PTCG: Persona-guided Tree-based Counterargument Generation

Anonymous ACL submission

Abstract

The ability to generate counterarguments is important for critical thinking and balanced discourse, yet existing approaches typically produce only a single counterargument, failing to capture the diversity and persuasiveness required in real-world debates. To address this limitation, we propose Persona-guided Tree-based Counterargument Generation (PTCG), a framework that combines Tree-of-Thoughts–inspired step-wise generation and pruning with speaker persona selection. By estimating the author's persona from the original argument and incorporating speaker personas representing distinct perspectives, PTCG operationalizes perspective-taking and enables the generation of diverse counterarguments. Results from LLM-as-a-Judge, classifier-based assessment, and human evaluations indicate that PTCG shows consistent improvements in both the diversity and persuasiveness of counterarguments compared to baseline methods.

1 Introduction

The ability to generate effective counterarguments is a growing area of interest in computational argumentation (Wang et al., 2023). A counterargument is not merely an expression of disagreement but a reasoned response that challenges an argument, exposing assumptions, logical gaps, or alternative perspectives. Engaging with counterarguments is widely recognized as a key mechanism for fostering critical thinking, as it encourages individuals to evaluate evidence, weigh competing viewpoints, and refine reasoning (Ennis, 2015; Dekker, 2020). The ability to generate effective counterarguments is especially important in contexts such as political debates, legal reasoning, and online discussions, where opposing views promote balanced discourse and better-informed decision-making (Li et al., 2020; Behrendt et al., 2025; Zhang and Ashley, 2025; Gray et al., 2025).

Existing counterargument generation approaches suffer from two critical limitations: First, most methods generate a counterargument focusing on a single point, failing to capture diverse perspectives on the subject. This limitation stems from their reliance on a single strategy, such as attacking a weak premise (Alshomary et al., 2021), directly refuting the conclusion (Alshomary and Wachsmuth, 2023), or pointing out logical flaws (Lin et al., 2023). Second, counterarguments produced by existing approaches, including those generated with large language models (LLMs), often lack persuasiveness. While Chen et al. (2024) demonstrate the potential of LLMs in argument generation, subsequent studies (e.g. Lu et al. 2025, Plenz et al. 2025) reveal that generated arguments often lack value-based reasoning and scenario-driven perspectives essential for persuasiveness in human argumentation. In addition, the aforementioned issue of narrow perspective also negatively impacts the overall persuasiveness of the counterargument. This suggests that persuasiveness across diverse audiences depends on presenting arguments from multiple perspectives, making the resolution of this limitation a central challenge for counterargument generation.

To address these issues, we propose Persona-guided Tree-based Counterargument Generation (PTCG), a framework grounded in the theory of perspective-taking. Perspective-taking, widely studied in social psychology, refers to the practice of imaginatively adopting others' standpoints when evaluating or constructing arguments. It has been shown to foster empathy, reduce bias, and encourage reasoning from viewpoints different from one's own (Batson et al., 1997; Green and Brock, 2000). PTCG operationalize perspective-taking by guiding counterargument generation with pre-defined personas. Specifically, after estimating the original author's persona from their argument, the framework selects personas from both similar and contrasting predefined persona clusters and uses them to guide

counterargument generation. The predefined persona clusters were created to organize about 50,000 personas, reducing redundancy and enabling efficient selection of diverse perspectives. This design enables models to move beyond default stances and generate counterarguments that reflect a wider range of perspectives. In addition, PTCG incorporates a Tree-of-Thoughts (ToT)–inspired reasoning procedure (Yao et al., 2023). Multiple candidate reasoning paths called *plans* for generating counterarguments are first generated, evaluated, and pruned. Only the most promising plans are then expanded into full counterarguments. This iterative process of generation and selection allows for diverse and persuasive reasoning paths. By combining persona conditioning with a step-wise reasoning process inspired by Tree-of-Thoughts (ToT) (Yao et al., 2023), the framework iteratively generates and selects candidate plans and counterarguments, ultimately producing multiple counterarguments that capture both diversity and persuasiveness.

To evaluate PTCG, we conduct experiments with multiple LLMs using the discussion threads from the ChangeMyView subreddit, which cover a diverse range of real-world topics. We combine LLM-as-a-Judge to assess the diversity and persuasiveness—general and targeted—as well as the stance and quality of the generated counterarguments. We further incorporate classifier-based metrics for the key dimension of persuasiveness, providing a more comprehensive evaluation. In addition, we conduct human evaluation, which not only complements the LLM-as-a-Judge results but also demonstrates persuasiveness across a diverse pool of evaluators, providing further evidence of applicability to the real audience. Across these evaluations, PTCG consistently outperforms the baselines, producing counterarguments that are more diverse, persuasive, and higher in overall quality.

2 Related Work

2.1 Argument Generation

Early work on argument generation framed the task as a largely symbolic or rule-driven process. Sato et al. (2015) proposed a pipeline-based debating system, but such rule-based approaches were often brittle and lacked scalability. To improve flexibility, Wachsmuth et al. (2018) employed hand-crafted rhetorical patterns, while Hua and Wang (2018) introduced neural argument generation augmented with retrieved evidence to ground arguments in factual content. Together, these approaches laid important groundwork for argument and counterargument generation.

2.2 Counterargument Generation

Recent studies on counterargument generation have primarily focused on exploiting explicit argument structures or predefined strategies. For instance, Alshomary et al. (2021) generate counterarguments by attacking weak premises, while Alshomary and Wachsmuth (2023) guide generation by modeling the conclusion of the original post. Lin et al. (2023) instead operate at the sentence level, producing concise counterarguments for individual statements. However, these approaches typically rely on a single argumentative strategy and generate only one counterargument, limiting their ability to explore diverse perspectives. Unlike prior work, our framework explicitly aims to produce multiple distinct counterarguments grounded in diverse personas and reasoning trajectories. To achieve this, we ground reasoning in diverse personas and explore alternative argumentative paths through tree-based step-wise generation.

While Hu et al. (2025) recently proposed a multi-agent framework in which personas debate to synthesize a single essay, our objective is fundamentally different. Rather than merging multiple voices into a unified output, our approach allows each persona to independently develop its own reasoning. As a result, diversity across personas is not merely an intermediate mechanism for producing a single response, but a primary objective of the framework itself. By systematizing persona selection based on distance, our approach generates multiple distinct and persuasive counterarguments that reflect a broader spectrum of perspectives.

2.3 Perspective-Taking

Psychological studies highlight the persuasive power of perspective-taking and narrative immersion. Batson et al. (1997) show that imagining how others feel fosters empathy and altruistic motivation. Green and Brock (2000) and Mar and Oatley (2008) suggest that narrative "transportation" enables simulated experience, which can influence attitudes more deeply than factual exposition. More recently, Bullock et al. (2021) argue that narratives are persuasive partly because they are processed more fluently than non-narrative formats. These findings motivate our use of perspective-taking-

based generation to induce perspectival engagement and simulate meaningful disagreement. Moreover, perspective-taking enables the incorporation of diverse viewpoints, making it possible to generate multiple counterarguments for a single post.

2.4 Diverse-Audience Persuasion

Lukin et al. (2017) show that persuasiveness depends on audience traits, motivating the use of personality-based analysis in argumentation. Building on this, studies have explored personalization in persuasive dialogue: Wang et al. (2019) adapt strategies based on user traits, and Al Khatib et al. (2020) incorporate debaters' characteristics to improve persuasiveness prediction. Recent work further demonstrates that LLMs can generate more persuasive messages when tailored to psychological profiles (Matz et al., 2024), modulate linguistic features according to personality cues (Mieleszczenko-Kowszewicz et al., 2024), and role-play personas to enhance empathy and strategy distribution (Yang et al., 2025). Building on this line of research, our work develops an approach that generates diverse counterarguments with awareness of audience diversity, aiming for greater persuasiveness.

3 Persona-guided Tree-based Counterargument Generation

3.1 Task Description

We define the task of **multiple distinct counterargument generation** to evaluate the ability to generate persuasive counterarguments that cover a broad spectrum of perspectives on a given argument. Specifically, the input is an *argument* consisting of a claim and one or more premises supporting it, and the output is a set of *counterarguments*, each presenting a distinct perspective that challenges the original argument. We detail the evaluation criteria in Section 4.3.

3.2 Perspective Taking: Persona Selection

To realize perspective-taking in counterargument generation, we draw on the notion of personas as proxies for diverse viewpoints. Prior work has shown that incorporating personas into dialogue models allows them to go beyond generic language modeling, producing responses that are more consistent, human-like, and reflective of underlying experiences or viewpoints (Zhang et al., 2018; Moon et al., 2024). We adopt the PersonaHub dataset (Ge et al., 2024), a large-scale repository of about 50,000 personas. Since using all personas is redundant and costly, we cluster similar ones into groups. Each persona is embedded using OpenAI's text-embedding-3-large (OpenAI 2024b; 3,072-dimensional), reduced to a 50-dimensional representation via UMAP (McInnes et al., 2018), and clustered into 39 groups[1] using HDBSCAN (McInnes et al., 2017). Subsequently, we leverage inter-cluster distances to select counterargument speakers from distinct groups, ensuring that the generated counterarguments embody genuinely different perspectives. To help readers better understand the semantic coherence and diversity encoded in these clusters, we provide representative persona examples from several clusters in Section A2. These examples illustrate how grouping similar personas enables systematic selection of viewpoints that truly differ in background, expertise, and worldview—an essential property for perspective-taking in counterargument generation.

