AMERICANO: Argument Generation with Discourse-driven Decomposition and Agent Interaction

Anonymous ACL submission

Abstract

Argument generation is a challenging task in natural language processing, which requires rigorous reasoning and proper content organization. Inspired by recent chain-of-thought prompting that breaks down a complex task into intermediate steps, we propose AMERICANO, a novel framework with agent interaction for argument generation. Our approach decomposes the generation process into sequential actions grounded on argumentation theory, which first executes actions sequentially to generate argumentative discourse components, and then produces a final argument conditioned on the components. To further mimic the human writing process and improve the left-to-right generation paradigm of current autoregressive language models, we introduce an argument refinement module that automatically evaluates and refines argument drafts based on feedback received. We evaluate our framework on the task of counterargument generation using a subset of Reddit/CMV dataset. The results show that our method outperforms both end-to-end and chain-of-thought prompting methods and can generate more coherent and persuasive arguments with diverse and rich contents.

1 Introduction

Argument generation is an essential task in natural language processing with wide applications, such as debates and essay writing (Toulmin, 2003). In this work, we study *counterargument generation* which aims to generate persuasive arguments to refute a given proposition on a controversial topic (Hua and Wang, 2018; Alshomary and Wachsmuth, 2023). However, generating counterarguments poses significant challenges for both humans and machines as it requires a profound comprehension of the original proposition, the ability to present a valid standpoint from an opposing perspective, and the competence to provide rigorous reasoning to justify the claim (Antaki and Wetherell, 1999; Grote et al., 1997; Walton et al., 2008; Wang et al., 2017).

Proposition: People shouldn't change gender

Counterargument:

① [Individuals should not be prohibited from changing their gender, as it is a personal choice that ought to be respected and constitutes a fundamental aspect of one's identity.]
② [Denying someone the right to undergo a gender transition infringes upon their basic human rights, as it disregards their autonomy and deprives them of their freedom to make decisions about their own body and identity. Research has demonstrated that the positive effects of gender transition have significantly increased in recent years, due to advancements in surgical techniques and enhanced social support.] ③ *[Although some may argue that changing one's gender is an act of rebellion and contravenes traditional values and norms, this argument does not negate the essential right of individuals to determine how they express themselves and live their lives as they deem fit.]* ④ [This is a fundamental human right that must be acknowledged and protected].

Figure 1: Sample counterargument that refutes the proposition. The argument structure consists of components including ① a *claim* serving as the main statement to attack the proposition, ② a *reasoning* that supports the claim, ③ a *concession* responding with potential rebuttals and ④ a *conclusion*.

Recent large language models (LLMs) have exhibited remarkable capabilities in addressing various tasks with human-alike result (Brown et al., 2020a; Ouyang et al., 2022; OpenAI, 2023; Chowdhery et al., 2022). However, the token-level autoregressive generation paradigm makes LLMs fall short of dealing with complicated tasks involving multiple actions due to the lack of *high-level planning* ability (Bubeck et al., 2023). Prior work shows that chain-of-thought (CoT) prompting can significantly boost the LLMs' ability on complex reasoning tasks by encouraging the model to decompose the task into a sequence of intermediate results (Wei et al., 2022). Later work further imposes automatic decision-making and action-executing to break down complex tasks leveraging LLMs (Shinn et al., 2023; Yao et al., 2022; Sun et al., 2023a).

Although the above methods achieve good performance in solving reasoning tasks, they still face

challenges when applied to argument generation. Generating arguments *not only requires rigorous reasoning but also demands deliberate discourse structures to enhance overall coherence and persuasion* (Musi et al., 2018; Hua and Wang, 2020). As shown in Figure 1, a counterargument comprises several discourse components, and generating a strong argument needs both to produce high-quality components and to properly organize the components to ensure overall quality. Nevertheless, decomposing the goal of argument generation into intermediate actions remains a non-trivial task. Moreover, *the left-to-right single-pass generation paradigm of current LLMs hinders them from tracking back and revising in previously generated text*. This limitation potentially depletes the soundness and coherence of the generated argument (Wang et al., 2018; Madaan et al., 2023; Hu et al., 2022a).

In this work, we propose **AMERICANO**, a novel framework for argument generation with discourse-driven decomposition and agent interaction, where a **generation agent** first produces an argument draft, and then an **evaluation agent** and **refinement agent** iteratively produce feedback and revise the draft. Inspired by argumentation theory and argumentative discourse structure (Van Eemeren and Grootendorst, 2004; Green, 2010; Palau and Moens, 2009), our argument generation agent decomposes the goal into predefined actions and sequentially generates each argumentative discourse component. Specifically, given a proposition and the goal of generating a counterargument, the sequential actions aiming to create high-quality discourse components include: (1) a *claim action* that produces a strong claim to refute the proposition; (2) a *reasoning action* that generates and revises a detailed logical reasoning to support the claim; (3) a *concession action* that creates potential acknowledgements of the original proposition. Following the generation of these intermediate discourse components, an *argument generation action* is executed to organize the intermediate contents and generate a final counterargument.

To further mitigate the drawback of left-to-right generation and incorporate feedback, we propose an argument refinement module with two agents - an evaluation agent and a refinement agent. Specifically, the argument draft is first evaluated by the evaluate agent to provide verbal feedback signals, and then the feedback is passed to the refinement agent to revise the draft. This process can be conducted iteratively until the evaluator is satisfied with the result. Both agents are operated by prompting LLMs without any model training. This is also akin to the human writing process of first composing a draft and then revising the draft (Flower and Hayes, 1981) to improve the quality.

We evaluate our framework on the task of zero-shot counterargument generation, with a subset of propositions collected from Reddit/CMV dataset. We leverage both LLM-based automatic evaluation and human evaluation to validate the model outputs. The results show that our method is able to produce high-quality counterarguments with better coherence and persuasiveness compared with end-to-end prompting and CoT prompting. Moreover, our system can generate more diverse results than baseline methods. [1]

2 Argument Generation with Discourse-driven Sequential Actions

The overall framework is shown in Figure 2, which consists of three agents that collaboratively perform task decomposition and refinement for argument generation. We first introduce the generation agent.

Argument generation can be modeled as $p(y|x)$, where x is an input proposition and y is an output counterargument. However, directly modeling this probability presents significant challenges, as generating arguments necessitates appropriate high-level planning, rigorous reasoning, and proper content organization. Instead of directly prompting LLMs for argument generation, we decompose the goal into a sequence of actions based on argumentative discourse structure (Stab and Gurevych, 2014; Madnani et al., 2012; Wambsganss and Niklaus, 2022). Each action tackles a subproblem based on the internal structure of an argument, which typically includes: a **claim** as the central statement the writer is trying to argue, a **reasoning** to support the claim, and an optional **concession/acknowledgement** to address potential dissenters and improve persuasion. [2]

Driven by this, we break down the generation into sequential actions that first generate the components and then produce a final argument: $p(y|x) = p(y|a, r, c, x)p(a|r, c, x)p(r|c, x)p(c|x)$, where c, r, a denotes claim, reasoning and acknowledge/concession respectively. Such modeling reduces the complexity of $p(y|x)$. All the actions

[1] We will release code and data upon publication.

[2] We do not explicitly include a conclusion as the main claim can often be restated as the conclusion.

