Fine-Tuning LLMs with Noisy Data for Political Argument Generation and Post Guidance

Svetlana Churina and **Kokil Jaidka,**
Centre for Trusted Internet & Community,
National University of Singapore,
Singapore

Abstract

The incivility implicit in social media discourse complicates interventions for post guidance and content moderation for politically sensitive content. Fine-tuning and prompting strategies are critical to mitigate toxicity in such contexts, yet their interplay is less understood. This study investigates the fine-tuning and prompting effects on GPT-3.5 Turbo using Twitter and Reddit datasets of political discussion posts labeled for their discussion quality characteristics. Fine-tuned models on Reddit data scored highest on discussion quality, while combined noisy data led to persistent toxicity. Prompting strategies reduced specific toxic traits, such as personal attacks, in the generated arguments. We develop and validate a new rubric for argument quality evaluation, and close with recommendations for LLM deployment in content authoring, moderation, and intervention contexts.

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

1 Introduction

Large language models (LLMs) play a pivotal role in shaping online discourse, acting as intermediaries in conversations, generating content, and influencing the tone and trajectory of discussions. Their benefits are tempered by pressing concerns, particularly in politically sensitive contexts. First, fine-tuning on noisy data, where noise is any characteristic the model aims to suppress—such as incivility—may inadvertently reinforce rather than mitigate its presence. This aligns with the bias-variance tradeoff in machine learning (Geman et al., 1992; Bishop and Nasrabadi, 2006), where fine-tuning on highly adversarial discourse (e.g., Twitter) risks overfitting to aggressive, incivil, or ideologically extreme rhetorical patterns, reducing generalization ability. The second concerns the role of synthetic data in steering model behavior. If fine-tuning introduces distortions without effective correction mechanisms, recursive training on model-generated data may lead to long-term degradation in discourse quality (Shumailov et al., 2024). In this study, we address these gaps with empirical findings on the efficacy of controlled text generation fine-tuned on political social media posts. Our research focuses on two key objectives:

- **Research Objective 1 (RO1):** To examine the discourse quality of political arguments generated by models fine-tuned on datasets constituting noise, i.e., incivility.
- **Research Objective 2 (RO2):** To examine the effectiveness of mitigative approaches, such as balancing the dataset or revising the prompts, on improving output quality.

To our knowledge, no prior work has systematically investigated how fine-tuning on noisy, political datasets affects the rhetorical coherence and deliberative quality of AI-generated arguments over time. Our work has the following contributions:

- We show that fine-tuned models exhibit platform-specific rhetorical biases. Incivility and adversarial framing are more pronounced in outputs from models fine-tuned on high-variance (e.g., Twitter) than high-structure discourse (e.g., Reddit).
- We develop and validate a novel LLM-assisted annotation pipeline for rhetorical analysis to assess how generated arguments compare in their ability to integrate argumentative elements to emphasize their points.
- We show that prompt-based steering techniques have limited efficacy in mitigating incivility once fine-tuning has reinforced it, while training on heterogeneous data also provides little to no improvements.

Our research focuses on generating political arguments; unlike generic comments, which may

be reactive, neutral, or descriptive, arguments are structured to be directed, evidence-based, and stance-taking (Bender et al., 2011). Arguments are also inherently rhetorical—they are crafted to persuade, counter, or reinforce a position rather than merely provide an observation (Rowe, 2015). In this context, a misstep in argument generation may compromise the integrity of discourse, where respect, compassion, and trust are paramount. Therefore, our findings highlight the long-term risks of training AI models on politically charged discourse and underscore the importance of dataset selection in developing AI-driven deliberative tools.

2 Related Work

Prior research has extensively explored AI-driven argumentation, fact-checking, and discourse quality assessment. Our research focuses on the deliberative quality of arguments. Figueras and Agerri (2024) introduced a novel framework for generating critical questions, illustrating the potential of LLMs to generate deliberative discourse. Lin et al. (2023) proposed a sentence-level counter-argument generation framework, demonstrating that concise, well-structured rebuttals are more persuasive than lengthy, diffuse responses. However, AI-driven argumentation remains fraught with challenges. Disparities in argument persuasiveness have been observed across ideological alignments, as Simmons (2023) found that political moral framing significantly influences argument reception. Additionally, El Baff et al. (2024) demonstrated that LLMs tend to favor liberal perspectives, leading to systematic imbalances in political discourse and motivating our work in characterizing these problems. We aim to examine ways that mitigate or exacerbate these trade-offs through our experiments.