3.3 Persona-guided Tree-based Counterargument Generation (PTCG) Framework

Our method, which we call Persona-guided Tree-based Counterargument Generation (PTCG; See Figure 1 and Algorithm 1), integrates the Original Poster (OP) persona estimation, clustering-based speaker persona selection, and a Tree-of-Thoughts (ToT) (Yao et al., 2023)-inspired step-wise generation process (Plan and Counterargument Generation and Selection).

Step 1: Original Poster (OP) Persona Estimation Given an original post, we estimate the OP persona using an LLM-based estimation prompt. Here, the OP refers to the author of the original post. This estimated persona represents the OP's beliefs, values, and worldview, and conditions subsequent counterargument generation. The detailed prompt is provided in Appendix Figure A4.

Step 2: Speaker Persona Selection To balance diversity with interpretability, we set the number of personas to three, supported by observations from the CMV dataset where posts with multiple delta-awarded comments—indicating diverse, high-

[1] As shown in Appendix Table A10, the configuration with 39 clusters was selected as it achieves the highest Silhouette Score (Rousseeuw, 1987) while also maintaining a strong Calinski–Harabasz Index (Caliński and Harabasz, 1974), indicating the best balance between cohesion and separation.

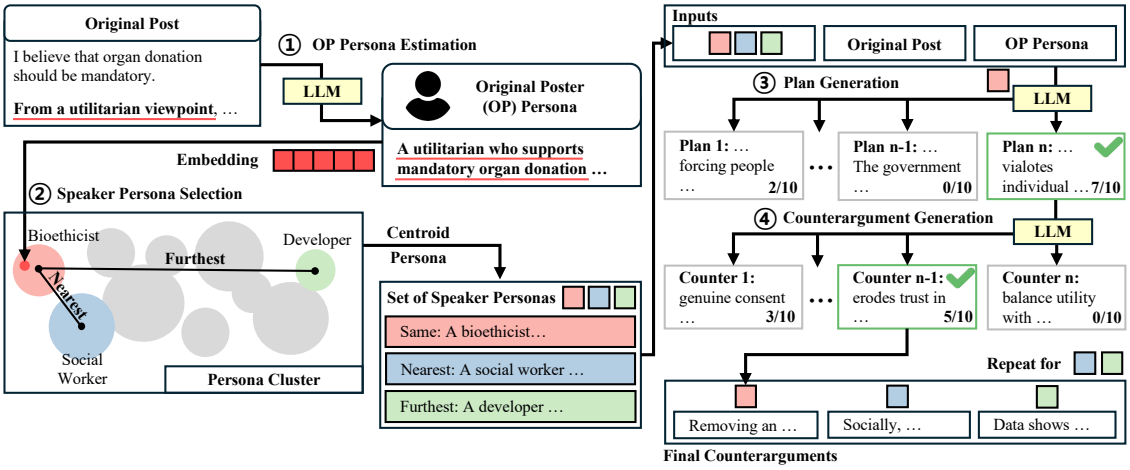

Figure 1: Persona-guided Tree-based Counterargument Generation (PTCG). An LLM first extracts the author's persona and conditions on distinct personas to produce multiple candidate reasoning plans. These plans are evaluated and pruned in a tree-based step-wise manner, and the most promising ones are expanded into three final counterarguments, each grounded in a different persona.

quality counterarguments—rarely exceeded three (see Appendix Figure A3). We select personas based on their cluster distance from the OP: one from the *same* cluster, one from the *nearest* cluster, and one from the *furthest* cluster. For each cluster, no persona exactly matched the centroid; therefore, we used the closest in embedding space. This setup ensures that the three chosen personas collectively reflect perspectives ranging from highly aligned to markedly divergent, thereby systematically probing how cluster distance influences counterargument generation and persuasiveness.

Step 3: Plan Generation and Selection Once the OP persona and speaker personas are determined, the generation process proceeds in a tree-based step-wise manner inspired by Tree-of-Thoughts (Yao et al., 2023). For each persona, the LLM generates three candidate plans, each outlining a persuasive strategy for counterargument generation. A voting procedure then evaluates these candidates in terms of whether they effectively use the contrast between personas, apply a strong strategy, and present their reasoning clearly and logically. The most promising plan is selected among the three, resulting in one best plan per persona. For each speaker persona, one finalized plan is thus determined and passed to the subsequent generation stage. The prompts used for plan generation and selection can be found in Figure A5 and Figure A6, respectively.

Step 4: Counterargument Generation and Selection Building on the selected plans, the finalized plan for each speaker persona is explicitly used, along with the original post, OP persona, and the designated speaker persona, to guide counterargument generation. The LLM then generates three candidate counterarguments per persona, each following the selected plan while reflecting the persona's distinct perspective. A voting procedure evaluates these candidates based on whether they leverage the contrast between personas, directly challenge the original argument, and are specific, persuasive, and logically consistent. Among the three candidates, the most persuasive counterargument is selected as the final output for that persona. This process is repeated for each persona, with one best counterargument selected per persona, yielding three diverse counterarguments from the same, nearest, and furthest clusters that collectively capture multiple perspectives. The prompts for this step are identical to those in Step 3, shown in Figure A5 and Figure A6, respectively.

4 Experiments

4.1 Dataset

To assess the ability to generate diverse and persuasive counterarguments, we obtained the dataset of 847 *ChangeMyView* (CMV) subreddit[2] posts (i.e., arguments) from Academic Torrents[3], each paired with three comments (i.e., counterarguments) that have successfully persuaded the original poster as

[2] https://www.reddit.com/r/changemyview/
[3] https://academictorrents.com/details/1614740ac8c94505e4ecb9d88be8bed7b6afddd4

gold-standard persuasive counterarguments. On CMV, each post consists of a title summarizing the main claim and a body providing the premises supporting the claim. Original posters award a delta (Δ) to comments that successfully change their view, which we treat as quality counterarguments. We first collected 72,999 posts from CMV, spanning the years 2013 to 2023. Then, to ensure diversity, we filter for posts with three delta-awarded comments, yielding a dataset of 847 post-comments pairs.

4.2 Baselines

To establish meaningful comparisons, we adopt baselines from both prior counterargument generation research and representative large language models.

Backbone Language Models

- **LLaMA3.1-8B** (Grattafiori et al., 2024). An instruction-tuned model from the LLaMA3.1 family trained on large-scale publicly available datasets. In particular, we use `LLaMA3.1-8B-Instruct` as the primary backbone for our proposed method.

- **LLaMA3.2-1B** (Grattafiori et al., 2024). A more lightweight variant of the LLaMA family designed for efficient inference. We employ the `LLaMA3.2-1B-Instruct` model to evaluate performance on a significantly smaller parameter scale.

- **Qwen3-8B** and **Qwen-14B** (Team, 2025). A family of state-of-the-art large language models with strong reasoning and language capabilities. We evaluate both `Qwen3-8B` and `Qwen3-14B` models in non-thinking mode to provide comparisons across different architectures and model sizes.

Counterargument Generation Methods

- **Vanilla Generation (Vanilla)**. In this setting, the model is prompted to directly produce three counterarguments in a single pass, as illustrated in Figure A7.

- **Argument Undermining** (Alshomary et al., 2021). This method identifies weak premises in the original post and generates counterarguments by attacking them. For fair comparison, we first employ a weak-premise identification model to select the top three weak premises. We then generate one counterargument per premise, resulting in three outputs.

- **Joint One-seq** (Alshomary and Wachsmuth, 2023). This method infers multiple conclusions from the premises in the original post and uses them as the basis for counterargument generation. Following the original formulation, we sample three alternative conclusions and generate a counterargument for each.

4.3 Evaluation Metrics

We evaluate our method across two dimensions: Persuasiveness and Perspective Diversity. We employ *LLM-as-a-Judge* (Zheng et al., 2023; Gu et al., 2024), as it provides more reliable evaluations than traditional lexical-overlap-based metrics such as BLEU and ROUGE, which often diverge from human judgment (Celikyilmaz et al., 2020; Hu et al., 2024). we use GPT-4o-mini as the judge model. For targeted persuasiveness, which measures how well the counterarguments resonate with the original poster, we additionally use data-driven *classifier-based scores*.

Persuasiveness While general persuasiveness is important, it is equally crucial for counterarguments to be persuasive with respect to the original poster's context. To capture both, we evaluate *general persuasiveness* using only the title (claim) of the original post, and *targeted persuasiveness* using both the title and body (premises), which assesses resonance with the original poster's context and simulates realistic dialogue. Since each input yields multiple counterarguments, we report the average persuasiveness score to reflect overall persuasiveness. The prompts used for these two evaluations are provided in Appendix Figures A8 and A9.

Perspective Diversity Producing counterarguments that embody genuinely distinct viewpoints, beyond superficial rewordings, is difficult. True diversity requires capturing ideological, emotional, or experiential variation that reflects distinct perspectives. We therefore evaluate whether the outputs move beyond lexical variation and invoke deeper interpretive frames not explicitly given in the input. The prompt for evaluating perspective diversity is provided in Appendix Figure A10.