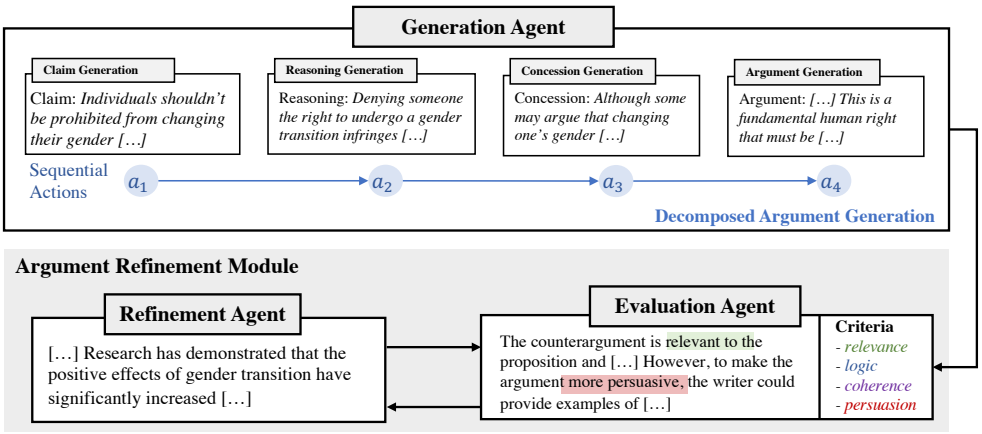

Figure 2: Overview of our framework. The generator first decomposes the task into a sequence of actions and produces an initial result. Then, a refinement module with two agents iteratively provides feedback and revises the result.

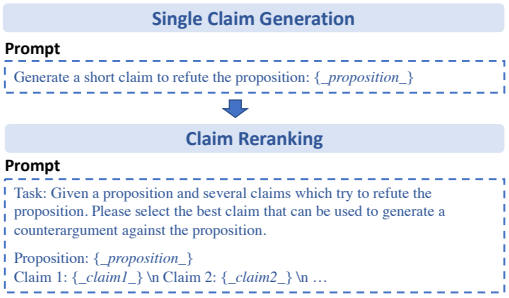

Figure 3: Prompts for claim generation.

are conducted by prompting the same LLM ($\mathcal{M}$), eliminating the costly model training.

2.1 Claim Generation Action

The claim is the central component of an argument. For counterarguments, it should express a different stance regarding the proposition. As shown in Figure 3, we prompt $\mathcal{M}$ to generate a potential claim. However, multiple valid claims may exist given an input proposition. Therefore, instead of executing the action only once, we prompt $\mathcal{M}$ multiple times to produce a set of claims and then introduce a claim reranking step to select the best one.

For claim reranking, we again utilize $\mathcal{M}$ to rank the claims based on the potential to generate a persuasive argument. To reduce variance and improve the self-consistency of the ranking, we further introduce a majority voting strategy by prompting $\mathcal{M}$ multiple times and selecting the claim that is ranked as topmost with the highest frequency. This simple strategy has proven effective in other tasks such as CoT prompting (Wang et al., 2022).

2.2 Reasoning Generation Action

Reasoning generation action aims to produce a comprehensive reasoning conditioned on both proposition (x) and the previously generated claim

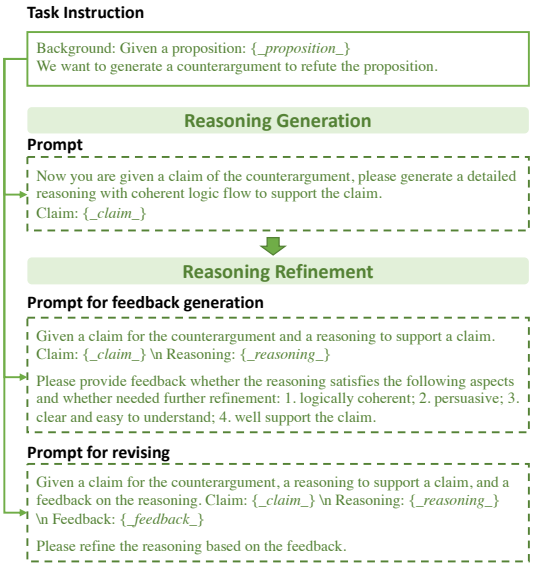

Figure 4: Prompts for reasoning generation.

(c). As illustrated in Figure 4, we first employ $\mathcal{M}$ to create an initial reasoning using the concatenation of the task instruction and prompt. Additionally, we leverage an off-the-shelf NLI model[3] to verify that generated reasoning entails the claim.

However, generating high-quality reasoning requires strict logical inference and internal consistency, which is difficult to achieve by only prompting LLMs once. Therefore, we leverage $\mathcal{M}$ as a critic to provide feedback and reinforce the generator to progressively revise the reasoning. We employ pre-defined criteria as verbal prompts, addressing aspects including logical coherence, persuasiveness, and whether the reasoning makes sense and well supports the claim. The generator then modifies the reasoning by additionally consuming the feedback. This process is conducted iteratively until no feedback is required or the maximum num-

[3] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

Argument Generation

Task Instruction

Background: Given a proposition: {*_proposition_*}
We want to generate a counterargument to refute the proposition.

Prompt

Generate a persuasive and coherent counterargument leveraging the given claim, reasoning and concession to refute the given proposition.

Requirements

- **Stance**: The counterargument should be against to the given proposition;
- **Coherence**: The counterargument should be well-structured and organized in a coherent manner. Use appropriate transitions to connect the ideas;
- **Clarity**: Do not just copy the given contents, but summarize and reorganize the contents to make the counterargument clear and persuasive;
- …

Figure 5: Prompts for argument generation.

ber of iterations is reached. This ensures a strong reasoning is generated, which can be utilized to enhance the subsequent counterargument generation.

2.3 Concession Generation Action

Concessions are considered as an argumentative strategy that enhances persuasion in discourse studies (Mann and Thompson, 1988; Musi et al., 2018; Antaki and Wetherell, 1999; Wolfe et al., 2009). A concession, or acknowledgement, is typically employed to produce trust and fortify one's position by addressing potential dissenters in an argument.

This action aims to generate a concession based on the proposition, the previously generated claim, and reasoning. Similarly, we utilize $\mathcal{M}$ for concession generation. It is crucial to note that the concession should not weaken the original counterargument, and thus we include the following instruction in the prompt:

"Note that the goal of the concession is not to weak the claim and reasoning, but to produce trust and make the counterargument more convincing and persuasive to the audience."

This instruction has proven effective in our experiments. The full prompt is in the Appendix B.

2.4 Counterargument Generation Action

Thus far, we have generated all the essential components of an argument, including a claim, reasoning and concession. Next, we generate the final counterargument based on these components. This step requires properly understanding the components and effectively organizing the content to produce a coherent outcome. We again rely on $\mathcal{M}$ to execute the action. As shown in Figure 5, besides the task instruction and prompt, we further include predefined requirements on aspects including stance, coherence, and clarity, to enhance overall performance and effectiveness.

3 Argument Refinement Module

Previous work has shown that producing an output on a single attempt is challenging for both machines and human beings (Hua and Wang, 2020; Hu et al., 2022a; Wang et al., 2018). Conventional autoregressive language models produce outputs from left to right at the token level, lacking the capacity to edit and revise previously generated content. Drawing inspiration from the human writing process that involves first creating an initial draft and subsequently refining it, we propose an argument refinement module to mimic this process. As illustrated in Figure 2, this module comprises an evaluation agent and a refinement agent. The evaluator first provides feedback on the current draft, and then the refinement agent takes the feedback and revises the draft. Two agents interact with each other interatively to formalize an optimization process for generation.

3.1 Evaluation Agent for Feedback Generation

Given a proposition and an initial counterargument draft, the evaluation agent first provides feedback on improving the counterargument. First, a valid counterargument should possess an opposing stance compared with the original proposition, and hence we leverage the same NLI model as in the reasoning generation to compute the relationship class s_{arg} between the proposition and the counterargument. This result will be used in later steps if the predicted label does not correspond to "CONTRADICTION".

Furthermore, we leverage $\mathcal{M}$ to assess the counterargument draft and generate feedback. The evaluation criteria for counterargument include aspects of **relevance**, **logical consistency**, **coherence of structure**, and **persuasion**. The detailed prompts can be found in the Appendix B.