Another key but underexplored challenge is the risk of fine-tuning on low-quality political discourse, which has broad implications for both argument generation and factual integrity. Giarelis et al. (2024) and Dykes et al. (2024) emphasized how training LLMs on noisy, low-quality data sources can reinforce biases and misinformation, ultimately diminishing the reliability of AI-generated arguments. Compounding this issue, recursive training on synthetic outputs can further degrade discourse quality over time. Shumailov et al. (2024) found that when models are iteratively trained on their outputs, they experience model collapse, where argument diversity decreases and biases become more deeply entrenched. We aim to examine these effects through our experiments.

Efforts to directly evaluate the deliberative quality, such as work by Behrendt et al. (2024), highlights the need to consider multiple indices such as civility, rationality, and reciprocity in characterizing political deliberation. Unlike Behrendt et al. (2024), our focus is not on assessing discourse quality but on *applying* deliberative concepts to generate political arguments. Specifically, we examine how AI-generated arguments exhibit **justification**—whether they are grounded in personal experiences, values, or factual references—and how they facilitate **reciprocity**, reflecting engagement in dialogue (Steenbergen et al., 2003). To this end, we leverage the datasets from Jaidka (2022b), which were originally designed for classification tasks, providing annotations on deliberative quality dimensions. While these datasets have primarily been used to analyze and classify discourse patterns, they offer a promising opportunity to extend their application to text generation, allowing us to model justification and reciprocity within AI-generated arguments. This shift from classification to generation enables a deeper exploration of how computational models can identify deliberative features and also reproduce them in structured argumentation.

3 Method

We have assessed the influence of training data on fine-tuned models through comparative analyses. First, we analyzed the outputs of models fine-tuned on training samples curated from different platforms, identifying patterns associated with incivility.

Prompting Strategies: Political arguments were generated using multiple configurations of the GPT-3.5 turbo model, incorporating zero-shot, few-shot, and fine-tuned variants, with additional directives on platform, style, and tone. To enrich argument content, keyphrases from validation sets were included as input. Additionally, we evaluate the impact of dataset preprocessing—such as filtering uncivil data points—and varying prompt formulations, including explicit instructions to reduce incivility. We test these interventions across zero-shot, few-shot, and fine-tuned models to determine their effects on discourse quality and argumentative structure. Given that our goal is to generate political discourse reflective of real-world discussions

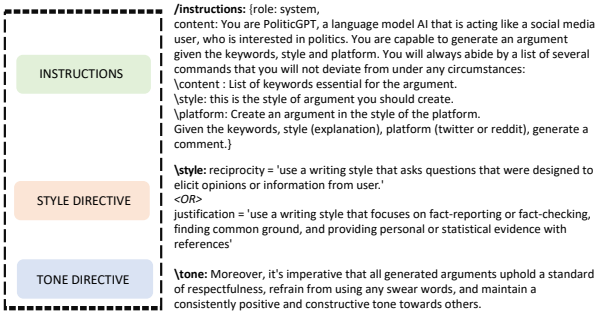

Figure 1: Example prompt for generating a political argument.

on platforms like Twitter and Reddit, we primarily focused on general prompting strategies rather than structured reasoning techniques such as Chain of Thought (CoT). Since political arguments in online discussions are typically concise and direct rather than step-by-step logical derivations, CoT is less applicable in this setting. Instead, we experimented with different phrasings and instruction styles to guide model outputs. Arguments were generated using keyword inputs across two rhetorical styles (Justification, Reciprocity) and six prompting strategies (zero-shot, few-shot, and fine-tuning with and without the tone directive). Figure 1 illustrates a sample prompt for argument generation.

3.1 Evaluation

We use automated and LLM-assisted evaluations to analyze the effects of fine-tuning on noisy political discourse, following established frameworks for assessing argument quality, toxicity, and rhetorical alignment:

- **Fine-tuning effects on discourse quality**: To evaluate the impact of training data on argumentation style, we use the Perspective API to score model-generated arguments on key quality dimensions, including Respect, Compassion, Curiosity, Affinity, and Toxicity.