Backbone	Method	*Persuasiveness*		*Perspective Diversity*
		General	Targeted	
LLaMA3.1-8B	Vanilla	$7.79_{\pm 0.0031}$	$7.59_{\pm 0.0023}$	$3.78_{\pm 0.0021}$
	Argument Undermining	$4.23_{\pm 0.0040}$	$4.34_{\pm 0.0021}$	$2.76_{\pm 0.0015}$
	Joint One-seq	$4.62_{\pm 0.0055}$	$4.69_{\pm 0.0067}$	$3.15_{\pm 0.0066}$
	PTCG (Ours)	$\mathbf{8.32}_{\pm 0.0025}$	$\mathbf{7.93}_{\pm 0.0032}$	$\mathbf{4.28}_{\pm 0.0024}$
LLaMA3.2-1B	Vanilla	$6.63_{\pm 0.0031}$	$6.44_{\pm 0.0039}$	$3.34_{\pm 0.0032}$
	Argument Undermining	$3.95_{\pm 0.0034}$	$3.92_{\pm 0.0040}$	$2.49_{\pm 0.0043}$
	Joint One-seq	$3.56_{\pm 0.0056}$	$3.44_{\pm 0.0027}$	$2.54_{\pm 0.0048}$
	PTCG (Ours)	$\mathbf{7.11}_{\pm 0.0022}$	$\mathbf{6.84}_{\pm 0.0016}$	$\mathbf{3.61}_{\pm 0.0029}$
Qwen3-8B	Vanilla	$8.18_{\pm 0.0024}$	$7.94_{\pm 0.0029}$	$3.59_{\pm 0.0069}$
	Argument Undermining	$5.54_{\pm 0.0033}$	$5.59_{\pm 0.0041}$	$2.13_{\pm 0.0051}$
	Joint One-seq	$3.81_{\pm 0.0009}$	$3.80_{\pm 0.0062}$	$2.61_{\pm 0.0062}$
	PTCG (Ours)	$\mathbf{8.72}_{\pm 0.0026}$	$\mathbf{8.22}_{\pm 0.0025}$	$\mathbf{4.71}_{\pm 0.0070}$
Qwen3-14B	Vanilla	$8.33_{\pm 0.0022}$	$8.02_{\pm 0.0031}$	$4.19_{\pm 0.0051}$
	Argument Undermining	$5.38_{\pm 0.0016}$	$5.56_{\pm 0.0030}$	$2.17_{\pm 0.0035}$
	Joint One-seq	$4.10_{\pm 0.0060}$	$4.16_{\pm 0.0027}$	$2.79_{\pm 0.0024}$
	PTCG (Ours)	$\mathbf{8.73}_{\pm 0.0031}$	$\mathbf{8.29}_{\pm 0.0008}$	$\mathbf{4.28}_{\pm 0.0062}$

Table 1: Evaluation results of baselines and PTCG. All scores are averaged over five runs and reported with standard deviations. *Persuasiveness* is reported for general and targeted settingsa (1–10 scale), *Perspective diversity* evaluates the variety of viewpoints reflected in generated counterarguments (1–5 scale).

5 Results and Analysis

5.1 Rating-scale Evaluation

The reported scores were averaged across five runs to mitigate random variation. As shown in Table 1, PTCG consistently outperforms baseline methods across key dimensions, persuasiveness and perspective diversity, when compared within the same backbone model. The persuasiveness scores indicate that PTCG-generated counterarguments are more effective for both a general audience and the original poster (OP).

While other baselines like Argument Undermining and Joint One-seq often show a significant drop in persuasiveness, PTCG maintains strong performance, surpassing the Vanilla baseline. The improvement in perspective diversity is also clear. While other methods struggle to maintain diversity, PTCG consistently stays above the Vanilla scores. These results show that PTCG is a robust and scalable solution that effectively improves both the quality and the variety of perspectives across different model scales.

5.2 Classifier-based Evaluation

To complement LLM-based judgments, we use a Delta classifier trained on CMV delta-award annotations to assess targeted persuasiveness. Results are shown in Table A7. Overall, PTCG demonstrates strong performance across all backbones and consistently outperforms prior counterargument generation methods.

Backbone	Method	Avg.	Max
LLaMA3.1-8B	Vanilla	**0.84**	0.88
	Argument Undermining	0.64	0.83
	Joint One-seq	0.42	0.78
	PTCG (Ours)	0.81	**0.90**
LLaMA3.2-1B	Vanilla	0.80	0.86
	Argument Undermining	0.69	0.87
	Joint One-seq	0.32	0.74
	PTCG (Ours)	**0.82**	**0.88**
Qwen3-8B	Vanilla	0.83	0.86
	Argument Undermining	0.84	0.90
	Joint One-seq	0.49	0.79
	PTCG (Ours)	**0.88**	**0.91**
Qwen3-14B	Vanilla	0.85	0.88
	Argument Undermining	0.81	0.88
	Joint One-seq	0.55	0.81
	PTCG (Ours)	**0.88**	**0.90**

Table 2: Targeted persuasiveness is evaluated using the delta classifier. The score ranges from –1 to 1 and represents the predicted probability of receiving a delta.

Notably, PTCG achieves the best Avg and Max scores on most backbones, including LLaMA3.2-1B, Qwen3-8B, and Qwen3-14B, indicating its ability to generate highly persuasive counterarguments. For LLaMA3.1-8B, although the Avg score is slightly lower than the Vanilla baseline, PTCG attains the highest Max score (0.90), suggesting that it can produce particularly strong persuasive arguments among its generated candidates.

Overall, these results confirm that PTCG effectively improves targeted persuasiveness across diverse backbone models.

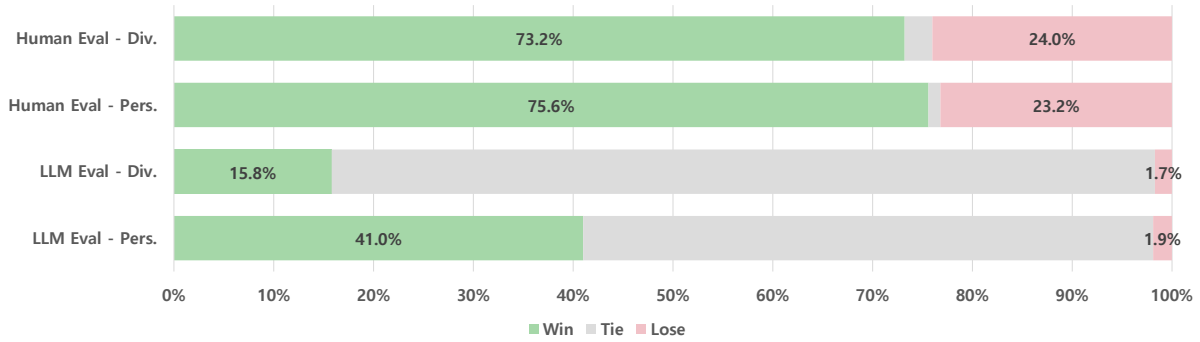

Figure 2: Win/Tie/Lose analysis of diversity (Div.) and persuasiveness (Pers.) across Human Evaluation (top) and LLM-based Evaluation (bottom). The figure compares our method (PTCG) with baselines using LLaMA3.1-8B as the main backbone model, showing the proportion of cases where each system's counterarguments were judged as more diverse or persuasive (Win), equivalent (Tie), or less effective (Lose).

5.3 Pairwise Evaluation

To assess our framework's effectiveness, we conduct human and LLM-based pairwise evaluations, focusing on the persuasiveness and diversity of generated counterarguments.

For the human evaluation, we randomly sampled 95 inputs[4] and asked five human evaluators to rate the outputs from PTCG and the baselines. Details are in Appendix Section A7. As shown in Figure 2 (top), human judges consistently rated PTCG higher than the baseline across both diversity and persuasiveness.[5] Since both systems rely on the same LLaMA3.1-8B model, this comparison isolates the effect of the framework itself.

In terms of diversity, PTCG generated counterarguments that reflected a much wider variety of perspectives compared to the baseline. The high win rate in this metric demonstrates that our framework effectively avoids repetitive reasoning by exploring multifaceted logical paths. Meanwhile, the results for persuasiveness indicate that evaluators found the arguments from PTCG to be significantly more compelling. This suggests that the generated content is not only diverse but also structurally sound and convincing. Taken together, these distinct gains show that PTCG is not only capable of exploring a diverse solution space but also excels at generating highly persuasive counterarguments that resonate with human evaluators.

We also conducted pairwise comparisons using GPT-4o-mini (OpenAI, 2024a) as the evaluator, assessing persuasiveness and diversity. To mitigate potential ordering effects (Zheng et al., 2023), we presented the counterargument sets in both orders[6]. We also allowed for *Tie*, which captures cases where the evaluator either explicitly selected a "Hard" option or produced inconsistent preferences when the order of the two sets was reversed, reflecting potential ordering effects inherent in LLM-based pairwise judgments. Figure 2 (bottom) presents the pairwise comparison results between PTCG and its Vanilla version using LLaMA3.1-8B as the backbone model. LLM-based pairwise evaluations tend to produce relatively high tie rates due to potential ordering bias. Nevertheless, the results show that PTCG achieves a higher proportion of wins in both diversity and persuasiveness, further supporting the improvements observed in the human evaluation. Additional LLM-based evaluation results across other backbone models are provided in Appendix Section A8.

Overall, these results demonstrate that PTCG consistently improves both perspective diversity and persuasiveness compared to the vanilla baseline across all evaluations.

5.4 Qualitative Analysis

Analysis of the generated counterarguments[7] shows that PTCG moves beyond the abstract ethical reasoning commonly produced by baseline models by incorporating richer social and practical contexts.

In the organ donation debate, baseline outputs largely repeat general arguments about autonomy and utilitarianism. In contrast, PTCG grounds its

[4]This evaluation scale is larger than those in prior studies (Hu et al., 2025; Alshomary and Wachsmuth, 2023).