3.2 Refinement Agent

The refinement agent takes as input the feedback and generates a revised version of the counterargument in each iteration. Concretely, it first verifies the stance based on the prediction of the NLI model: if the NLI label is not "CONTRADICTION", it first utilizes $\mathcal{M}$ to adjust the draft so that its stance aligns with a valid counterargument that refutes the original proposition, with the prompt: *"The stance is wrong. The counterargument should be against the statement."*. Subsequently, it refines the coun-

terargument by addressing the feedback from the evaluator to enhance the overall quality. The two agents work together in a loop until the evaluator is satisfied with the result. In practice, we bound the process by a maximum number of iterations.

Our refinement module distinguishes itself from Self-refine (Madaan et al., 2023) in the way that they leverage the same LLM instance to serve as the generator, evaluator, and revisor, without any task decomposition. In contrast, our generation agent features a sequence of actions designed to produce high-quality initial results, offering a superior starting point for the refinement process, ultimately resulting in enhanced efficiency and effectiveness.

4 Experiment Setup and Evaluation

4.1 Task Setup

We evaluate our framework on the task of counterargument under a zero-shot setting, where the model is asked to generate a counterargument to refute a given proposition on a controversial topic. We randomly sample 50 propositions from Reddit/CMV dataset (Hua et al., 2021; Hu et al., 2022b), which is a counterargument generation dataset with samples collected from Reddit/ChangeMyView. All propositions are in the politics and policy domains.

4.2 Model Implementations and Baselines

As we study zero-shot argument generation, we compare our model with unsupervised baselines by prompting LLMs. We use GPT-3.5 (text-davinci-003) as the base LLM. We consider the following baselines: (1) End-to-end generation (E2E) which directly prompts the LLM to generate a counterargument without any intermediate steps; (2) Plan-based CoT generation (PlanCoT) that first generates a chain of plans as intermediate content planning, and then produces the counterargument based on the plan; (3) Our model variant without refinement module. All the baseline models use the same GPT-3.5 version as our framework. More details are in Appendix A.

4.3 Evaluation Metrics

We employ both automatic and human evaluations in our experiments. Automatically evaluating open-ended text generation tasks is a challenging task (Celikyilmaz et al., 2020). Recent work leverages LLMs to conduct reference-free text generation evaluation on tasks such as summarization and dialogue generation (Liu et al., 2023; Fu et al., 2023). Therefore, we propose a LLM-based counterargument evaluation method leveraging GPT-4 (OpenAI, 2023). We mainly evaluate two aspects: **coherence** and **persuasion**.

4.3.1 LLM-based Automatic Evaluation

We leverage GPT-4 to evaluate coherence and persuasion by scoring the outputs on a scale of 1 to 5, with the higher score signifying superior quality. To reduce randomness, we evaluate each sample 5 times and average the scores.

The prompts used for evaluation are designed with specific task instructions and a comprehensive list of detailed criteria. For coherence, we concentrate on assessing both logical and discourse coherence, measuring the score jointly based on clarity, relevance to the proposition, logical consistency and soundness of reasoning. For persuasion, we appraise the outputs according to language and rhetoric usage, the ability to address opposing viewpoints, credibility of evidence, and the overall effectiveness to persuade the audience. Each aspect comes with a detailed explanation. To improve stability, we prompt model to first generate a detailed rationale and then predict the score. The specific prompts and more details are in Appendix C.

4.3.2 Human Evaluation

For human evaluations, we hire three proficient English speakers as judges to evaluate output quality on the following aspects: **Appropriateness**-measures if an output is clear, readable and logical consistent; **Content Richness**-represents the amount of informative talking points; and **Overall Quality**. Given an input proposition and several model outputs, the judges are asked to rank the outputs according to each aspect. In addition, we ask the judges to identify **valid** counterarguments of high quality, disregarding relative rankings. [4] We select 30 random instances for evaluation.

5 Results and Analysis

5.1 Automatic Results

The LLM-based evaluation results on coherence and persuasion are displayed in Figure 6. As can be seen, our method outperforms all baselines in terms of persuasion and coherence, demonstrating the effectiveness of our framework in generating high-quality arguments. Specifically, for coherence, we observe that decomposing the generation

[4] Detailed guidelines are in Appendix F.

Method	Appropriateness (↑)	Content Richness (↑)	Overall Quality (↑)	% Validness (↑)
E2E	1.81 / 6.7%	1.54 / 0.0%	1.78 / 4.4%	45.00%
PlanCoT	2.00 / 14.4%	1.74 / 6.7%	2.99 / 10.0%	53.33%
$\text{Ours}_{\text{w/o Refine}}$	3.00 / 34.4%	3.16 / 32.2%	3.03 / 28.9%	91.67%
Ours	**3.12 / 44.4%**	**3.44 / 61.1%**	**3.31 / 56.7%**	**86.67%**

Table 1: Human evaluation results. For appropriateness, informativeness, and overall quality, the first score is computed based on the relative ranking position, and the second value represents the frequency of the output being ranked as the topmost. For validity, we present the percentage of results that are considered valid results of high quality.

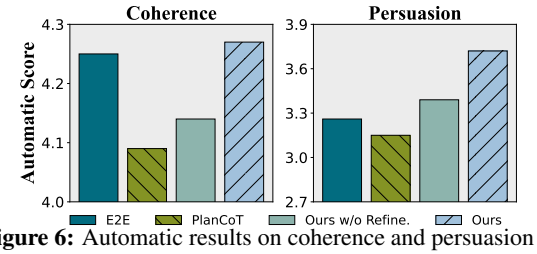

Figure 6: Automatic results on coherence and persuasion by GPT4-based evaluation. A larger score means better quality.

(Ours w/o Refine) results in reduced coherence compared with E2E. One possible reason is that generating a final argument based on argumentative discourse components requires a deep understanding of each component and proper content organization, posing challenges when executing the argument generation action only once. Especially, our decomposed generation tends to produce longer outputs [5], further complicating the task of generating a coherent result in a single step. However, incorporating the refinement module significantly boosts the coherence score, proving the importance of the refinement module in improving coherence. For persuasion, both our model and the decomposed generation achieve higher scores compared to E2E and PlanCoT. Our manual inspections show that our model outputs tend to include more talking points in the arguments, thus making the results more persuasive.

5.2 Human Evaluation Results

The human evaluation results are shown in Table 1. As the evaluation of appropriateness, informativeness and overall quality are ranking-based, we convert the ranks into scores determined by subtracting its position in the ranking from the total number of candidates, with higher scores indicating better quality. We also present the percentage of times the result is considered the top one for each aspect.

First, $\text{Ours}_{\text{w/o Refine}}$'s results are ranked higher on all aspects compared with E2E and PlanCoT. This demonstrates that breaking down the E2E generation helps to maintain high-quality discourse components, ultimately leading to improved quality of final arguments. Second, the better content richness of our decomposed generation indicates that our model can produce outputs with more informative talking points to support the claim. This can be attributed to the reasoning generation action's ability to revise reasoning and make it stronger. Third, our discourse-driven sequential actions are more effective at improving the results compared with PlanCoT's content-based plans, making them better suited for argument generation.

By incorporating the refinement module to further reinforce the generation process, the results exhibit substantial improvements across all aspects. Our manual inspection reveals that, on the one hand, the refinement module can reorganize argument content and reform the draft to achieve a better discourse structure, increasing the readability and coherence of the argument. On the other hand, during the refinement process, the model tends to add more detailed examples to support the claim, enhancing content richness and overall persuasion.

We also ask human judges to determine whether a generated result qualifies as a valid high-quality counterargument. While only 45% of E2E generation results are considered valid, introducing CoT improves the outcomes, showing that decomposing complex goals is beneficial. $\text{Ours}_{\text{w/o Refine}}$ with discourse-driven actions achieves significantly better results, with almost 92% of samples considered valid counterarguments, validating the effectiveness of incorporating discourse information into sequential actions. Interestingly, our full model with refinement scores approximately 87% in validity, which is slightly lower than $\text{Ours}_{\text{w/o Refine}}$. The manual inspection reveals that, occasionally, the refinement process may introduce redundancies to the outputs and diminish overall readability, which we leave for future work.