- **Fine-grained toxicity features**: In line with prior work on toxicity analysis (Fortuna et al., 2021), we conduct a focused examination of subdimensions of toxicity, including insults, profanity, sexually explicit content, threats, flirtation, attacks on authors or commenters, incoherence, inflammatory remarks, obscenity, and unsubstantial content. This allows us to quantify rhetorical degradation and identify potential shifts in argumentative incivility across fine-tuned models.

- **Argument structure and rhetorical alignment**: To assess how fine-tuning affects argumentative coherence, we develop a new LLM-assisted annotation pipeline for rhetorical analysis to assess how generated arguments compare against the baselines (human-written arguments from the datasets). The Alignment and Authority in Wikipedia Discussions (AAWD) corpus (Bender et al., 2011) provided a basis the development of our argument annotation pipeline along three key rhetorical dimensions:

  – Alignment: Whether the argument maintains a clear stance.
  – Experiential grounding: The extent to which arguments incorporate personal experiences, narratives, or user perspectives.
  – External authority: The use of credible sources, factual claims, or references to institutional knowledge.
  – Social expectations: Whether arguments adhere to community norms or invoke shared moral or ethical standards.

Integrating automated toxicity detection with rhetorical assessment provides a comprehensive evaluation of argument quality, incivility, and persuasive efficacy. This dual-pronged approach also enables us to assess whether fine-tuned models enhance deliberative discourse or reinforce adversarial argumentation patterns.

Table 1: Subsets of the CLAPTON Dataset Used for Fine-Tuning and Political Argumentation Analysis

Dataset	Number of Data Points	Justification (%)	Reciprocity (%)	Political Content (Count)	Non-Political Content (Count)	Incivility (%)
Reddit	8,682	30.4	25.7	6,667	2,015	14.8
Twitter	16,845	64.2	34.2	8,019	8,826	20.6

3.2 Datasets

In this study, we analyze political discussions sourced from two distinct social media platforms with **contrasting content characteristics**: **Reddit**, which provides higher-content, lower-noise discussions, and **Twitter (now X)**,[1] which tends to generate lower-content, higher-noise exchanges. Both also have a moderate to high occurrence of

[1]As the dataset was curated when X was still Twitter, we have stuck to the former term to preserve its provenance.

Table 2: Excerpts from examples of cases marked positive for different deliberative attributes from the Twitter and Reddit datasets (source: Jaidka (2022b)).

Justification	
Twitter	• @USER #morningjoe @USER @USER Aft Sen <name> mtg confirmed what we all KNEW: "I didn't expect an epiphany"! Yeah, he be
Reddit	• 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.
Reciprocity	
Twitter	• @USER Why are you sponsoring legislation to stop Russia investigation?
Reddit	• 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.
Incivility	
Twitter	• @USER #Paid #Ass #Kisser = #Prostitute ?! • @USER "Best treatment" eh? You hypocrit. No Obamacare for you - you're too special for that. No VA care either. SOB
Reddit	• 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. • 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.

justification and **reciprocity** in their argumentative styles. These datasets, compiled from prior research (Jaidka, 2022a,b), are human-annotated with discussion quality facets, enabling a structured evaluation of argumentation quality in diverse social media environments. They were used to curate four treatment conditions to evaluate how varying the data source and removing incivility affected the fine-tuned models' performance. Dataset characteristics are provided in Table 1. The datasets were split into training and validation sets, where the training sets were used to fine-tune GPT-3.5-turbo models on inputs with style labels. The held-out validation sets were first pre-processed using KeyBERT to ensure that the model did not infer style or tone directives through the input text. Examples from these sources are provided in Table 2.

4 Results

4.1 Data characteristics

As the first step, we aimed to better understand the content-noise characteristics of the Twitter and Reddit datasets. As seen in Table 1, political content constitutes a larger proportion of the Reddit dataset (76.8%) compared to Twitter (47.6%), suggesting that in these samples, the Reddit discussions are more politically focused than those on Twitter. While both platforms frequently employ Justification and Reciprocity in political arguments, their prevalence differs significantly. On Twitter, 64.2% of posts exhibit Justification, while 34.2% contain Reciprocity. In contrast, Reddit posts display 30.4% Justification and 25.7% Reciprocity.