[5]Inter-annotator agreement, assessed via Fleiss' Kappa (0.619) and Krippendorff's Alpha (0.612), indicates substantial consistency (Landis and Koch, 1977) among annotators.

[6]Each evaluation pair was tested in both AB and BA sequences, and the final results were aggregated accordingly.

[7]Please refer to Table A8 and Table A9 for examples.

Configuration	*Persuasiveness*		*Perspective Diversity*
	General	Targeted	
LLaMA3.1-8B	$7.79_{\pm 0.0031}$	$7.59_{\pm 0.0023}$	$3.78_{\pm 0.0021}$
+ Persona	$7.79_{\pm 0.0014}$	$7.16_{\pm 0.0050}$	$4.24_{\pm 0.0040}$
+ CoT-based Gen. (n=1; linear chain)	$7.98_{\pm 0.0015}$	$7.75_{\pm 0.0018}$	$1.99_{\pm 0.0097}$
+ Tree-based Gen. (n=3; tree search)	$8.01_{\pm 0.0019}$	$7.77_{\pm 0.0024}$	$2.25_{\pm 0.0075}$
+ Persona + Tree-based Gen. (PTCG)	$\mathbf{8.32}_{\pm 0.0025}$	$\mathbf{7.93}_{\pm 0.0032}$	$\mathbf{4.28}_{\pm 0.0024}$

Table 3: Ablation study of PTCG. All scores are averaged over five runs and reported with standard deviations.

counterarguments in more concrete considerations, such as societal trust in healthcare systems and the broader implications of state control over bodily decisions. This framing makes the arguments more relatable and easier for audiences to evaluate in real-world terms.

Similarly, in the scientific theory patenting example, the baseline focuses on a single general claim about innovation, whereas PTCG introduces multi-faceted perspectives through diverse personas. This shift from abstract reasoning to human-centered context enables PTCG to produce more nuanced and persuasive responses.

Consequently, PTCG achieves stronger performance by grounding its arguments in concrete, human-centered reasoning that baseline models fail to capture. Our framework naturally produces longer and more detailed outputs due to its multi-stage, persona-grounded reasoning process. These excerpts demonstrate how persona conditioning influences the depth, structure, and domain-specific reasoning styles of each counterargument, ultimately contributing to richer and more nuanced responses compared to vanilla LLMs.

5.5 Ablation Study

We analyze the contribution of each component through ablation experiments, which evaluate each module individually and in combination. Starting from the baseline LLaMA3.1-8B model, we incrementally introduce persona grounding, CoT-based, and tree-based generation.

Persona grounding alone substantially boosts perspective diversity by introducing various social backgrounds and viewpoints. However, it fails to improve persuasiveness and even leads to a decrease in targeted persuasiveness, suggesting that diverse viewpoints alone do not inherently guarantee a compelling argument. In contrast, compared to the LLaMA 3.1 baseline, tree-based generation contributes most significantly to persuasiveness by exploring multiple reasoning paths. However, this structured search process leads to a sharp decline in perspective diversity. This indicates that while the search process optimizes for argumentative strength, it tends to converge on a narrow set of logical paths. Ultimately, PTCG, which integrates both persona grounding and tree-based generation, achieves the highest scores across all metrics. It successfully recovers the diversity lost during the tree search, while simultaneously attaining the highest levels of both general and targeted persuasiveness. This demonstrates that PTCG effectively translates diverse persona-driven signals into robust persuasive arguments through structured reasoning.

6 Conclusion

In this work, we proposed and addressed the task of generating multiple distinct and persuasive counterarguments grounded in realistic personas. Inspired by the Tree-of-Thoughts approach, we proposed a tree-based step-wise generation with a pruning process to improve reasoning quality, while integrating personas grounding through distance-based persona selection. This design enables our framework to explore diverse argumentative perspectives while maintaining a strong persuasive structure, overcoming the inherent limitations of standard LLM generation. For evaluation, we conducted comprehensive experiments combining human evaluation, LLM-based judgments, and classifier-based assessments. The results demonstrate that our persona-grounded and tree-based generation approach significantly improves both the diversity and persuasiveness of counterarguments. Our work highlights the importance of structured reasoning and persona-driven perspective-taking for counterargument generation, suggesting a promising direction for building LLM systems that support critical thinking, balanced debates, and informed decision-making.

Ethical Considerations

This study makes use of the publicly available Reddit ChangeMyView dataset, which has also been widely adopted in prior work (Alshomary and Wachsmuth, 2023; Lin et al., 2023) on computational argumentation. The dataset was used strictly for research purposes. We are aware that generating persuasive counterarguments with large language models raises ethical concerns, particularly the risk of misuse in manipulative or coercive contexts. To reduce these risks, our work is limited to academic exploration, with the goal of examining diversity of perspectives rather than promoting adversarial persuasion.

In addition, to prevent the risk of outputs being mistaken for human speech in deceptive ways, we restricted the use of first-person generation. To protect the privacy of the original authors, we anonymized all user-identifiable information by replacing original usernames with generic identifiers (e.g., User 1, User 2). Furthermore, we paraphrased the examples cited in this paper while preserving their semantic intent. These modifications were applied to the illustrative examples, whereas the actual computational analysis was conducted on the original anonymized data. Thus, our study focused on the substantive content and patterns of the arguments rather than the identities of individual authors.

Additionally, our reliance on English-based Reddit data may introduce demographic biases and limit the generalizability of our findings to other languages or cultural rhetorical traditions. We also emphasize the need for future work to consider safeguards and responsible use practices when deploying such systems.

Limitations

While our approach shows significant improvements, we identify several limitations that provide opportunities for future work. First, this work focuses on generating counterarguments for a single opinion. However, debates often unfold over multiple rounds, with each round consisting of an exchange between participants (Durmus and Cardie, 2019; Li et al., 2020). Extending this framework to multi-round debates would enable models to engage in interactive and dynamic exchanges, where counterarguments are refined over multiple turns. This line of research could further extend to multi-party debates, where multiple participants interact and compete (Sia et al., 2022). Second, the current framework is limited to text-based personas derived from clustering. Incorporating richer user signals, such as value-based attributes, may further improve the persuasiveness of generated counterarguments (Lukin et al., 2017).

References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967, Dubrovnik, Croatia. Association for Computational Linguistics.

C Daniel Batson, Shannon Early, and Giovanni Salvarani. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin*, 23(7):751–758.

Maike Behrendt, Stefan Sylvius Wagner, Carina Weinmann, Marike Bormann, Mira Warne, and Stefan Harmeling. 2025. Natural language processing to enhance deliberation in political online discussions: A survey. *arXiv preprint arXiv:2506.02533*.

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

Olivia M Bullock, Hillary C Shulman, and Richard Huskey. 2021. Narratives are persuasive because they are easier to understand: examining processing fluency as a mechanism of narrative persuasion. *Frontiers in Communication*, 6:719615.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large

language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.

Teun J Dekker. 2020. Teaching critical thinking through engagement with multiplicity. *Thinking Skills and Creativity*, 37:100701.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Robert H Ennis. 2015. Critical thinking: A streamlined conception. In *The Palgrave handbook of critical thinking in higher education*, pages 31–47. Springer.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Morgan Gray, Li Zhang, and Kevin D Ashley. 2025. Generating case-based legal arguments with llms. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 160–168.

Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. AMERICANO: Argument generation with discourse-driven decomposition and agent interaction. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 82–102, Tokyo, Japan. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yi-Long Lu, Jiajun Song, Chunhui Zhang, and Wei Wang. 2025. Mind the gap: The divergence between human and llm-generated tasks. *arXiv preprint arXiv:2508.00282*.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Raymond A Mar and Keith Oatley. 2008. The function of fiction is the abstraction and simulation of social experience. *Perspectives on psychological science*, 3(3):173–192.

Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

Leland McInnes, John Healy, Steve Astels, and 1 others. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Wiktoria Mieleszczenko-Kowszewicz, Dawid Płudowski, Filip Kołodziejczyk, Jakub Świstak, Julian Sienkiewicz, and Przemysław Biecek. 2024. The dark patterns of personalized persuasion in large language models: Exposing persuasive linguistic features for big five personality traits in llms responses. *arXiv preprint arXiv:2411.06008*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: September 23, 2025.

OpenAI. 2024b. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: September 23, 2025.

Moritz Plenz, Philipp Heinisch, Janosch Gehring, Philipp Cimiano, and Anette Frank. 2025. From argumentation to deliberation: Perspectivized stance vectors for fine-grained (dis)agreement analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1525–1553, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Suzanna Sia, Kokil Jaidka, Hansin Ahuja, Niyati Chhaya, and Kevin Duh. 2022. Offer a different perspective: Modeling the belief alignment of arguments in multi-party debates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11939–11950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023. Argument and counter-argument generation: A critical survey. In *International conference on applications of natural language to information systems*, pages 500–510. Springer.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. 2025. Psyplay: Personality-infused role-playing conversational agents. *arXiv preprint arXiv:2502.03821*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Li Zhang and Kevin D Ashley. 2025. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation. *arXiv preprint arXiv:2506.02992*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213,

Melbourne, Australia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A1 LLM Usage

We employed GPT-4o to support the manuscript preparation process. Specifically, the model was used to assist in identifying relevant prior work and refining the clarity and readability of the draft. This usage was limited to auxiliary scholarly support—such as improving grammar and reducing stylistic inconsistencies.