5.3 Analysis on Output Diversity

We follow previous work and analyze output diversity by calculating the average number of distinct bigrams (Li et al., 2016) and content words in each output. The results are in Figure 7. Our method

[5] We provide additional analysis on the impact of output length in Appendix D.

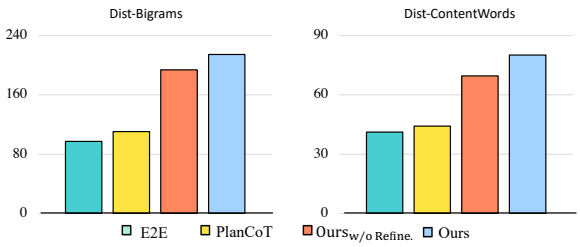

Figure 7: Average number of distinct bigrams (Dist-Bigrams) and content words (Dist-ContentWords).

generates the most distinct bigrams and content words, demonstrating its ability to produce more diverse outputs. In contrast, E2E produces the least diverse outputs. We hypothesize that this may result from the fact that an LLM trained with RLHF tends to produce safer outputs without directly optimization for diversity. After applying chain-of-thought, PlanCoT generates more diverse results. Leveraging our discourse-driven actions further improves scores, indicating that decomposing argument generation based on discourse components effectively enhances output diversity and content richness. We also present commonly generated verbs and nouns in Appendix E.

5.4 Further Analysis on Discourse

Usage of Discourse Markers. Discourse markers are key features for modeling coherence (Callaway, 2003; Grote and Stede, 1998) in various tasks (Eckle-Kohler et al., 2015; Sharma et al., 2019; Samy and González-Ledesma, 2008). We adopt the discourse markers from Sileo et al. (2019) and extract frequent discourse markers that appear more than 10 times for each model, and present the results alongside their frequency in Figure 8.

While only 4 common discourse markers are observed in the E2E results, applying chain-of-thought increases the usage of more markers. This is consistent with previous observations that chain-of-thought contributes to improved diversity. Moreover, both our methods, with or without refinement, leverage significantly more discourse markers. Compared with E2E and PlanCoT, our model variants employ discourse markers such as "though" and "regardless" because of the inclusion of concession components. Furthermore, the use of "for example" implies that our model variants learn to include more examples to support the claim, making the overall argument more persuasive.

Analysis of RST Tree. The discourse structure provides insight into the high-level organization of text. We analyze the Rhetorical Structure Theory (RST) tree of the generated arguments. Concretely,

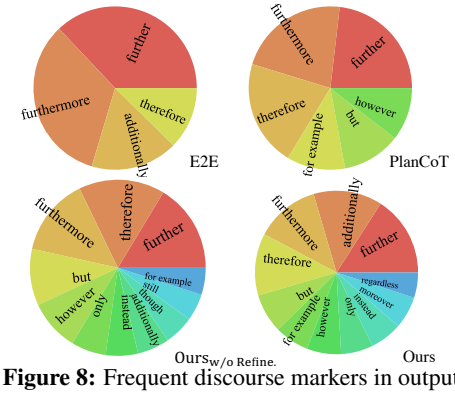

Figure 8: Frequent discourse markers in outputs.

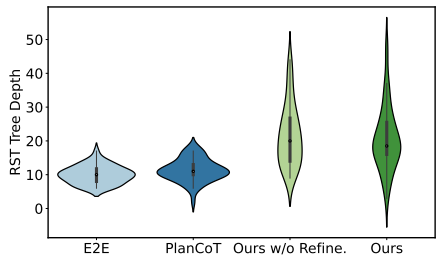

Figure 9: Distribution of RST tree depth of arguments.

we utilize an off-the-shelf discourse parsing tool (Ji and Eisenstein, 2014) to convert the arguments into RST trees and analyze the depth of the trees, illustrated in Figure 9. As can be seen, our model variants yield arguments with deeper structures. Moreover, a noticeable diversity in depth can be observed, spanning a broader spectrum compared to both E2E and PlanCoT models This further implies that our model can produce arguments with more diverse and complicated structures.

5.5 Sample Output Analysis

We present a sample output with intermediate results in Figure 10. Given a proposition, our sequential actions first properly generate each discourse component and then organize them to form a coherent argument. Moreover, the final argument conforms to the discourse components and faithfully reflects each reasoning point, further indicating the strong controllability of the decomposed generation. It also implies our model's potential to be applied to *interactive writing* wherein users could modify intermediate components and let the model organize the contents and generate final results. We leave this to future work. Notably, during the refinement process, our model effectively adds *examples in the concession part* to strengthen the whole argument. This proves that our refinement process can gradually improve the results. In addition, the underlined discourse markers in the final argument show that our model can properly utilize discourse markers and generate coherent outputs. More samples can be found in Appendix G.

Proposition: The fact that voting isn't mandatory is a good thing.

Claim: Voting should be mandatory because it is a civic duty that ensures that the people's voices are heard and that their opinions are represented in government.

Reasoning: [...] ①Mandatory voting is a way of ensuring that all citizens are engaged in the electoral process and that their voices are taken into consideration when decisions are being made. ②It also helps to reduce the influence of special interests and ensures that the government truly reflects the will of the people. ③Furthermore, mandatory voting can help to reduce the influence of money in politics by ensuring that all citizens are represented and can participate in the democratic process. [...]

Concession: While it is true that mandatory voting can ensure that the voices of all citizens are heard, it is important to consider how to make the process accessible and fair for those who are unable to vote due to physical or financial limitations.

Counterargument: The fact that voting isn't mandatory is not a good thing. ①Mandatory voting should be implemented because it is a fundamental part of democracy and is an essential way to ensure that citizens are engaged in the electoral process and their voices are taken into consideration when decisions are made. ②By requiring citizens to vote, it can help to reduce the influence of money in politics, as more people are participating in the democratic process and their opinions are being represented. ③Additionally, mandatory voting can help to ensure that the government truly reflects the will of the people, as all citizens have an equal say in the political process.

However, it is important to consider how to make the process accessible and fair for those who are unable to vote due to physical or financial limitations. *For example, measures such as providing transportation to and from polling stations, offering absentee ballots, and introducing flexible voting hours can help to make the process easier and more accessible for those who have difficulty voting in person.* Additionally, lawmakers should focus on creating laws that promote fairness and accessibility for all citizens, regardless of their socio-economic status.

Overall, mandatory voting is a way to guarantee that all citizens have a voice in the political process and to ensure that the government is truly representative of the people's wishes. By implementing measures that make the voting process easier and more accessible for all citizens, mandatory voting can help ensure that everyone is given an equal chance to participate and have their voices heard

Figure 10: Sample output of our method with the intermediate discourse components.

6 Related Work

Task Decomposition and Reasoning. LLMs have achieved impressive results in solving various tasks with prompting (OpenAI, 2023; Brown et al., 2020b; Anil et al., 2023; Bubeck et al., 2023). However, the token-level left-to-right generation process limits the model's ability to tackle more complex tasks (Bubeck et al., 2023). To further improve model ability on complex reasoning tasks, recent work involves sampling intermediate reasoning steps (Wei et al., 2022; Nye et al., 2021; Wang et al., 2022) or decomposing the complicated goal into actions (Sun et al., 2023b; Hao et al., 2023; Zhou et al., 2022; Chen et al., 2023). In this paper, we focus on the specific task of argument generation and decompose the goal into a sequence of predefined actions based on the argumentative theory to generate each discourse component.

Argument Generation. Argument generation requires text planning, logical reasoning, and content organization (Carenini and Moore, 2006; Hua and Wang, 2018). Hua et al. (2019) propose a planning-based model with a retrieval module for counterargument generation. Schiller et al. (2021) utilize keywords to control the content of arguments. Bao et al. (2022) introduce a dual-decoder model to improve content planning. Different from previous work, we leverage LLMs and introduce a framework with multi-agents for counterargument generation. Our method effectively decomposes argument generation into subproblems and prompts LLMs for each action without model training.

