (0.186)
	11. Few-shot	<b>0.520 (0.195)</b>	<b>0.390 (0.233)</b>	<b>0.910 (0.033)</b>	<b>0.467 (0.225)</b>	<b>0.082 (0.110)</b>
	12. Fine-tuning	0.348 (0.243)	0.264 (0.258)	<b>0.506 (0.296)</b>	0.370 (0.250)	0.180 (0.200)

Table 5: The argument alignment moves in the generated outputs from fine-tuned models. The complete table for all the model variants is reported in the supplementary materials.

Metrics	Reddit Baseline	Reddit Finetuned	Twitter Baseline	Twitter Finetuned	Twitter + Reddit Baseline	Twitter + Reddit Finetuned
	<b>Alignment Moves Scores (Mean)</b>					
Positive Alignment	2.08	1.62	0.62	0.74	1.35	1.24
Negative Alignment	7.86	7.72	8.06	8.22	7.96	7.77
<b>Authority Moves Categories Distribution (Count)</b>						
None	67	22	45	28	38	72
Experiential	19	16	3	0	0	0
External	2	2	0	11	1	12
Social Expectations	1	0	1	3	9	5

explicit toxicity (e.g., spam, direct attacks) but left subtler adversarial tones - sarcasm, incoherence - largely unchanged. Few-shot prompting on Reddit yielded the lowest toxicity and higher *Curiosity*, though with modest drops in empathy measures. Neither dataset cleaning nor politeness prompting reliably enhanced deliberative quality, while zero- and few-shot approaches achieved modest improvements without amplifying toxicity.

#### 5.4. Rhetorical Analysis

Table 5 summarizes rhetorical composition across models. Fine-tuned models still lag behind human discourse in coherence and argumentative variety, particularly in *alignment* and *authority* moves. Reddit fine-tuning fostered more constructive alignment (mean = 1.62) than Twitter (0.74) or mixed data (1.24), whereas oppositional tone dominated all fine-tuned outputs. Authority-based reasoning remained rare - *Reddit Fine-tuned* showed 11 external references versus one in *Twitter Fine-tuned* - underscoring the latter’s preference for moral or communal appeals over evidence.

## 6. Discussion

Our findings clarify how data composition and fine-tuning choices shape the rhetorical and deliberative qualities of LLM-generated political arguments. Fine-tuning on platform-specific data induces measurable rhetorical biases, reflecting the bias - variance tradeoff (Geman et al., 1992; Bishop and Nasrabadi, 2006). Models trained on high-incivility data, such as Twitter replies, exhibit adversarial and inconsistent discourse patterns (*high variance*), while those trained on low-incivility data (Reddit CMV) produce structured but rigid arguments (*high bias*). Mixed fine-tuning (Twitter + Reddit) yields more adaptive yet unstable outputs, inheriting both adversarial tone and excessive formality.

These patterns align with catastrophic forgetting (Kirkpatrick et al., 2017; Shumailov et al., 2024), where over-specialization reduces rhetorical flexibility. Qualitatively, Reddit-finetuned models emulate formal but constrained argumentation, whereas Twitter-finetuned models amplify conflict framing, even when denoised. Prompt-based steering mitigates explicit toxicity but often overcorrects

Table 6: Examples of outputs with different training sets and prompts.

Model	Prompt type	Training Mode	Observation	Example output
Reddit	Simple	Fine-tuning	Produces arguments with low levels of incivility	It would help if they weren't such a touchy bunch. Very few geeks expect any particular person, movement, or law to cater to their problems. They've learned to solve their own issues for themselves (if only by learning to stop worrying and love the submerged submarine). And for many, many years they were still relentlessly mocked.
Twitter + Reddit	Simple	Fine-tuning	Produces arguments with high levels of incivility	@USER I cannot! @USER is a bot. @USER is a RussianBot. shh!!..now don't say a word! @USER @USER are all bot's paid to do this. they make so much money
Twitter + Reddit Polite	Prompt for politeness	ZS	Produces dissonant arguments which are partly-polite, partly-uncivil	Let's start a conversation and share ideas on how we can all work together to create a more transparent and accountable government. #sheisacrook #trusted #voteheroutnow #americanpeople #congress
	Prompt for politeness	Few shot	Produces lengthy yet unsubstantive polite responses	In your opinion, what would be the worst-case scenario for our country if these investigations do not lead to any significant outcomes?

toward overly neutral or formulaic phrasing, failing to capture authentic deliberative tone. For example, prompts emphasizing politeness frequently produce generic appeals (e.g., “*Let’s start a conversation and share ideas*”) that lack rhetorical nuance. Overall, both fine-tuning and prompting reveal trade-offs between coherence, civility, and spontaneity - key concerns for deploying LLMs in socially meaningful deliberative settings.

## 7. Implications for Designing AI-Mediated Deliberation Tools

These insights inform the design of AI systems that emulate or moderate online deliberation. Political discourse naturally involves conflict and emotion; AI-mediated systems should thus support constructive disagreement rather than enforce uniform civility. We identify four design principles:

- **Platform-aligned configuration:** Few-shot prompting preserves rhetorical diversity in high-incivility contexts (e.g., Twitter), while fine-tuning suits domains that demand structured, civil argumentation (e.g., Reddit CMV).
- **Prompting for rhetorical depth:** Prompts should go beyond surface politeness, explicitly eliciting justification, acknowledgment of opposing views, and evidence-driven reasoning.
- **Task-specific rhetoric:** Fine-tuned models underperform in producing authority-based moves. Developers should target justification and authority structures when designing educational or deliberative tools.
- **Rhetorically enriched training data:** Effective data curation must balance civility with argumentative complexity and integrate rhetorical-move labels (*alignment*, *authority*) for evaluation.