Despite the presence of deliberative styles, incivility remains a persistent challenge, even within Justification-based arguments. In the Twitter dataset, 20.6% of posts labeled with Justification also exhibit uncivil language (e.g., abusive, racist, threatening, or exaggerated rhetoric). In comparison, 14.8% of Reddit's Justification-based posts contain incivility, reinforcing prior observations that Twitter discourse tends to be noisier and more adversarial than Reddit discussions. In Table 2, we can observe that incivility is implicit across all the rows for the Twitter dataset; furthermore, the incivility in those posts appears to be more targeted at other users. On the other hand, those from Reddit that include general profanity do not attack co-discussants.

While the dataset sizes for Twitter (16.8k posts) and Reddit (8.6k posts) seem imbalanced, our actual training data is smaller due to our focus on Reciprocity: 2.9k Twitter and 1.4k Reddit posts. Additionally, Reddit posts are significantly longer (600 vs. 117 words on average), resulting in a total word count of 173k for Reddit and 66k for Twitter. Simply downsampling Reddit would lead to content loss, making balancing impractical. Instead, we retain the natural distribution to preserve the discourse differences between platforms.

These differences inspired our exploration of whether the dataset could be fine-tuned to generate authentic political arguments and thereby augment the dataset while offering an opportunity to benchmark model outputs against ground truth data.

4.2 Effects on discourse quality

First, for **RO1**, Table 3[2] provides a detailed comparison of the quality metrics across different models and prompting strategies, while the "Baseline" row reflects the quality metrics of human-authored messages in the original dataset. First, a comparison between the baselines comprising human arguments from the low-noise, high-politics Reddit (row 1) and the high-noise, low-politics Twitter (row 6) reveals that the Twitter dataset displays significantly

[2]* indicates an effect size (Cohen's d) >= 0.3 (small to medium effect size) in comparison to the baseline.

Table 3: Discussion quality and toxicity measurements for outputs from the different generative and prompt settings, in order of increasing incivility in the training data. $*$ indicates a significant difference (Cohen's d $>=$ 0.3) in comparison to the baseline in the same set.

Model	**Type of prompt**	**Automatic Quality Metrics**				
		Respect	Compassion	Curiosity	Affinity	Toxicity
Reddit	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)	0.468*(0.248)	0.939*(0.022)	0.511*(0.266)	0.064*(0.064)
	3. Fine-tuning	0.520*(0.210)	0.530*(0.234)	0.760*(0.178)	0.630*(0.286)	0.210 (0.160)
Twitter + Reddit	4. Zero-shot	0.550*(0.210)	0.423 (0.255)	0.929*(0.030)	0.507 (0.260)	0.066 (0.070)
(no incivility)	5. Few-shot	0.546*(0.200)	0.372*(0.250)	0.927*(0.033)	0.531 (0.228)	0.076*(0.100)
Twitter + Reddit	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	0.544*(0.195)	0.414 (0.245)	0.927*(0.032)	0.500 (0.229)	0.072*(0.096)
	8. Few-shot	0.538*(0.190)	0.429 (0.242)	0.929*(0.029)	0.489*(0.246)	0.073*(0.090)
	9. Fine-tuning	0.440 (0.235)	0.430 (0.286)	0.614*(0.296)	0.500 (0.280)	0.176 (0.153)
Twitter	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	0.520*(0.195)	0.390*(0.233)	0.910*(0.033)	0.467*(0.225)	0.082*(0.110)
	12. Fine-tuning	0.348(0.243)	0.264(0.258)	0.506*(0.296)	0.370 (0.250)	0.180 (0.200)

lower quality scores on Respect (Cohens' d = 0.50), Compassion (Cohens' d = 0.85), Curiosity (Cohens' d = 0.63), and Affinity (Cohens' d = 0.63).

We can also compare the two fine-tuning models reported in this table, i.e., fine-tuning in the Reddit (row 3 vs. row 1) and the Twitter (row 10 vs. row 12) cases. On the one hand, we observe that fine-tuning with low-noise, high-content data in Reddit (row 3 vs. row 1) produces significantly lower scores in quality metrics (e.g., Respect and Compassion) (Cohen's d = 0.3) and higher Toxicity compared to the few-shot version (Cohen's d = 0.8 in row 3 vs. row 2), corroborating our concerns about the limited efficacy of prompting when fine-tuning LLMs with small datasets. Therefore, the Table suggests no perceptible improvement in discourse quality metrics after fine-tuning.