Feedback and Refinement for Text Generation. Previous work refines text generation by directly revising outputs without feedback (Wang et al., 2018; Hu et al., 2022a) or masking contents with low probability (Hua and Wang, 2020). Recent work utilizes LLMs to provide feedback and reinforce language agents to improve model performance (Shinn et al., 2023; Sun et al., 2023a; Madaan et al., 2023; Liang et al., 2023). In this work, we introduce a refinement module with specifically designed criteria for argument refinement. Different from Madaan et al. (2023) which uses only one LLM instance for generation, evaluation, and refinement, our system consists of multiple agents that decompose generation with sequential actions, thus providing a better starting point for the refinement module and further encouraging divergent thinking of LLMs.

7 Conclusion

In this work, we present a novel framework for argument generation with agent interaction. Our framework consists of a generation agent that decomposes argument generation into a sequence of predefined actions driven by argumentative theory to produce a draft, and then a refinement module with an evaluator and a refinement agent to iteratively provide feedback and refine the draft. All parts are implemented leveraging LLMs with zero-shot prompting. Both automatic and human evaluations show that our framework can generate more coherent and persuasive results with better diversity in counterargument generation.

Limitations

Argument generation is a challenging task in natural language processing. In this work we propose a multi-agent based framework utilizing LLMs for counterargument generation. However, there are several directions to explore in the future work. First, in our system, the refinement module only revises the argument draft without directly modifying the actions in the generation agent. However, the feedback can be incorporated to further improve actions for initial argument draft generation. Second, debating is an interactive process where two sides can interactively debate with each other in a conversational way. Future work might study interactive argumentation with multiple debating agents. Third, decomposing generation into intermediate steps can enable users to collaboratively modify the intermediate results, which we do not tackle in this paper and leave to future work.

Ethics Statement

We recognize that our framework may generate fabricated and potentially harmful content due to the systematic biases of pre-training using heterogeneous web corpora and the open-ended generation characteristics of the argumentative text generation tasks. As our method utilizes large language models and does not require model training, the generated outputs may contain harmful and biased contents as the generation of language models can not be fully controlled. We urge the users to carefully examine the ethical influence of the generated outputs and cautiously apply the system in real-world applications.

References

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.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Charles Antaki and Margaret Wetherell. 1999. Show concessions. *Discourse studies*, 1(1):7–27.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Charles B. Callaway. 2003. Integrating discourse markers into a pipelined natural language generation architecture. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 264–271, Sapporo, Japan. Association for Computational Linguistics.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.

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

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*,

pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Nancy L Green. 2010. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24:181–196.

Brigitte Grote, Nils Lenke, and Manfed Stede. 1997. Ma (r) king concessions in english and german. *Discourse processes*, 24(1):87–117.

Brigitte Grote and Manfred Stede. 1998. Discourse marker choice in sentence planning. In *Natural Language Generation*, Niagara-on-the-Lake, Ontario, Canada. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Zhe Hu, Hou Pong Chan, and Lifu Huang. 2022a. MOCHA: A multi-task training approach for coherent text generation from cognitive perspective. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10324–10334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022b. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. 2021. DYPLOC: Dynamic planning of content using mixed language models for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6408–6423, Online. 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.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2018. Changemyview through concessions: Do concessions increase persuasion? *arXiv preprint arXiv:1806.03223*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma,

David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Doaa Samy and Ana González-Ledesma. 2008. Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic- Spanish-English). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023a. Adaplanner: Adaptive planning from feedback with language models. *arXiv preprint arXiv:2305.16653*.

Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2023b. Pearl: Prompting large language models to plan and execute actions over long documents. *arXiv preprint arXiv:2305.14564*.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Frans H Van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Paper abstract writing through editing mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 260–265, Melbourne, Australia. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Christopher R Wolfe, M Anne Britt, and Jodie A Butler. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Experimental Details

In our experiments, all modules of our methods and baselines are implemented by prompting GPT-3.5 (text-davinci-003) [6] as the based LLM. For hyper-parameters, we set temperature as 0.7 and top-p as 1, the maximum tokens are set as 2048. For claim generation, we set the number of claims to be generated as 5. We set the maximum of iteration for refining reasoning and final argument as 3 and 1 respectively, considering the cost of API. For automatic evaluation, we use GPT4 (gpt-4-0314) [7] as the base model.

Discourse Markers. For the result analysis on discourse markers, we select markers from Sileo et al. (2019). Some common markers such as "and", "or" are removed from the list. The complete list is presented in Figure 11.

B Detailed Prompts

Here we provide detailed prompts for each module. Specifically, the prompt for concession generation is presented in Figure 15. For the argument refinement module in our framework, the detailed prompt for the evaluation agent is presented in Figure 17, and the detailed prompt for the refinemnt agent in shown in Figure 18.

The prompt for PlanCoT is presented in Figure 16. For PlanCoT, we match the content of Counterargument as the final results, and do not use the plan in our experiments.

C Automatic Evaluation with GPT4

For GPT4-based automatic evaluation as described in Section 4.3.1, the prompts used for evaluation are designed with specific task instructions and a comprehensive list of detailed criteria, as in Figure 13. We present the detailed criteria for coherence and persuasion in Figure 19. The description of the criteria is concatenated with the task instruction as the final prompt.

In our initial experiments, we find that the GPT-4 predictions are not very stable. This observation is consistent with prior work (Shen et al., 2023; Wang et al., 2023). To mitigate this problem, instead of directly prompting GPT-4 to predict a score, we first ask it to provide a detailed rationale on evaluation, and then predict the score, which is similar to chain-of-thought prompting (Wei et al., 2022). By this strategy, we find the stability of predictions improves by a large margin.

Another observation is that when evaluating coherence, GPT-4 evaluator tends to prefer shorter results or longer outputs with multiple paragraphs. This is a possible reason that in Figure 6, our decomposed generation receives a lower score on coherence compared with E2E. However, our model with refinement achieves a higher coherence score, as during the refinement process, the refinement agent tends to produce longer outputs with more

absolutely, accordingly, actually, additionally, admittedly, afterward, again, already, alternately, alternatively, although, altogether, amazingly, anyway, apparently, arguably, as a result, basically, because of that, because of this, besides, but, by comparison, by contrast, by doing this, by then, certainly, clearly, coincidentally, collectively, consequently, conversely, curiously, currently, elsewhere, especially, essentially, eventually, evidently, finally, first, firstly, for example, for instance, fortunately, frankly, frequently, further, furthermore, generally, gradually, happily, hence, historically, honestly, hopefully, however, ideally, immediately, importantly, in contrast, in fact, in other words, in particular, in short, in sum, in the end, in the meantime, in turn, incidentally, increasingly, indeed, inevitably, initially, instead, interestingly, ironically, lastly, lately, later, likewise, locally, luckily, maybe, meaning, meantime, meanwhile, moreover, mostly, namely, nationally, naturally, nevertheless, next, nonetheless, normally, notably, obviously, occasionally, oddly, often, on the contrary, on the other hand, once, only, optionally, originally, otherwise, overall, particularly, perhaps, personally, plus, preferably, presently, presumably, previously, probably, rather, realistically, really, recently, regardless, remarkably, sadly, second, secondly, separately, seriously, significantly, similarly, simultaneously, slowly, sometimes, soon, specifically, still, strangely, subsequently, suddenly, supposedly, surely, surprisingly, technically, thankfully, then, theoretically, thereafter, thereby, therefore, third, thirdly, though, thus, together, traditionally, truly, truthfully, typically, ultimately, undoubtedly, unfortunately, unsurprisingly, usually, yet

Figure 11: List of discourse markers for result analysis.