A2 Persona Cluster Examples

To illustrate how the persona clusters are organized, we present representative examples from three clusters: (1) Sports and Physical Education, (2) Finance and Marketing, and (3) Historians. These examples demonstrate the semantic coherence within each cluster, reflecting domain-specific interests and professional backgrounds.

Sports and Physical Education

- A high school physical education teacher seeking to incorporate Paralympic history and achievements into the curriculum to inspire and educate students about inclusivity in sports.
- A sports scientist researching the biomechanics and physics of tennis, focusing on how racket specifications impact performance and injury risks.
- A sports journalist covering the history of ice hockey and its impact on national identity in Poland.
- An elementary school teacher who enjoys incorporating diverse sports stories in her curriculum to inspire students.
- A football coach seeking to learn from successful strategies and team management in various leagues.

Finance and Marketing

- A financial analyst specializing in Asian markets and wealthy individuals, interested in tracking the investments and philanthropic activities of billionaires like Gerald Chan.
- A quantitative analyst with expertise in financial modeling and algorithmic trading, seeking to develop and implement systematic value investment strategies.
- A digital marketing specialist interested in innovative aggregator models that consolidate search results from multiple sources.
- A marketing specialist for a tech company, looking for innovative ways to engage with pop culture and fandoms to promote new products and services.
- A business strategist for Arriva UK Bus, interested in exploring opportunities and challenges related to subsidiary operations and company restructuring.

Historians

- An Iowa historian focusing on the development and growth of townships in Jones County.
- A historian specializing in 19th-century British architecture, with a focus on the works of notable architects in Lancashire.
- A local historian specializing in the political and business development of Marlborough, Massachusetts in the 19th century.
- A historian specializing in the late medieval and early modern history of France and the Iberian Peninsula, with a focus on power dynamics, family strategies, and women's roles in politics.
- A local historian or genealogist researching the history of small communities and families in Fremont County, Iowa.

A3 Implementation Details

Our models were developed and evaluated using multiple GPU resources. Specifically, the argument undermining and joint one-seq methods were trained on an NVIDIA A100 GPU (80GB). For inference and testing, the vanilla baseline and our PTCG were deployed on two NVIDIA RTX A6000 GPUs (48GB each) using vLLM (v0.8.5) (Kwon et al., 2023) to ensure efficient processing. During the generation process, we set the temperature to 0.8 and top-p to 0.95 to balance diversity and coherence, with the maximum token length limited to 2,048. Additionally, API-based models, including gpt-4o-mini and text-embedding-3-large, were accessed via OpenAI's API.

A4 LLM-based Evaluation: Quality

Prior studies have established diverse criteria for evaluating the quality of arguments. Alshomary et al. (2021) consider grammaticality and content richness as key factors in assessing generated arguments, while Lin et al. (2023) emphasize appropriateness, grammaticality, and logic, aligning GPT-based and human evaluation through shared criteria. Moreover, Wachsmuth et al. (2017) provide a comprehensive taxonomy of argument quality, highlighting clarity, appropriateness, and Relevance as core components under effectiveness

Backbone	Method	App.	Cla.	Gra.	Rel.
LLaMA3.1-8B	Vanilla	4.40	4.27	4.74	4.70
	AU	2.46	2.36	2.37	2.80
	Joint One-seq	3.02	3.06	3.22	3.22
	PTCG (Ours)	**4.55**	**4.34**	**4.87**	**4.75**
LLaMA3.2-1B	Vanilla	**3.95**	**3.90**	**4.32**	4.18
	AU	2.55	2.46	2.61	2.59
	Joint One-seq	2.33	2.31	2.48	2.30
	PTCG (Ours)	3.90	3.83	4.29	**4.19**
Qwen3-8B	Vanilla	**4.71**	**4.55**	**4.94**	**4.90**
	AU	3.40	3.30	3.45	3.53
	Joint One-seq	2.42	2.40	2.50	2.59
	PTCG (Ours)	4.28	4.27	4.84	4.63
Qwen3-14B	Vanilla	**4.79**	**4.62**	**4.96**	**4.94**
	AU	3.35	3.25	3.43	3.55
	Joint One-seq	2.61	2.60	2.68	2.81
	PTCG (Ours)	4.70	4.60	**4.96**	4.93

Table A1: Evaluation results of baselines and PTCG measured by GPT-4o-mini. Reported scores have been averaged over five runs. For *quality*, we report appropriateness (App.), clarity (Cla.), grammaticality (Gra.), and relevance (Rel.), each on a 1–5 scale.

Configuration	App.	Cla.	Gra.	Rel.
LLaMA3.1-8B	4.40	4.27	4.74	4.70
+ Persona	3.82	3.81	4.38	3.99
+ CoT-based Gen. (n=1; linear chain)	4.69	4.43	**4.93**	4.82
+ Tree-based Gen. (n=3; tree search)	**4.70**	**4.44**	**4.93**	**4.83**
+ Persona + Tree-based Gen. (PTCG)	4.55	4.34	4.87	4.75

Table A2: Ablation study of PTCG. For *quality*, we report appropriateness (App.), clarity (Cla.), grammaticality (Gra.), and relevance (Rel.), each on a 1–5 scale.

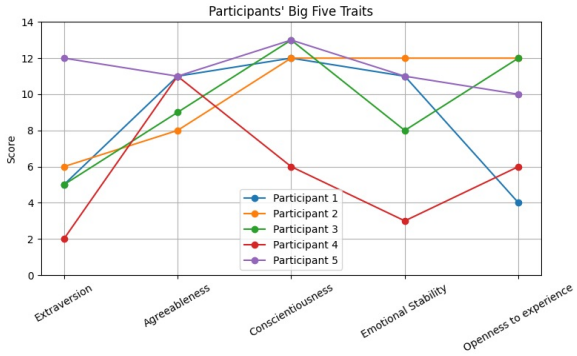

Figure A1: Big Five Personality trait distribution of recruited evaluators. The heterogeneous profiles helped ensure diverse perspectives in human evaluation.

and reasonableness. Based on these findings, we adopt four criteria for LLM-based evaluation of counterarguments:

Appropriateness: Whether the language and tone are suitable for the context and proportional to the issue. Inappropriate tone (e.g., overly aggressive or dismissive) lowers the score.

Clarity: Whether the writing is clear, well-organized, and free from ambiguity or unnecessary complexity, allowing the reader to easily follow the reasoning.

Grammaticality: Whether the text follows standard grammar conventions, including punctuation, sentence structure, and syntax. This ensures the counterargument reads fluently without errors.

Relevance: How directly the counterargument engages with the original post and addresses its key points. Superficial or off-topic content would reduce relevance.

A5 Quality Scores Across Backbone Models

As presented in Table A1, PTCG demonstrates robust performance in terms of generation quality across various backbone models. Overall, PTCG consistently maintains high scores across all four dimensions. While minor fluctuations in scores are observed compared to the Vanilla baseline, these variations remain within a highly acceptable range. Such marginal differences are interpreted as a natural trade-off arising from the model's effort to adhere to specific persona constraints while maintaining linguistic universality.

A6 Ablation Study on Quality Scores

The ablation study in Table A2 provides a deeper insight into how PTCG balances persona integration with output quality. Our analysis reveals that simply injecting persona conditioning leads to a noticeable decline in quality scores, as the model prioritizes persona-specific traits over general clarity. However, the integration of the tree-based stepwise generation and selection effectively mitigates this issue, successfully rebounding the scores back to levels comparable to the original LLaMA 3.1 baseline. These results empirically demonstrate that the tree-based search mechanism serves as a critical quality-control layer, ensuring that the diversity gained through persona conditioning does not come at the expense of linguistic or logical integrity. Consequently, PTCG achieves an optimal equilibrium between diverse perspectives and generation quality.

A7 Human Evaluation

Recruitment We recruited five evaluators comprising both undergraduate and graduate students. They were compensated at a rate of 15,000 KRW

Algorithm 1 Persona-guided Tree-based Counterargument Generation (PTCG)

Require: Original post x, persona cluster centroids $\mathcal{C}$
Ensure: Final counterarguments $\mathcal{Y} = \{y_1, y_2, y_3\}$

Step 1: Original Poster (OP) Persona Estimation
$p_{\text{op}} \leftarrow \texttt{LLM_Estimate_Persona}(x)$ ▷ Prompt: Figure A4
$c_{\text{op}} \leftarrow \texttt{Nearest_Centroid}(p_{\text{op}}, \mathcal{C})$

Step 2: Speaker Persona Selection
Select three speaker personas from $\mathcal{C}$:
  p_{same} = centroid persona from c_{op} ▷ Same cluster
  p_{nearest} = centroid persona from nearest cluster
  p_{furthest} = centroid persona from furthest cluster
$\mathcal{P}^\star = \{p_{\text{same}}, p_{\text{nearest}}, p_{\text{furthest}}\}$

Step 3: Plan Generation and Selection
for each $p \in \mathcal{P}^\star$ **do**
  $\{r_1, r_2, r_3\} \leftarrow \texttt{LLM_Generate_Plans}(x, p_{\text{op}}, p)$ ▷ Prompt: Figure A5
  $r_p^\star \leftarrow \texttt{LLM_Select_Plan}(\{r_1, r_2, r_3\}, p, p_{\text{op}})$ ▷ Prompt: Figure A6
end for

Step 4: Counterargument Generation and Selection
for each $p \in \mathcal{P}^\star$ **do**
  $\{y_1^p, y_2^p, y_3^p\} \leftarrow \texttt{LLM_Generate_Counters}(x, p_{\text{op}}, p, r_p^\star)$ ▷ Prompt: Figure A5
  $y_p^\star \leftarrow \texttt{LLM_Select_Counter}(\{y_1^p, y_2^p, y_3^p\}, p, p_{\text{op}})$ ▷ Prompt: Figure A6
  Add $y_p^\star$ to $\mathcal{Y}$
end for

return $\mathcal{Y}$

per hour, which exceeds the local minimum wage. In selecting participants, we referred to prior work showing that argument persuasiveness can be influenced by individual personality traits (Lukin et al., 2017), particularly as measured by the Big Five Personality framework. To ensure diversity in personalities, participants completed the Ten Item Personality Inventory (TIPI) survey (Gosling et al., 2003), and we considered the distribution of their scores during recruitment. This ensured a heterogeneous pool of evaluators, encompassing a wide range of personality traits, as illustrated in Figure A1.