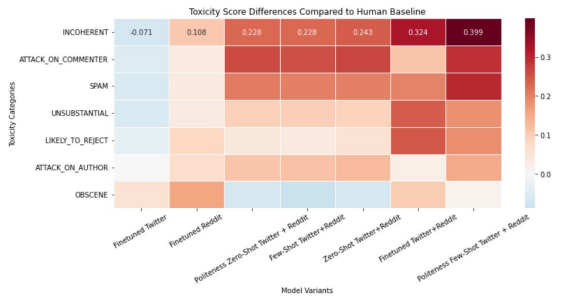

Figure 2: Performance of different model variants in predicting toxicity scores, compared to human baselines. Rows represent different toxicity categories, while columns represent model variants trained under different conditions. The values indicate the difference in toxicity scores relative to human baselines, with deeper shades of red reflecting higher overestimation and deeper shades of blue reflecting higher underestimation.

4.3 Fine-grained toxicity features

For **RO2**, we assess the effectiveness of our efforts to mitigate the noise, such as in the case of using the Twitter + Reddit dataset and further denoising it for training. In the case of attempting to denoise the Twitter + Reddit dataset, we found that the implicit incivility that remained in the dataset made it challenging to curate a sizeable denoised dataset for fine-tuning. For other variants, we can zoom in on the differences in the fine-grained toxicity scores from the Perspective API in Figure 2 where darker shades in the heatmap indicate an increase in the toxicity as compared to the baselines, i.e., the human-authored messages from the dataset. The figure suggests three main takeaways:

- **Politeness-trained models still generate high toxicity in certain categories.** Despite training on more polite data, "Politeness Twitter + Reddit" models exhibit increased toxicity as compared to the mean, for the *INCOHERENT* and *ATTACK_ON_COMMENTER* categories.
- **Finetuned models produce less toxic outputs overall.** "Finetuned Twitter" and "Finetuned Reddit" models generate content that is closer to human baselines, with lower toxicity in the *TOXICITY* and *INCOHERENT*.
- **Few-Shot and Zero-Shot models generate more harmful language.** "Few-Shot" and "Zero-Shot" models consistently produce more toxic content, particularly in *SPAM* and *ATTACK_ON_COMMENTER*.

The findings offer more nuance to Table 3 as they suggest that, contrary to the conflated 'Toxicity' metric in the latter, finetuning may reduce the

Table 4: 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
Alignment Scores (Mean)						
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
Alignment Categories Distribution (Count)						
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

likelihood of toxic completions. In understanding the role of prompting strategies, the Table suggests that the few-shot prompting approach on the Reddit (row 2) subset achieved significantly lower Compassion (Cohens' d = 0.63) and Affinity (Cohen's d = 0.63) as compared to the baseline (row 1), while also achieving significantly lower Toxicity scores (Cohen's d = 0.86) and higher Curiosity scores (Cohen's d = 1.2), suggesting a large deviation from the platform norms. On the other hand, prompting for politeness may still generate harmful language in ambiguous contexts, a finding that is corroborated from the Figure. The poor results with zero- and few-shot prompting suggests that models with limited exposure to safe training data struggle to regulate toxic language. Finally, from the Table 3, we also note no perceptible improvements in the metrics for prompt-based approaches (row 5 vs. row 8), suggesting that the extra effort of hand-curating denoised data was not helpful toward generating more reciprocating nor substantive outputs.

In summary, while our findings suggest the limited utility of fine-tuning with denoised data to better adhere to platform norms in generating political arguments, we instead recommend the use of zero-shot and few-shot methods in noisy contexts, which are more likely to generate higher-quality political arguments than fine-tuned models. For instance, a few-shot approach on the Twitter dataset (row 11 vs row 10) marginally improves the quality on Curiosity (Cohen's d = 0.6) and Affinity (Cohen's d = 0.3) for Twitter + Reddit (row 8 vs. row 6).

4.4 Rhetorical analysis

The rhetorical analysis reported in Table 4 reveals that even fine-tuned models struggle to match the richness of human-written arguments, particularly in alignment and authority moves. The prompts used to obtain LLM-labeled rhetorical moves for the generated outputs are provided in the online repository. Findings highlight the importance of fine-tuning in generating arguments with richer authority moves:

- Alignment Moves: Fine-tuned models exhibit limited argumentative diversity on average, with Reddit Finetuned demonstrating significantly more positive alignment (1.62) than Twitter Finetuned (0.74) and Twitter + Reddit Finetuned (1.24). Negative alignment moves are more prevalent across all finetuned models, with Twitter Finetuned (8.22) leading in negative moves, suggesting a stronger tendency toward argumentative opposition rather than constructive engagement.
- Authority Moves: Authority moves, which capture social expectations, external claims, and forum-based argumentation, also vary significantly across finetuned models. The "None" category dominates across all setups, especially in Twitter + Reddit Finetuned (72 instances), showing that even after fine-tuning, models struggle to integrate authoritative reasoning structures. External claims appear most frequently in Reddit Finetuned (11 instances) but rarely in Twitter Finetuned (1 instance), reinforcing the idea that Reddit's discourse tends to incorporate more external validation. Social expectation moves are highest in Twitter Finetuned (9 instances) but remain modest in Reddit Finetuned (3) and Twitter + Reddit Finetuned (5).

These patterns highlight how platform-specific discourse norms influence the rhetorical framing of model-generated arguments, with Reddit-trained models producing more balanced alignment strategies compared to Twitter-trained models.

Additionally, we conducted a human validation of the alignment scores assigned by the LLM. As shown in Table 6, the agreement between LLM and human annotations is high, with ICC values of 0.974 and 0.986 for negative and positive align-

Table 5: 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 canot! @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 (no incivility)	**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?

ment, respectively. This suggests that LLM annotations are reliable and aligned with human judgments, supporting their use in our analysis. In the full results in the appendix, we have observed that the rhetorical analysis of zero-shot models show the highest variability, with Twitter zero-shot outputs producing the strongest positive alignment (2.12) but failing to integrate deeper argumentative structures (e.g., forum-based engagement, external claims, and social expectations). Few-shot models slightly improve at alignment moves, they introduce an additional risk: prompting choices significantly influence rhetorical strategy, sometimes amplifying biases and incivility.

Table 6: Human validation of LLM rhetorical annotations for Section 4.4

Metric	Negative Alignment	Positive Alignment
Intraclass Correlation Coefficient (ICC2/k)	0.974	0.986
Pearson Correlation (r)	0.949	0.975
Spearman Correlation (ρ)	0.834	0.925
Mean Absolute Error (MAE)	0.850	0.200

5 Discussion and Qualitative Insights

Our findings corroborate the concerns that irrespective of prompting or fine-tuning approaches, data quality during fine-tuning critically influences model performance. Fine-tuning on platform-specific datasets leads to argumentative bias, where models overfit to the dominant rhetorical strategies present in the training data. This aligns with the bias-variance tradeoff in machine learning (Geman et al., 1992; Bishop and Nasrabadi, 2006), where models trained on high-noise data (e.g., Twitter) exhibit high variance and models trained on low-diversity data (e.g., Reddit) display high bias. Additionally, the loss of rhetorical diversity in fine-tuned models mirrors catastrophic forgetting (Kirkpatrick et al., 2017; Shumailov et al., 2024), suggesting that repeated training on platform-specific arguments may degrade general argumentative capabilities over time.

Qualitative analyses (Table 5) suggests that fine-tuning on Reddit results in outputs that closely mimic Reddit's moderated, discussion-oriented style, avoiding overtly hateful language. However, this higher-bias model sacrifices rhetorical diversity, failing to incorporate authority moves, such as social expectations or external claims—essential for persuasive argumentation. This suggests that over-reliance on a structured, low-noise dataset leads to rigid, under-generalized outputs, as also illustrated through the difference in rhetorical complexity of human-authored arguments vs. model-generated responses, particularly in their ability to incorporate alignment moves. These disparities align with the bias-variance tradeoff (Geman et al., 1992; Bishop and Nasrabadi, 2006), where fine-tuning on different datasets produces divergent generalization failures—either overfitting to adversarial discourse (high variance) or underfitting by failing to engage with natural argumentative complexity (high bias). On the other hand, fine-tuning on Twitter + Reddit produces more dynamically adaptive outputs, but at the cost of higher variance, as it inherits both the conversational tone and incivility present in the Twitter subset. The tradeoff between generalization and adversarial speech patterns is evident: Reddit-trained models generate structured but less engaging arguments, while Twitter-trained models risk amplifying toxic discourse.

So what should researchers do? Our third takeaway is regarding the role of prompting strategies.