Method	Coherence	Persuasion
E2E	3.84	3.64
PlanCoT	3.83	3.58
Ours	**3.87**	**3.78**

Table 2: GPT4-based automatic evaluations of generated arguments under length constraints.

[6] https://platform.openai.com/docs/model-index-for-researchers

[7] https://platform.openai.com/docs/models/gpt-4

Method	Overall Quality
E2E	24.7%
Ours	51.7%
PlanCoT	17.8%
Ours	70.0%

Table 3: Pairwise human evaluations on overall quality. We report percentage of times the results considered as better.

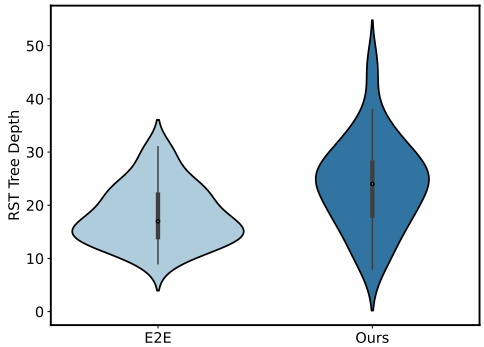

Figure 12: Distribution of RST tree depth of generated arguments under length constraints.

paragraphs (e.g., average 2.44 paragraphs for ours v.s. 2.12 paragraphs for ours w/o refine). We leave further analysis to future work.

D Analysis on Model Performance Under Length Constraints

In our experiments, we do not impose specific length constraints on the generated outputs due to the open-ended nature of the argument generation task. Our model variants with decomposed generation tend to produce longer outputs than baseline methods (i.e., on average 310 words for our model v.s. 120 words for E2E). In this section, we further analyze the influence of introducing length constraints by specifying the desired output length. In particular, we explicitly include "*counterargument in around 300 words*" in the prompts for all methods and further analyze the results. By doing so, the average output lengths of our model, E2E, and PlanCoT change to around 378, 303, and 240 words, respectively.

The GPT4-based automatic scores are shown in Table 2. As can be seen, our model outperforms both E2E and PlanCoT in terms of coherence and persuasion scores. The results are consistent with the observations where no length constraints are imposed. These findings confirm that our approach, with decomposed generation and subsequent refinement, is highly effective in producing high-quality outputs.

We then conduct human evaluations of the model

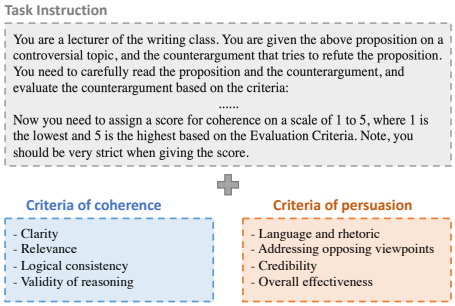

Figure 13: LLM-based automatic evaluation.

outputs using pairwise comparison. Specifically, we ask three human annotators to rate the overall quality of the outputs. Given an input, they are shown two outputs, with one generated by our model and one generated by a baseline method, presented in random order. They are asked to select which one is better, and ties are allowed if the outputs are not distinguishable. The results are summarized in Table 3. Our model results are considered as better with more times than both baselines, underscoring our model effectiveness even when operating under length constraints.

Discourse Diversity. We also analyze the output diversity by visualizing the RST trees of outputs. As shown in Figure 12, although E2E generates significantly longer outputs under length constraints, the distribution of RST tree depth is still less diverse compared to our model. This further demonstrate our model ability to produce outputs with greater diversity in rhetorical structure.

E Visualization of Common Words

We present the visualization top 50 most common verbs and nouns of our model-generated results with word cloud, as displayed in Figure 14. The larger word means the higher frequency. Overall, we can see that our model can generate quite diverse surface formats in the results. Notably, most nouns are policy-relevant, and this is because our dataset is in the politics and policy domains.

F Details for Human Evaluation

We present 30 random samples for human evaluation. To reduce bias, we anonymize the model names and present the outputs in a random order. We evaluate model outputs on the following aspects:

- **Appropriateness**: whether the content is expressed clearly, without ambiguity, vagueness, or grammatical errors; whether it has a good overall structure and strong readability, and the overall

Figure 14: The top 50 most common verbs and nouns in the arguments generated by our method.

Background:

Given a statement: [_proposition_]

We want to generate a counterargument to refute the statement.

Task:

Given a claim and a reasoning of the counterargument, we want to generate a short and brief concession to deal with potential dissenters and predict problems that might weaken the claim and reasoning.

Claim: [_claim_]

Reasoning: [_reason_]

Note: the goal of the concession is not to weaken the claim and reasoning, but to strengthen them by demonstrating that you have considered multiple perspectives and can respond to opposing viewpoints effectively. By acknowledging valid points from the opposing side, you build credibility and show that you are open to a fair and balanced discussion. A potential solution might be included in the concession. The concession should be in one sentence.

Concession:

Figure 15: Prompt for concession generation.

Task:

Given a proposition: _proposition_

We want to generate a coherent and persuasive counterargument to refute the proposition. You should first generate a brief plan, and then produce the counterargument based on the plan. The output should be in the format of:

Plan:the generated plan here

Counterargument: the generated counterargument here

Figure 16: Prompt for PlanCoT.

logic is smooth, consistent, and complete, with no internal contradictions or incoherence, and the main conclusion can be strongly supported by sub-arguments;

- **Content Richness:** whether the output is abundant, with sufficient points and evidence to effectively understand and refute the original input perspective; whether the expression is diverse, with varied diction and different forms of argumentation;

Proposition: [_proposition_]

Counterargument: [_argument_]

Task: Assume you are a professional writer. Given the statement and the counterargument aiming to refute the statement, please evaluate the counterargument based on the following aspects:

* Relevance: The counterargument should directly address the main claim or statement being challenged, rather than introducing tangential or irrelevant points;

* Correct Stance: The counterargument should have a different stance, in order to refute the given statement;

* Logical consistency: The counterargument should be logically consistent and not contain any contradictions or fallacies that weaken its credibility;

* Coherence of structure: The counterargument should have a clear and well-structured progression, with each idea logically flowing from the one before it;

* Persuasiveness: The counterargument should be strong enough to successfully challenge the original statement, which means it should be backed up by solid evidence, clear reasoning, and logical consistency.

Please return a one-paragraph suggestion on how to improve it based on the above criteria.

Suggestions:

Figure 17: Prompt for evaluation agent in the refinement module.

- **Overall Quality:** this is a general assessment that whether you think the output ranks higher than other candidates.

G Additional Sample Outputs

We present additional examples with different model outputs from Figure 20 to Figure 23.

Proposition: [_proposition_]

Counterargument: [_argument_]

Feedback: [_feedback_]

Task: Assume you are a professional writer. Given the statement, a counterargument on the opposing stance to refute the statement, and a feedback on how to improve the counterargument. Please revise the counterargument based on the feedback.

Revised counterargument:

Figure 18: Prompt for refinement agent in the refinement module.

Coherence:

* Clarity: The counterargument should be expressed clearly, with a well-defined structure that is easy to follow. Ambiguity or vagueness can detract from the argument's coherence;

* Relevance: The counterargument should directly address the proposition and stay focused on the topic. Irrelevant points or anecdotal evidence can detract from its coherence;

* Validity of reasoning: Evaluate the clarity and coherence of the counterargument's reasoning. Is the line of reasoning easy to follow? Does it present a clear cause-and-effect relationship or logical progression? A well-structured and coherent counterargument should present a logical flow of ideas;

* Logical consistency: Assess the counterargument for internal consistency. It should not contain any contradictory statements or logical fallacies that undermine its coherence. Look for logical connections and coherence between the counterargument's claims, evidence, and reasoning

Persuasion:

* Language and rhetoric: The counterargument should effectively use persuasive language and rhetoric techniques, such as appeals to logic, emotions, or ethics, to sway the reader's opinion;

* Addressing opposing viewpoints: A persuasive counterargument should acknowledge and address the main points of the given proposition, demonstrating an understanding of the opposing view and refuting it effectively;

* Credibility of evidence: The counterargument should be supported by credible evidence or sources. Unsupported claims or anecdotal evidence will not be as persuasive as a counterargument based on solid evidence;

* Overall effectiveness: The counterargument should effectively challenge the initial proposition and provide a convincing alternative viewpoint, and is likely to persuade the reader to reconsider their initial position.