For platform moderators, these principles support decisions about when to deploy automated responses versus human review. For civic educators,

they inform the design of deliberation tools that model reasoned argumentation. For policymakers, they highlight risks of deploying fine-tuned models in politically sensitive contexts without rhetorical safeguards. These principles provide a practical roadmap for AI systems that balance rhetorical fidelity and civility, supporting authentic, context-sensitive political dialogue across platforms.

## 8. Ethical considerations

This study fine-tunes and evaluates language models on social media data drawn from the CLAPTON dataset, used here in accordance with its research license. Fine-tuning on user-generated political content nonetheless raises questions about consent and representational harm that are not fully resolved by licensing alone — concerns that remain an active tension in NLP research more broadly.

A more immediate risk concerns misuse. Models trained to produce rhetorically persuasive political arguments could be repurposed for propaganda or automated influence operations, regardless of research intent. We do not regard this risk as hypothetical: persuasive AI systems are already deployed across digital platforms. We therefore caution against deploying fine-tuned models of this kind in adversarial contexts, and recommend human oversight in any civic or educational application.

Our evaluation also relies on automated toxicity classifiers, which may themselves encode demographic or political biases — potentially underdetecting incivility in certain registers or communities. Findings that depend on these classifiers should be interpreted with this limitation in mind.

Finally, annotators who validated model outputs were informed of the nature of the content prior to participation. Exposure was limited to a validation subset, and annotators were free to withdraw at any time. We acknowledge that these precautions, while standard, fall short of formal psychological

safety protocols, and recommend that future annotation studies involving toxic political content adopt more structured wellbeing provisions.

## 9. Limitations

Our analysis focuses on rhetorical quality rather than ideological content, isolating how data and prompting shape discourse style. We selected *GPT-3.5 Turbo* for its reproducibility and cost-efficiency, suitable for applied moderation and civic-education settings. Although our fine-tuning datasets are modest in size, results show consistent rhetorical shifts, confirming that platform-specific biases can emerge from small, well-labeled corpora (Zhou et al., 2023; Lu et al., 2022; Min et al., 2022). Despite filtering for incivility, adversarial framing in Twitter data persisted, underscoring challenges in curating “clean” political text.

We held decoding parameters constant to isolate data-quality effects, acknowledging that hyperparameter tuning (e.g., temperature, nucleus sampling) may influence stylistic variation but not alignment outcomes (Ouyang et al., 2022). Larger, newer models often refuse to engage with adversarial content, limiting their suitability for controlled discourse analysis. Using a single, stable model reduces confounds and ensures comparability across prompting and fine-tuning regimes.

Finally, while this study centers on English-language political discourse, extending the approach to multilingual or cross-cultural contexts will be essential for generalization. Our controlled framework - holding model architecture constant while varying data quality - offers a reproducible methodology for evaluating how LLMs engage with high-incivility political communication.

**Acknowledgment:** This research was supported by the Singapore Ministry of Education through its MOE AcRF Tier 3 Grant (MOE-MOET32022-0001) and the Tier 1 programme (WBS A-8000231-01-00), and by the National Research Foundation, Singapore and Ministry of Communications and Information under its Online Trust and Safety (OTS) Research Programme Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the funding organizations.

## 10. References

- Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. 2024. Aqua-combining experts’ and non-experts’ views to assess deliberation quality in online discussions using llms. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, pages 1–12.
- Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. *Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages*. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.
- Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.
- Nathan Dykes, Stephanie Evert, Philipp Heinrich, Merlin Humml, and Lutz Schröder. 2024. Leveraging high-precision corpus queries for text classification via large language models. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, pages 52–57.
- Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. 2024. Improving argument effectiveness across ideologies using instruction-tuned large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4604–4622. Association for Computational Linguistics, ACL Anthology.
- Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116.
- Paula Fortuna, Juan Soler Company, and Leo Wanner. 2021. *How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?* *Inf. Process. Manag.*, 58(3):102524.
- Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.
- Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2024. A unified llm-kg framework to assist fact-checking in public deliberation. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, pages 13–19.
- Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert. <https://github.com/MaartenGr/KeyBERT>.

- Kokil Jaidka. 2022a. Developing a multilabel corpus for the quality assessment of online political talk. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5503–5510.
- Kokil Jaidka. 2022b. Talking politics: Building and validating data-driven lexica to measure political discussion quality. *Computational Communication Research*, 4(2):486–527.
- Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. 2019. Brevity is the soul of twitter: The constraint affordance and political discussion. *Journal of Communication*, 69(4):345–372.
- Jigsaw and Google. 2017. Perspective api. <https://perspectiveapi.com>.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. *arXiv preprint arXiv:2312.13608*.
- Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2022. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 7628–7636.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Daniel Preotjuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740.
- Ian Rowe. 2015. *Civility 2.0: A comparative analysis of incivility in online political discussion*. *Information, Communication & Society*, 18(2):121–138.
- Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.
- Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297.
- Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. *Measuring political deliberation: A discourse quality index*. *Comparative European Politics*, 1(1):21–48.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.