Enforcing politeness in Twitter + Reddit outputs results in overly formal and often dissonant responses. For example, the output

> "Let's start a conversation and share ideas on how we can all work together"

lacks the spontaneity and engagement typical of natural social media discourse, as seen in Table 2. This overcorrection suggests that prompting alone does not adequately balance argumentative richness and civility—a challenge exacerbated by the high variance of Twitter-trained models, which exhibit strong fluctuations in tone and engagement based on prompt constraints. Our anecdotal observations are further corroborated by quantitative metrics in Table 3, which demonstrate substantial differences in argumentative quality across few-shot and fine-tuned model outputs. The experiments with Twitter + Reddit show that while zero-shot and few-shot prompts can degrade the discussion quality on average, they improve the quality of the arguments as compared to the baseline in the case of Twitter, the noisiest dataset. They may also amplify the effect of fine-tuning to yield greater improvements in discussion quality, but this effect is conditional on the data quality.

6 Implications for Post guidance

Our study on fine-tuning LLMs for political argument generation offers practical insights into enhancing online discourse quality. Table 9 presents a structured overview of these approaches. Complementing automated assessments with rhetorical analysis provides deeper insights into argumentative structure and integrity. Key approaches for practice are summarized in the framework reported in the appendix and include, in order of priority:

- **Configuring models for noisy domains:** Use zero-shot models for limited data, few-shot models to build on examples, and fine-tuning for platform-specific stylistic alignment.
- **Designing prompts:** Tailor prompts with clear stylistic instructions to promote deliberative discourse (Bender et al., 2011).
- **Understanding task context**: If the task requires structured political arguments, apply strategies emphasizing justification and reciprocity (Steenbergen et al., 2003).
- **Managing dataset quality:** Use filtering techniques to reduce incivility while retaining argumentative richness, yet be mindful that these efforts are only effective at large data sizes (Dykes et al., 2024).
- **Evaluating discourse quality:** Combine automated tools and rhetorical analysis to track and improve model performance (Behrendt et al., 2024).

These approaches would support the development of LLMs capable of generating high-quality, civil political arguments that encourage constructive engagement.

7 Conclusion and Recommendations

Our findings emphasize that few-shot prompting can improve the quality of political arguments in noisy contexts such as Twitter. On the other hand, fine-tuning with noisy data can adhere to the civil and rhetorical expectations of social media platforms. Our work offers nuance to prior work suggesting that targeted training data from similar platforms is crucial for effective task generalization, clarifying that "clean data" is a necessary conditional closely tied to the fine-tuning objective. The distribution of labels, such as incivility indicators, plays a pivotal role in fine-tuning outcomes.

Ultimately, we offer the following recommendations for platforms aiming at designing moderation interventions for social media platforms. First, we recommend that fine-tuning with curated datasets is appropriate for platform-specific assistants to align with their unique tonal and conversational norms. For example, a Twitter-focused assistant could address brevity and reduced toxicity, while an Instagram version could focus on empathy and community building. Second, on the authoring side, we observed that LLM-generated posts score significantly higher on Curiosity; therefore, they could be used to seed conversations that promote curiosity and empathy. Authoring tools could also be pre-configured for post guidance for specific purposes, such as casual conversation versus professional or political discussions, or for different tones such as "respectful debates" or "fact-driven explanations." Finally, in moderation, auto-moderation tools in noisy contexts could use few-shot prompting to flag potentially harmful posts. In discussions on contentious topics, their suggestions can guide conversations toward respectful and constructive discourse.

8 Limitations

Our focus was not on the ideological nor social biases that may guide argument content, as in this study, we focused only on its quality dimensions. Fine-tuning and reporting results with GPT-3.5 was considered appropriate due to its balance of performance and computational efficiency, its smaller size and lower resource requirements, which make it more practical for applications in cost-constrained environments and accessible for reproducibility by other researchers.

One potential concern is data contamination—whether the model has seen the CLAPTON dataset during training. However, this is not an issue in our case, as CLAPTON was released in 2022, while GPT-3.5 Turbo's training data cutoff was before January 2022. This ensures that the model had no prior exposure to our dataset. Additionally, we specifically chose GPT-3.5 over newer models to prevent reliance on any potential memorized content, ensuring that argument generation is based on learned generalization rather than direct recall. Future work could explore the impact of using more recent models while applying safeguards against contamination.

When applying these models to detect or generate political arguments in different cultural settings, there are risks associated with inaccurate predictions and stereotypical content generation. Exploring newer or domain-specific models, while potentially fruitful, was outside the scope of this work and is identified as an avenue for future research. Finally, while we only evaluated GPT-3.5 in a constrained context for greater control, broader comparisons across diverse models or datasets could provide further generalizability.