Figure 19: Specific criteria for GPT4-based automatic evaluation on coherence and persuasion.

Proposition: The fact that voting isn't mandatory is a good thing.

E2E: Mandatory voting could help to increase voter turnout, which is important for achieving a fair and representative democracy. Low voter turnout can lead to a government that does not accurately reflect the interests of the people, and can result in policies that are not beneficial for all citizens. Mandatory voting could also help to ensure that all voices are heard and that all perspectives are taken into consideration.

PlanCoT: The fact that voting isn't mandatory may lead to a decrease in voter turnout, which can have negative effects on our democracy. Mandatory voting can ensure that all citizens take part in the democratic process, and that no one's opinion is silenced. Furthermore, the government has a responsibility to ensure that all citizens are informed about the importance of voting and the issues at stake, which can encourage more citizens to vote. Ultimately, voting is a crucial part of any democracy, and making it mandatory ensures that everyone has a say in the political process.

Ours$_{\text{w/o Refine}}$: The fact that voting isn't mandatory is not a good thing. Mandatory voting should be implemented because it is a civic duty that ensures that the people's voices are heard and that their opinions are represented in government. It is a fundamental part of democracy and is an essential way to ensure that citizens are engaged in the electoral process, their voices are taken into consideration when decisions are made, and that the government truly reflects the will of the people. Furthermore, mandatory voting can help to reduce the influence of money in politics by ensuring that all citizens are represented and can participate in the democratic process. While it is true that mandatory voting can ensure that the voices of all citizens are heard, it is important to consider how to make the process accessible and fair for those who are unable to vote due to physical or financial limitations. Ultimately, mandatory voting is a way to guarantee that all citizens have an equal say in the political process and that the government is truly representative of the people's wishes.

Ours: The fact that voting isn't mandatory is not a good thing. Mandatory voting should be implemented because it is a fundamental part of democracy and is an essential way to ensure that citizens are engaged in the electoral process and their voices are taken into consideration when decisions are made. By requiring citizens to vote, it can help to reduce the influence of money in politics, as more people are participating in the democratic process and their opinions are being represented. Additionally, mandatory voting can help to ensure that the government truly reflects the will of the people, as all citizens have an equal say in the political process.

However, it is important to consider how to make the process accessible and fair for those who are unable to vote due to physical or financial limitations. For example, measures such as providing transportation to and from polling stations, offering absentee ballots, and introducing flexible voting hours can help to make the process easier and more accessible for those who have difficulty voting in person. Additionally, lawmakers should focus on creating laws that promote fairness and accessibility for all citizens, regardless of their socio-economic status.

Overall, mandatory voting is a way to guarantee that all citizens have a voice in the political process and to ensure that the government is truly representative of the people's wishes. By implementing measures that make the voting process easier and more accessible for all citizens, mandatory voting can help ensure that everyone is given an equal chance to participate and have their voices heard.

Figure 20: Sample output of different model outputs.

Proposition: The increase in gender and sexuality vocabulary is not needed and ultimately hurts social progress on the topic.

E2E: The increase in gender and sexuality vocabulary is absolutely necessary and is an important step in helping to advance social progress on the topic. By expanding our language and terminology to accurately reflect the diversity of gender and sexuality, we are creating a more inclusive environment that allows people to identify and express themselves authentically. This in turn can lead to greater acceptance, understanding, and respect of all gender and sexuality identities. Additionally, the use of more specific language can lead to better and more effective conversation about the topic, allowing for deeper understanding and openness to different perspectives.

PlanCoT: The increase in gender and sexuality vocabulary is actually essential for social progress on the topic. It allows us to be more precise in our language, which in turn allows us to better understand each other and progress in our discussions. For example, the introduction of terms such as non-binary, gender fluid, and genderqueer has been crucial in allowing us to move away from the binary of male and female, and to recognize and appreciate the spectrum of gender identities and expressions. This has led to more inclusive and understanding conversations, and a greater respect for gender and sexual minorities.

Furthermore, the increased language can help to empower individuals, as it provides them with the words to express and identify their own gender and sexuality. This can be especially important for those who have felt invisible in the past, and it can create a sense of connection and understanding amongst members of the community.

Though there may be some drawbacks to increased language, such as confusion or misinterpretation, these drawbacks are far outweighed by the benefits. As language evolves, so do our understandings of gender and sexuality, and this is essential for making progress on the topic.

Ours$_{\text{w/o Refine}}$:
The given proposition that the increase in gender and sexuality vocabulary is not needed and ultimately hurts social progress on the topic is false. An increase in gender and sexuality vocabulary is crucial to promoting social progress on gender and sexuality topics, as it allows people to more accurately and effectively communicate their ideas and identities. By introducing new terms and expanding existing ones, individuals can more accurately and effectively communicate their thoughts, feelings, and identities. This helps to eliminate misunderstandings and create a more inclusive and open dialogue. For example, by introducing terms such as "non-binary" and "genderqueer", individuals who don't identify with traditional gender labels are able to express their gender identity in a more accurate way. This helps to create a more inclusive and accepting environment for people to discuss gender and sexuality issues. Furthermore, the use of new and more precise language allows individuals to better explain and articulate their thoughts and feelings, which can lead to more meaningful conversations and a better understanding of the issues at hand. By increasing gender and sexuality vocabulary, we can create a more open and accepting dialogue and move closer to achieving social progress on these issues.

However, it is also important to recognize that some people may find it difficult to understand and embrace these new terms. Thus, it is important to create accessible resources that explain these terms in a non-threatening and easy to understand way. This can help to ensure that individuals have the tools they need to understand and accept the new terms, and to participate in meaningful conversations about gender and sexuality issues.

In conclusion, an increase in gender and sexuality vocabulary is essential to fostering social progress on gender and sexuality topics. It is a critical tool for allowing individuals to more accurately and effectively communicate their ideas and identities, which can lead to more meaningful conversations and a better understanding of the issues at hand. Therefore, it is clear that the increase in gender and sexuality vocabulary is not only needed, but also beneficial in promoting social progress on the topic.

Ours: The given proposition that the increase in gender and sexuality vocabulary is not needed and ultimately hurts social progress on the topic is false. An increase in gender and sexuality vocabulary is essential to promoting social progress on gender and sexuality topics, as it allows people to more accurately and effectively communicate their ideas and identities. By introducing new terms and expanding existing ones, individuals can more accurately and effectively communicate their thoughts, feelings, and identities, and eliminate misunderstandings. This can lead to more meaningful conversations and a better understanding of the issues at hand. For example, by introducing terms such as "non-binary" and "genderqueer", individuals who don't identify with traditional gender labels are able to express their gender identity in a more accurate way. This helps to create a more inclusive and accepting environment for people to discuss gender and sexuality issues. Additionally, studies have found that the more precise language used to discuss gender and sexuality topics is associated with higher levels of acceptance of diverse gender identities and sexual orientations.

Furthermore, the use of new and more precise language allows individuals to better explain and articulate their thoughts and feelings, which can lead to more meaningful conversations and a better understanding of the issues at hand. By increasing gender and sexuality vocabulary, we can create a more open and accepting dialogue and move closer to achieving social progress on these issues. However, it is also important to recognize that some people may find it difficult to understand and embrace these new terms. Thus, it is important to create accessible resources that explain these terms in a non-threatening and easy to understand way, like glossaries or guides. This can help to ensure that individuals have the tools they need to understand and accept the new terms, and to participate in meaningful conversations about gender and sexuality issues.