Procedure Participants were formally warned beforehand that the Reddit-sourced data might contain offensive or sensitive content. After being informed of the study's purpose and data usage, all participants provided explicit consent prior to the task. Each participant reviewed several original posts and corresponding counterarguments generated by different methods. Inspired in part by the evaluation setup of Chung et al. (2025), where annotators assessed outputs on quality and diversity in set-based presentations, we adopted a similar approach. Specifically, from the full test set we sampled 95 instances, and participants were asked to perform two tasks: In the *persuasiveness task*, participants selected the most persuasive counterargument for each original post. In the *diversity task*, participants compared two sets of counterarguments and judged which set demonstrated greater perspective diversity.

To mitigate ordering effects, the order of the counterargument sets was randomized, and each session was conducted individually. For example, if one evaluation compared the systems in the order A then B (AB), another was conducted in the reverse order (BA) to balance potential bias.

A8 Additional LLM-based Evaluation Results

To further examine the robustness of our framework across different model architectures, we conduct additional LLM-based pairwise evaluations using GPT-4o-mini as the judge. In addition to the main

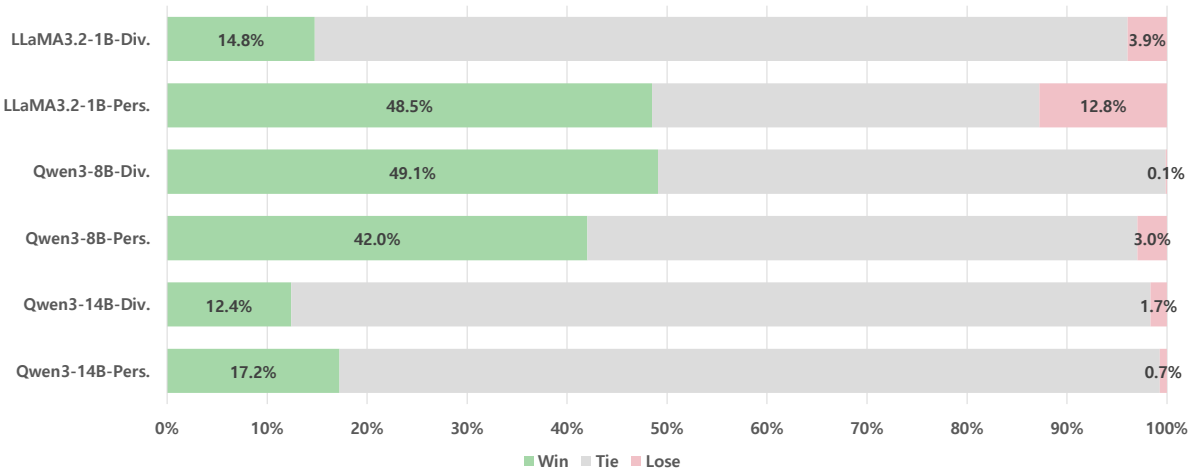

Figure A2: Win/Tie/Lose analysis of diversity (Div.) and persuasiveness (Pers.) in LLM-based evaluation using GPT-4o-mini. The figure compares our method (PTCG) with their Vanilla counterparts across different backbone models (LLaMA3.2, Qwen3-8B, and Qwen3-14B), showing the proportion of cases where each system's counterarguments were judged as more diverse or persuasive (Win), equivalent (Tie), or less effective (Lose).

backbone model reported in the main paper, we evaluate our method against the Vanilla baseline using LLaMA3.2-1B, Qwen3-8B, and Qwen3-14B. Figure A2 reports the Win/Tie/Lose ratios for persuasiveness and diversity. Overall, our method rarely underperforms the Vanilla baseline across all backbone models. In particular, the lose rates remain very low (below 5%), indicating that our framework consistently maintains comparable or better performance. The improvements are most notable for Qwen3-8B, where our method achieves win rates of 42.0% in persuasiveness and 49.1% in diversity. For larger models such as Qwen3-14B, most comparisons result in ties, suggesting that our framework maintains competitive performance while still providing improvements in a subset of cases. These results further confirm that the benefits of persona-grounded, tree-based generation generalize across different backbone models.

A9 Scalability and Efficiency Analysis

To evaluate the computational implications of our multi-stage reasoning framework, we conducted a detailed latency and throughput analysis across different generation settings. As expected, PTCG introduces additional overhead compared to single-pass LLaMA3.1-8B due to its structured planning, persona integration, and tree-based multi-branch generation steps.

Our results show that the end-to-end runtime increases with the number of branches n, reflecting the inherent cost of generating multiple candidate plans and counterarguments. For instance, PTCG with n=3 requires nearly twice the total time of the single-pass baseline (Tables A3, A4). This is consistent with the added computation introduced by the planning stage and the evaluation of multiple branches per input. However, an interesting and non-trivial observation emerges when examining per-response latency. Although PTCG increases total computation time, the average latency per generated counterargument is lower for PTCG (n=3) than for vanilla LLaMA (n=1). This occurs because PTCG leverages batched multi-branch generation, effectively improving batch utilization during decoding. By generating multiple reasoning paths within a single forward pass, the model amortizes computational cost and produces more responses without a proportional increase in latency.

Ablation experiments further clarify which components contribute most to computational overhead. The results in Tables A5 and A6 show that adding only the persona module increases latency modestly (about 12%), while tree-based generation contributes the largest increase (about 111%). When combined, the full framework reaches about 239% of the baseline cost. This decomposition confirms that tree-structured exploration and evaluation—rather than persona integration—drive the majority of the additional computation.

Despite the additional computation introduced by multi-step reasoning, the increase in runtime is a natural consequence of performing more detailed reasoning during generation. Compared

Model	Avg end-to-end latency/OP (s)	Per response (s)	Percentage (%)
LLaMA3.1-8B(n=1)	$3.16_{\pm 0.033}$	$3.16_{\pm 0.033}$	100.00
LLaMA3.1-8B (n=3)	$3.27_{\pm 0.014}$	$1.09_{\pm 0.011}$	34.49
LLaMA3.1-8B + PTCG (n=1)	$4.94_{\pm 0.016}$	$4.94_{\pm 0.016}$	156.33
LLaMA3.1-8B + PTCG (n=3)	$7.84_{\pm 0.022}$	$2.61_{\pm 0.007}$	82.59
LLaMA3.1-8B + PTCG (n=5)	$10.59_{\pm 0.073}$	$3.53_{\pm 0.024}$	111.71

Table A3: Latency results (Part 1): End-to-end latency, per-response latency, and relative percentage.

Model	Throughput (tokens/s)	Total Time (s)
LLaMA3.1-8B (n=1)	$266.33_{\pm 1.874}$	$2680.95_{\pm 21.812}$
LLaMA3.1-8B (n=3)	$267.63_{\pm 2.130}$	$2772.15_{\pm 14.660}$
LLaMA3.1-8B + PTCG (n=1)	$135.37_{\pm 0.510}$	$4240.53_{\pm 14.683}$
LLaMA3.1-8B + PTCG (n=3)	$257.23_{\pm 0.962}$	$6697.21_{\pm 19.632}$
LLaMA3.1-8B + PTCG (n=5)	$317.30_{\pm 2.288}$	$9029.98_{\pm 62.639}$

Table A4: Latency results (Part 2): Throughput and total time.

Ablation	Avg end-to-end latency/OP (s)	Per response (s)	Percentage (%)
LLaMA3.1-8B	$3.27_{\pm 0.014}$	$1.09_{\pm 0.011}$	100
+ Persona	$3.67_{\pm 0.046}$	$1.22_{\pm 0.015}$	111.93
+ Tree-based Gen.	$6.90_{\pm 0.024}$	$2.30_{\pm 0.008}$	211.01
+ Persona + Tree-based Gen.	$7.84_{\pm 0.022}$	$2.61_{\pm 0.007}$	239.45

Table A5: Ablation results (Part 1): End-to-end latency, per-response latency, and relative percentage.