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.

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.

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*.

Ian Rowe. 2015. Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2):121–138.

Ilia 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.

A Instructions for Alignment and Authority Annotation

Annotators (researchers from our lab with expertise in NLP and argumentation analysis) were provided with generated text from either Twitter or Reddit. Their task was to assign **positive and negative alignment scores** (ranging from 0 to 12) and categorize the **authority claim** used in the argument. The following task description was provided to the annotators:

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

1. Read the argument carefully.
2. Assign **alignment scores**:
   - **Positive Alignment (0-12):** Measures the degree of agreement, support, or acknowledgment expressed towards another participant's viewpoint.
   - **Negative Alignment (0-12):** Measures the degree of disagreement, opposition, or criticism expressed towards another participant's viewpoint.
3. Identify the **authority claim** used in the argument:
   - **Forum Claim:** The argument references rules, policies, or contextual norms of a platform, institution, or specific community. *Example: "Reddit's guidelines prohibit misinformation, so this post should be removed."*
   - **External Claim:** The argument cites an external authority, such as a law, book, research study, or expert opinion. *Example: "According to a study from Harvard, this policy is ineffective."*
   - **Social Expectation Claim:** The argument references beliefs, intentions, or expectations of groups beyond the immediate discussion. *Example: "Most people believe that education should be free and accessible."*
   - **None:** The argument does not reference any authority claim.

Guidelines for Alignment Scoring:

- 0 = No alignment present(neutral, off-topic, or lacking engagement).
- 1-4 = Weak alignment (minor agreement or disagreement).
- 5-8 = Moderate alignment (clear but not extreme support or opposition).
- 9-12 = Strong alignment (explicit agreement, praise, or strong criticism/insult).

Final Notes:

- If both positive and negative alignment are present, score both accordingly.
- Can be selected more than one authority claim per argument.
- Be consistent—similar arguments should receive similar scores.

Table 7: The argument alignment scores in the generated outputs.

Metrics	RedditFinetuned	TwitterFinetuned	Twitter + RedditFinetuned	Zero-ShotTwitter	Zero-ShotReddit	Few-ShotTwitter
			Alignment Scores (Mean)			
Positive Alignment	1.62	0.74	1.24	2.12	1.36	1.16
Negative Alignment	7.72	8.22	7.77	1.76	0.92	1.04

Table 8: Detailed argument alignment categories in the generated outputs.

Alignment Category	RedditFinetuned	TwitterFinetuned	Twitter + RedditFinetuned	Zero-ShotTwitter	Zero-ShotReddit	Few-ShotTwitter
None	28	38	72	45	22	-
External	11	1	12	2	2	-
Forum	2	2	1	1	1	-
Social Expectations	3	9	5	1	1	-
Forum, External	2	0	4	-	-	-
External, Social Expectations	2	0	2	-	-	-

Table 9: Framework for Fine-Tuning LLMs for Political Argument Generation

S/N	Step	Description
1	**Understanding Task Context**	**Identify the task requirements.** Determine if the task involves structured political arguments requiring justification and reciprocity (Steenbergen et al., 2003). If yes, proceed to Step 2. If not, simpler models may suffice.
2	**Managing Dataset Quality**	**Evaluate and preprocess data.** Use filtering techniques to reduce incivility while retaining argumentative richness (Shumailov et al., 2024). Ensure sufficient dataset size, as filtering is more effective with larger datasets (Dykes et al., 2024).
3	**Configuring Models for Noisy Domains**	**Select an appropriate model configuration.** - Use zero-shot models for limited datasets. - Use few-shot models to build on examples. - Fine-tune models for platform-specific stylistic alignment (Jaidka, 2022b).
4	**Designing Prompts**	**Craft clear, purpose-driven prompts.** Tailor prompts with explicit stylistic instructions to promote deliberative discourse (Bender et al., 2011). Focus on rhetorical features such as justification, reciprocity, alignment, and experiential grounding.
5	**Evaluating Discourse Quality**	**Assess the model's performance.** Combine automated tools like the Perspective API with rhetorical analysis to track and improve performance (Behrendt et al., 2024). Evaluate key discourse dimensions: Respect, Compassion, Curiosity, Affinity, and Toxicity.