In conclusion, an increase in gender and sexuality vocabulary is essential to fostering social progress on gender and sexuality topics. It is a critical tool for allowing individuals to more accurately and effectively communicate their ideas and identities, which can lead to more meaningful conversations and a better understanding of the issues at hand. Furthermore, the evidence shows that the more precise language used to discuss gender and sexuality topics is associated with higher levels of acceptance of diverse gender identities and sexual orientations. Therefore, it is clear that the increase in gender and sexuality vocabulary is not only needed, but also beneficial in promoting social progress on the topic.

Figure 21: Sample output of different model outputs.

Proposition: I don't think the duty of child raising should belong to the biological parents.

E2E: While it is true that in some cases, parents may not be the best suited to raise a child, it is important to remember that biological parents are often the most prepared to do so. They typically have a strong bond with their child due to their genetic connection and can provide the necessary emotional, physical, and financial support for their child's development and growth. Additionally, biological parents often have a better understanding of their child's specific needs and can make decisions that are in the best interest of their child. Thus, while it is important to consider alternative options for child-raising, it is generally in the best interest of the child to be raised by their biological parents.

PlanCoT:
Biological parents have a unique connection to their children that cannot be replicated by anyone else. They understand their children better and are better equipped to provide the emotional support and guidance needed to raise healthy, well-adjusted children. When the duty of child raising is not handled by the biological parents, there might be a lack of understanding and emotional connection, which can lead to the child feeling neglected or not understood. This, in turn, can have a detrimental effect on the child's development and well-being. Therefore, it is important that the duty of child raising still belongs to the biological parents, as they are in the best position to provide the love and support needed to raise a healthy and happy child.

Ours$_{\textbf{w/o Refine}}$: I strongly disagree with the proposition that the duty of child raising should not belong to the biological parents. The biological parents are uniquely suited to provide the best possible upbringing for their children, and as such, it is their responsibility and privilege to fulfill the duty of child raising.

The biological parents have a unique understanding of their children's physical and mental needs, which enables them to provide the best possible care. For example, parents will be able to recognize the signs of illness or distress in their children quicker than anyone else, and be able to provide the right type of care in a timely manner. Furthermore, parents are in the best position to provide the emotional and psychological support that children need to grow and develop. Parents will be able to tailor the type of guidance and discipline to the individual needs of their children, and provide an environment that will best nurture them and help them to reach their full potential. Thus, it is the unique privilege and responsibility of the biological parents to fulfill the duty of child raising.

That said, I acknowledge that there may be cases in which non-biological parents may provide similar or even better care for a child, and in those cases, it may be appropriate to entrust the duty of child raising to them. This does not, however, negate the fact that, in most cases, it is the biological parents who are best suited to provide the best possible upbringing for their children, and therefore it is their responsibility and privilege to fulfill the duty of child raising.

Ours: I strongly disagree with the proposition that the duty of child raising should not belong to the biological parents. The biological parents are uniquely suited to provide the best possible upbringing for their children, and as such, it is their responsibility and privilege to fulfill the duty of child raising. The biological parents have a unique understanding of their children's physical and mental needs that enables them to provide the best possible care. For example, parents are more likely to recognize signs of distress in their children quicker than anyone else due to their familiarity with their children's behaviors, and they can provide the right type of care in a timely manner. Furthermore, parents are in the best position to provide the emotional and psychological support that children need to grow and develop. As such, parents are able to tailor the type of guidance and discipline to the individual needs of their children, and provide an environment that will best nurture them and help them to reach their full potential.

To further illustrate this point, research has shown that children who grow up with both biological parents experience better physical and mental health outcomes, higher levels of educational achievement, and better economic success later in life. This is further proof of the unique role that biological parents have in providing the best possible upbringing for their children. Thus, it is the unique privilege and responsibility of the biological parents to fulfill the duty of child raising. That said, I acknowledge that there may be cases in which non-biological parents may provide similar or even better care for a child, and in those cases, it may be appropriate to entrust the duty of child raising to them. Nonetheless, in most cases, it is the biological parents who are best suited to provide the best possible upbringing for their children, and therefore it is their responsibility and privilege to fulfill the duty of child raising.

Figure 22: Sample output of different model outputs.

Proposition: There shouldn't be a death penalty for any crime but there should be a death option.

E2E: The death penalty should remain as a punishment option for some of the most serious crimes, as it serves as a deterrent to potential offenders and sends a clear message that certain actions will not be tolerated. By introducing the death option, we are sending the wrong message to would-be criminals, and it would encourage more people to commit serious crimes. Furthermore, the death option is not an effective way of achieving justice, as it does not allow for the possibility of rehabilitation or redemption for those who have committed a crime.

PlanCoT: The death penalty should not be an option for any crime. Firstly, it is uncertain whether it is an effective deterrent for crime. Studies have shown that the death penalty does not have a greater effect than other punishments, such as life imprisonment, in reducing crime rates. Secondly, there is a risk of wrongful executions, as mistakes can be made in the judicial process. Finally, while it may be argued that for some serious crimes imprisonment is not an appropriate punishment, attention should instead be directed to reforming the legal system to provide alternative forms of justice.

Ours$_{\textbf{w/o Refine}}$: The death penalty is an essential part of any justice system, and should remain as a punishment for serious crimes. There is no place for a 'death option' which would undermine the seriousness of serious criminal acts. The death penalty provides a clear and unambiguous deterrent for serious criminal acts, and ensures that those who commit such acts are held accountable for their actions. It also serves as a powerful symbol of the justice system's commitment to protecting society from the most serious offenders. Without the death penalty, there is a risk that offenders could escape appropriate punishment for their crimes. The death option proposed in the original statement would undermine the seriousness of serious criminal acts and create a loophole for those who commit them to escape punishment. It would also reduce the deterrent effect of the death penalty, as offenders would know that they could opt for the death option and avoid the full consequences of their actions. This could lead to more crime and a less safe society.

While we should consider alternative sentencing options that provide a fair punishment while also protecting society from further harm, the death penalty is an important part of any justice system. It is essential that those who commit serious crimes are held accountable for their actions, and the death penalty serves as a powerful signal of the justice system's commitment to upholding the law. There is no place for a 'death option' which would undermine the seriousness of serious criminal acts and create a loophole for those who commit them to escape punishment. Therefore, the death penalty should remain as a punishment for serious crimes, and any attempt to introduce a 'death option' should be strongly opposed.

Ours:The death penalty is an essential part of any justice system, and should remain as a punishment for serious crimes. It serves as a powerful symbol of the justice system's commitment to protecting society from the most serious offenders, and provides a clear and unambiguous deterrent for serious criminal acts. Furthermore, the death penalty ensures that those who commit such acts are held accountable for their actions. Introducing a 'death option' would undermine the seriousness of serious criminal acts and create a loophole for those who commit them to escape punishment. It would also reduce the deterrent effect of the death penalty, as offenders would know that they could opt for the death option and avoid the full consequences of their actions. This could lead to more crime and a less safe society.

Moreover, the death option proposed could also have a negative impact on society by providing less of an incentive for rehabilitation. Offenders may become less likely to seek rehabilitation if they know that they can opt for the death option and avoid the full consequences of their actions. This could lead to further criminal activity and a less safe society.

Therefore, while we should consider alternative sentencing options that provide a fair punishment while also protecting society from further harm, the death penalty is an important part of any justice system. It is essential that those who commit serious crimes are held accountable for their actions, and the death penalty serves as a powerful signal of the justice system's commitment to upholding the law. There is no place for a 'death option' which would undermine the seriousness of serious criminal acts and create a loophole for those who commit them to escape punishment. Therefore, the death penalty should remain as a punishment for serious crimes, and any attempt to introduce a 'death option' should be strongly opposed.

Figure 23: Sample output of different model outputs.