Ablation	Throughput (tokens/s)	Total Time (s)
LLaMA3.1-8B	$267.63_{\pm 2.130}$	$2772.15_{\pm 14.660}$
+ Persona	$262.47_{\pm 2.038}$	$3113.87_{\pm 38.347}$
+ Tree-based Gen.	$237.40_{\pm 0.854}$	$5899.39_{\pm 21.210}$
+ Persona + Tree-based Gen.	$257.23_{\pm 0.962}$	$6697.21_{\pm 19.632}$

Table A6: Ablation results (Part 2): Throughput and total time.

to single-pass LLaMA, PTCG requires additional steps for structured planning and tree-based exploration, which inevitably increases the total computation time. However, the tree-based generation process allows multiple candidate counterarguments to be produced in parallel. As a result, while the total runtime increases, the cost of producing multiple counterarguments does not grow proportionally. More importantly, the additional computation leads to substantial improvements in counterargument quality. The structured reasoning process enables the model to explore diverse candidate perspectives and refine them through evaluation, which directly improves both persuasiveness and diversity. From this perspective, the computational overhead represents a deliberate trade-off: the framework spends additional compute to obtain significantly stronger argumentative outputs.

Overall, although PTCG incurs a higher computational cost due to its structured reasoning process, the improved generation efficiency for multiple responses and the substantial gains in output quality demonstrate that the additional computation is justified for counterargument generation tasks.

Model	Precision	Recall	F1-score	Accuracy
RoBERTa	0.52	**0.94**	0.67	0.54
DeBERTa	0.62	0.67	0.65	0.64
Longformer	**0.63**	0.78	**0.70**	**0.66**

Table A7: Performance comparison of Delta classifiers built on different backbone models (RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and Longformer (Beltagy et al., 2020)). Scores are reported in terms of Precision, Recall, F1-score, and Accuracy. Longformer outperforms other baselines; therefore, we use the Longformer-based classifier to evaluate targeted persuasiveness.

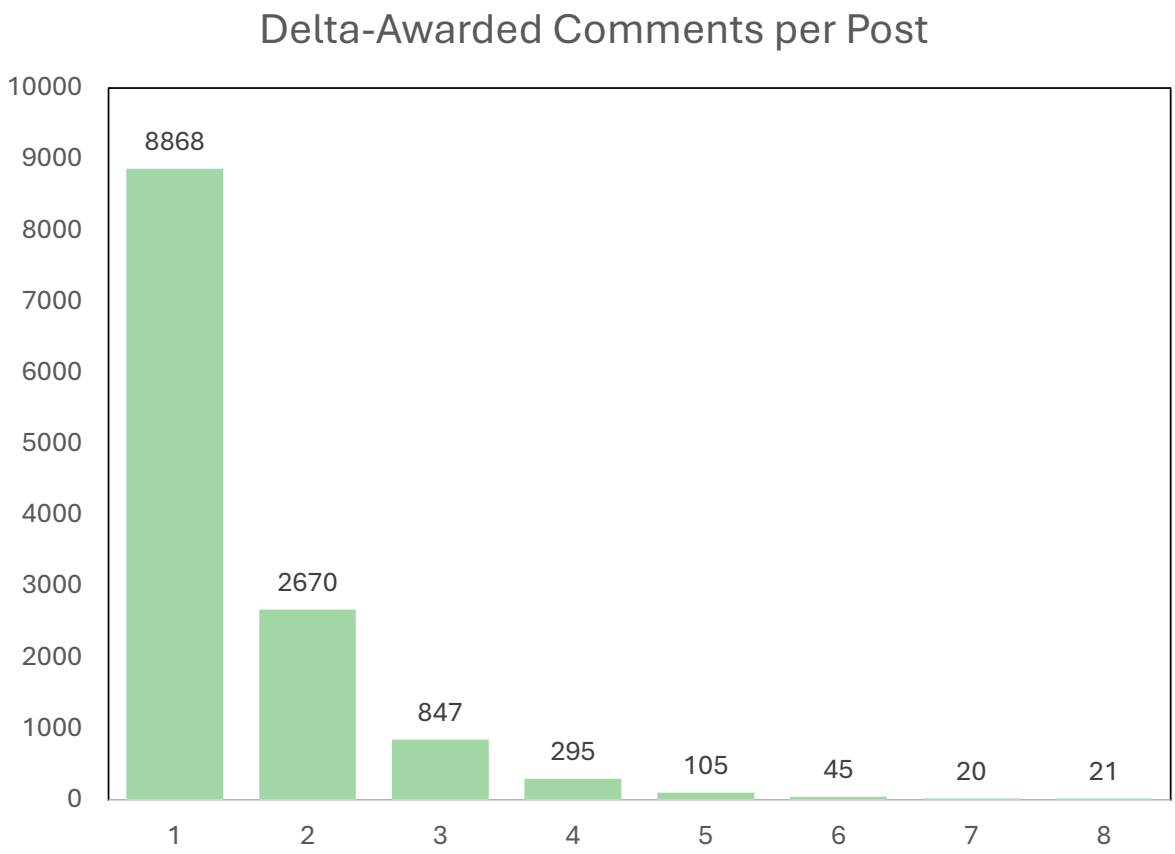

Figure A3: Distribution of delta-awarded comments per post. Most posts receive only one delta-awarded comment, while cases with more than three are rare. Since one or two comments provide limited diversity and instances of four or more are scarce, three were chosen as a practical setting to balance diversity with interpretability.

Original Poster (OP) Persona Estimation:
Given the following argument, infer the likely persona of the author.
Respond in exactly one sentence, focusing on the author's likely identity based on the argument alone.
Do not invent any information that is not grounded in the argument.

Argument:
{input}

Format your output as follows:

Author's Persona:
[Your response]

Figure A4: Prompt used for estimating the Original Poster (OP) persona from the given post. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Title (claim) **Organ donation should be mandatory.**
LLaMA3.1-8B-Instruct (Vanilla) **Counterargument 1:** Mandatory organ donation could be considered **an infringement on personal autonomy and bodily integrity**. The argument that someone would want their organs donated if they were in a similar situation is based on a flawed assumption that people's desires in hypothetical situations can be accurately predicted. ... **Counterargument 2:** The argument for **mandatory organ donation relies heavily on a utilitarian perspective**, which prioritizes the greatest good for the greatest number of people. However, this approach can overlook the individual experiences and needs of those who may be affected by mandatory organ donation. For example, certain communities or individuals may have strong cultural or personal objections to organ donation that cannot be ignored. ... **Counterargument 3:** The assumption that all people's organs should be made available for donation after death is based on **a flawed understanding of the complexities surrounding organ donation and transplantation.** In many cases, organs may be unsuitable for donation due to various medical conditions, such as cancer or HIV. ...
LLaMA3.1-8B-Instruct (PTCG) **Counterargument 1: Persona** A high school counselor specializing in alternative education, interested in exploring different educational models to better serve at-risk students. (Same cluster) While the idea of mandatory organ donation might seem appealing from a **utilitarian perspective, it is essential to consider the potential consequences of such a policy on the autonomy and dignity of individuals.** The notion that people's bodies should be treated as mere resources to be exploited for the greater good raises significant concerns about the commodification of human life. ... **Counterargument 2: Persona** A high school geography teacher who enjoys incorporating historical events into lessons to provide context for modern political and social issues. (Nearest cluster) By imposing this requirement, the state would be taking a profound step into the private lives of its citizens, potentially infringing upon deeply held beliefs and values. The argument that utilitarianism justifies mandatory donation overlooks the fact that the concept of "good" can be subjective and context-dependent. **What may be seen as a "good" outcome in one instance could be viewed as morally reprehensible in another.** Furthermore, the example of the liver transplant cited in support of mandatory donation oversimplifies the complexities of human relationships and the nuances of individual decision-making. ... **Counterargument 3: Persona** An entomologist specializing in Lepidoptera (moths and butterflies) with expertise in moth taxonomy and biogeography, particularly interested in Neotropical species. (Furthest cluster) ... the argument that a person would want to donate their organs if they were in need of a transplant ignores the fact that many individuals who would be affected by mandatory organ donation may not be aware of their preferences or may not have the capacity to make informed decisions. This could lead to a situation where individuals are forced to donate their organs against their will, which would be **a significant erosion of trust between individuals and the healthcare system.** ...

Table A8: Comparative Examples of Counterarguments on Mandatory Organ Donation Across Baseline Models and PTCG

Title (claim) **Scientists should be allowed to patent their theories.**
LLaMA3.1-8B-Instruct (Vanilla) **Counterargument 1:** Allowing scientists to patent their theories could stifle the advancement of knowledge by creating a barrier to entry for future researchers. Patenting theories would essentially turn scientific discovery into a competitive marketplace, where the primary goal is to secure intellectual property rights rather than to advance understanding. **This could lead to a situation where groundbreaking research is delayed or even abandoned due to concerns about potential patent infringement.** ... **Counterargument 2:** ... patenting theories would also create a new class of intellectual property owners who would have a vested interest in suppressing alternative explanations or discoveries that might challenge their patents. **This would lead to a situation where scientific progress is hindered by the need to navigate complex patent portfolios and negotiate licensing agreements.** ... **Counterargument 3: Patenting scientific theories would create a situation where individual scientists are rewarded for their contributions, while the broader scientific community is left to pick up the pieces.** Moreover, the fact that scientists are often motivated by a desire to make a discovery or advance our understanding of the world suggests that financial rewards are not the primary driver of scientific progress. In fact, many of the most groundbreaking discoveries in history were made by scientists who were driven by curiosity and a passion for knowledge, rather than a desire for financial gain. ...
LLaMA3.1-8B-Instruct (PTCG) **Counterargument 1: Persona** A science journalist with a focus on astronomy and space exploration, interested in reporting on observatories and their contributions to scientific knowledge. (Same cluster) ... patenting scientific theories would create an uneven playing field, where those with the resources to patent and litigate their ideas would have a significant advantage over those who do not. **This would lead to a lack of innovation and collaboration, as researchers would be more focused on protecting their own interests than on advancing the field as a whole.** ... **Counterargument 2: Persona** A medical researcher specializing in oncology, with a focus on exploring novel therapeutic approaches and evaluating clinical trial data for various cancer indications. (Nearest cluster) ... patenting theories could lead to the suppression of research findings that contradict established theories, **as scientists may be reluctant to publish their results for fear of losing their intellectual property.** The case of naturally occurring DNA sequences being patented is often cited as an example of the need for patenting scientific discoveries, but this is a misinterpretation. ... **Counterargument 3: Persona** A casting director for TV and film productions, looking for talented actors to fit specific roles in various projects. (Furthest cluster) ... patenting scientific theories would also lead to a surge in **frivolous lawsuits**, as scientists and researchers would be more likely to sue each other for allegedly infringing on their patents. This would not only be costly but also time-consuming, diverting resources away from actual scientific research. ... the patent office would need **to establish a new set of criteria for evaluating the patentability of scientific theories, which would be a daunting task.** ...

Table A9: Comparative Examples of Counterarguments on Restricting Children's Access to Movie Theaters at Night Across Baseline Models and PTCG

Min Cluster Size	# of Clusters	Silhouette Score	Calinski–Harabasz Index
25	280	0.6255	16008.35
50	147	<u>0.6416</u>	22752.07
100	58	0.6251	17977.72
200	39	**0.6513**	<u>24054.99</u>
300	27	0.6081	**26879.54**

Table A10: Clustering results under different minimum cluster size settings (dimensionality fixed to 50). Evaluation metrics include the Silhouette Score and Calinski–Harabasz Index.

Tree-based Step-wise Generation Prompt:
Write a persuasive counterargument against the given argument by considering the persona of the argument's author and responding from a given persona.

Given argument: {input}

Persona of the Argument's Author: {author_persona}
Given Persona: {given_persona}

Step 1 – Reasoning Plan
Explain how the given persona would respond based on their worldview and values.
The explanation should be clear, concise, and focused.
Limit your reasoning to 3–4 sentences.

Step 2 – Counterargument
Write a detailed and persuasive counterargument from the given persona's perspective.
Do not use any first-person language (e.g., "I", "we", "as a").
Let the persona's worldview emerge through tone, logic, and examples.
Aim to write in a way that could realistically persuade the original author, while keeping the tone respectful and well-reasoned.
You should consider the author when writing, but remember that the author persona is only an inferred estimation and should not be directly included in the content.
The counterargument must be at least 10 sentences and under 500 tokens.

You must format your output as follows:

Reasoning Plan:
[Your explanation here.]

Counterargument:
[Your counterargument here.]

Figure A5: Prompt used for tree-based step-wise generation of plan and counterarguments. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Tree-based Step-wise Selection Prompt:

Plan Selection:
You are given a given argument and multiple reasoning plans involving both an author persona and a counter persona.

Given argument:
{input}
Author persona: {author_persona}
Counter persona: {given_persona}

Choose the single plan most likely to yield a persuasive rebuttal.
Judge whether it makes effective use of the contrast between personas, applies a strong strategy, and presents its reasoning clearly and logically.

Do not explain your reasoning. Only output:
The best choice is X

Reasoning plan options:

Counterargument Selection:
You are given a given argument and multiple counterarguments involving both an author persona and a counter persona.

Given argument:
{input}
Author persona: {author_persona}
Counter persona: {given_persona}

Choose the most persuasive counterargument.
Judge whether it leverages the contrast between personas, directly challenges the argument, and is specific, persuasive, and logically consistent.

Do not explain your reasoning. Only output:
The best choice is X

Counterargument options:

Figure A6: Prompt used for tree-based step-wise selection of plan and counterarguments. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Write three diverse and persuasive counterarguments against the given argument.

Given argument: {input}

Instructions:
1. The counterargument must be logically sound, respectful in tone, and directly challenge the original argument.
2. Do not use any first-person language (e.g., "I", "we").
3. Instead, let the reasoning, tone, and examples naturally reflect a broader worldview or perspective.
4. Each counterargument should be at least 10 sentences long but under 500 tokens.
5. Focus on making the counterargument clear, well-reasoned, and persuasive to the original author.

Format your response as follows:

Counterargument 1:
[First counterargument]

Counterargument 2:
[Second counterargument]

Counterargument 3:
[Third counterargument]

Figure A7: Prompt used with a vanilla LLM to induce the generation of multiple counterarguments. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

General Persuasiveness:
[INSTRUCTION]
You are given an original opinion and a counterargument.
Rate the persuasiveness of the counterargument by focusing on its depth of insight and contextual authority.

A highly persuasive counterargument (8-10) should:
- Avoid "one-size-fits-all" or purely abstract logical responses.
- Speak from a specific and grounded position, using nuanced reasoning that suggests a deep, specialized understanding of the topic.
- Offer an "intellectual edge" by introducing a dimension of the issue that a general observer would likely miss.
- Maintain an authentic and compelling voice that feels human-like and deeply committed to its logic.

Output only a single integer between 1 and 10.

Score from 1–10:
1–2: Very weak. Feels robotic, predictable, or stays at a very superficial level of disagreement.
3–4: Limited. Addresses the point but lacks any distinctive "voice" or specialized angle; feels like a generic template.
5–6: Moderate. Logical and clear, but remains within the bounds of "common-sense" reasoning that anyone could provide.
7–8: Strong. Driven by a vivid and focused perspective; the argument feels authentic, specialized, and remarkably grounded.
9–10: Exceptional. The specialized depth and unique sharpness of the viewpoint are so profound that it adds a layer of realism and authority rarely seen in standard responses.

[Original Opinion]
{original_opinion}

[Counterargument]
{counterargument}

Persuasiveness Score (1–10):

Figure A8: Prompt used for evaluating general persuasiveness, where the counterargument is assessed against only the *title* of the original post (i.e., the main claim). Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Targeted Persuasiveness:
[INSTRUCTION]
You are given an original opinion and a counterargument.
Rate the persuasiveness by assessing how effectively the counterargument broadens the author's frame of reference.

A truly persuasive counterargument does more than just oppose; it reframes the issue through a specialized lens. High scores must be given to arguments that reveal a "blind spot" in the author's original view by introducing a nuanced and authoritative line of reasoning that feels lived-in and intellectually sharp.

Give a single score from 1 to 10, focusing on the authenticity and depth of the unique insight provided.

Score from 1–10:
1–2: Very weak. Uses broad or detached logic that fails to challenge the author's underlying assumptions.
3–4: Limited. Relies on standard "textbook" debate points; lacks a fresh or specialized insight that would prompt reflection.
5–6: Moderate. Offers a clear alternative view, but the reasoning remains within predictable and generic boundaries.
7–8: Strong. Successfully reveals a major blind spot in the original opinion by applying a remarkably sharp and focused perspective.
9–10: Definitive. Provides such a vivid and authoritative re-framing of the issue that it compels the author to fundamentally reconsider their entire standpoint.

Output only a single integer between 1 and 10.

[Original Opinion]
{original_opinion}

[Counterargument]
{counterargument}

Persuasiveness Score (1–10):

Figure A9: Prompt used for evaluating targeted persuasiveness, where the counterargument is assessed against both the *title* and *body* of the original post. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Perspective Diversity:
[INSTRUCTION]
Below is an original opinion and three counterarguments written in response.
Evaluate the overall diversity among the counterarguments using the criterion below.
Be as objective and concise as possible.
Provide only a score from 1 (very low diversity) to 5 (very high diversity).

[Original Opinion]
{original_opinion}

[Counterargument 1]
{counterargument_1}

[Counterargument 2]
{counterargument_2}

[Counterargument 3]
{counterargument_3}

[Evaluation Criterion]
Diversity: Assess whether the three counterarguments approach the original opinion from clearly different perspectives, drawing on distinct social identities, belief systems, or lived experiences. High scores should be given when each response plausibly reflects the worldview of a different kind of individual. Low scores indicate surface-level variation or repetition of the same underlying reasoning.

Evaluation Form (scores ONLY):

- Diversity:

Figure A10: Prompt used for evaluating perspective diversity, assessing whether the generated counterarguments reflect distinct viewpoints. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.

Quality:
[INSTRUCTION]
Below is an original opinion and a counterargument written in response.
Evaluate the counterargument based on the criterion below.
Be as objective as possible.
For each aspect, provide only score from 1 (worst) to 5 (best).

[Original Opinion]
{original_opinion}

[Counterargument]
{counterargument}

[Evaluation Criteria]
{criteria}: {criteria_description}

Evaluation Form (scores ONLY):

- {criteria}:

criteria: criteria_description
Grammaticality: Evaluate whether the text adheres to standard grammar conventions, focusing on the mastery of sophisticated sentence structures and precise syntax that support complex argumentation.

Appropriateness: Evaluate whether the language and tone are suitable for the context, specifically checking if the intellectual depth and professional register are proportional to the significance of the issue.

Relevance: Evaluate how directly the counterargument engages with the original opinion, rewarding responses that penetrate the underlying assumptions and address key points with multi-faceted insights.

Clarity: Evaluate whether the writing is clear and well-organized, ensuring that advanced concepts are articulated through a systematic structure without unnecessary ambiguity.

Figure A11: This prompt is used to evaluate quality, covering *appropriateness*, *clarity*, *grammaticality*, and *relevance*. The variables `criteria` and `criteria_description` are defined in detail below. Curly-braced placeholders (e.g., {}) are replaced with task-specific variables.