Plan Dynamically, Express Rhetorically: A Debate-Driven Rhetorical Framework for Argumentative Writing

Xueguan Zhao[1,3], **Wenpeng Lu**[1,3]*, **Chaoqun Zheng**[1,3], **Weiyu Zhang**[1,3], **Deyu Zhou**[2], **Jiasheng Si**[1,3]*

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing, China
[3]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China
`xueguan.zhao@foxmail.com, {wenpeng.lu, jiashengsi}@qlu.edu.cn`

Abstract

Argumentative essay generation (AEG) is a complex task that requires advanced semantic understanding, logical reasoning, and organized integration of perspectives. Despite showing a promising performance, current efforts often overlook the dynamical and hierarchical nature of structural argumentative planning, and struggle with flexible rhetorical expression, leading to limited argument divergence and rhetorical optimization. Inspired by human debate behavior and Bitzer's rhetorical situation theory, we propose a debate-driven rhetorical framework for argumentative writing. The uniqueness lies in three aspects: (1) it dynamically assesses the divergence of viewpoints and progressively reveals the hierarchical outline of arguments based on a depth-then-breadth paradigm, improving the *perspective divergence* within argumentation; (2) simulates human debate through iterative defender-attacker interactions, improving the *logical coherence* of arguments; (3) incorporates Bitzer's rhetorical situation theory to flexibly select appropriate rhetorical techniques, enabling the *rhetorical expression*. Experiments on four benchmarks validate that our approach significantly improves logical depth, argumentative diversity, and rhetorical persuasiveness over existing state-of-the-art models[1].

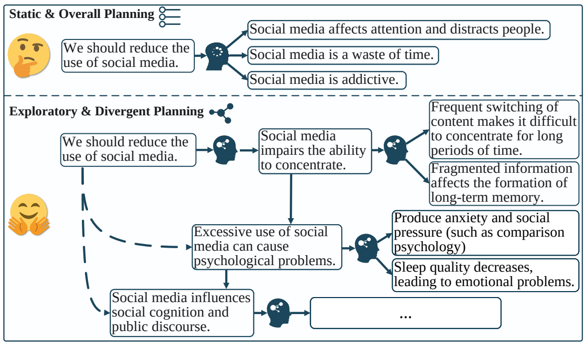

Figure 1: Comparison of two different plannings. **Head Icons** represent the two strategies—static (fixed structure) vs. divergent (dynamic, recursive thinking). **Solid arrows** show logical flow between arguments, while **dashed arrows** represent divergent arguments.

1 Introduction

Argumentative essay generation (AEG), literally the task of writing long-form perspective essays to express objective and diverse viewpoints or arguments on controversial issues, thus attempting to realize the persuasive intention (Bao et al., 2022). Due to its nature in high capability requirements of understanding social topics, integrating coherent arguments, and planning structural discourse, AEG poses significant challenges and attracts considerable attention in the NLP community (He et al., 2024; Xiao et al., 2024; Hu et al., 2025). In essence, AEG can be viewed as the instantiation of cognitive linguistics theory, offering a practical application of how argument and thought interact in the shape of persuasive and organized essays through rhetorical expression (Sarafyazd and Jazayeri, 2019).

Currently, multiple LLM-based works have been devised and exhibited notable performance through the *plan-then-write* paradigm (Yao et al., 2019; He et al., 2024; Liang et al., 2024b; Hu et al., 2025), involving structured skeleton planning and long-form content generation. Despite yielding promising performance by embedding ontologies of argumentative theory, existing works still face critical limitations. One crucial issue is that their reliance on *static and overall planning* strategy potentially raises a problematic trade-off. Which is, overemphasizing the "correct" logical rigor between central argument and evidence to align with predefined argumentative structure (Bao et al., 2022), while suppressing the dynamic divergence and derivability of the expressed opinion. This potentially re-

* Corresponding authors.

[1]Code and data are available at `https://github.com/zxg-x/DARE`

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9551–9573
November 4-9, 2025 ©2025 Association for Computational Linguistics

sults in the tendency to prioritize homogeneity in argumentation, risking issues of "perspective collapse"—biased towards the prominent and singular viewpoint. However, persuasive reasoning in cognitive linguistics is inherently dynamic and hierarchical, requiring not only logical rigor but also the depth and breadth of thoughtful viewpoints to engage a broad audience (Wu and Lytinen, 1990; Budzyńska and Kacprzak, 2008). For example in Fig.1, the central opinion presents sequentially divergence in its depth and breadth. Such a situation necessitates more exploratory and divergent thinking in argumentative planning.

Additionally, recent efforts indicate the importance of rhetorical language expression for persuasiveness, often by explicitly wording the "rhetoric" tokens in prompts (Xiao et al., 2024; Bao et al., 2022). However, this static requirement is unlike human writing, which typically involves strategic rhetorical flexibility in persuasive discourse. From the sociocultural and cognitive view, the writing context is never static (Cole et al., 1978; Prior, 2005; Ferretti and Graham, 2019). Writers possess a nuanced understanding of audience psychology, cultural social situation, and personal linguistic knowledge, to dynamically select the appropriate rhetorical strategies in shaping their writing. As indicated by Bitzer's rhetorical situation theory (Bitzer, 1968): rhetorical discourse is shaped by rhetorical choice and its context, guiding when and how rhetoric should be used, ensuring clarity, emotional resonance, and logical soundness.

In this paper, we reexamine the aforementioned issues and highlight three critical aspects in argumentative writing: **perspective divergence**, **logical consistency**, and **rhetorical expression**, aiming to *explore the influential factors in aligning the argumentative process with human behavioral patterns*. Specifically, *perspective divergence*: argumentative viewpoints should diverge from controversial topic, presenting analytic opinions in depth and breadth, forming a hierarchical and coherent structure (e.g., *argument→sub-arguments*); *logical consistency*: each perspective should be logically supported by evidence, warrant, or data, while mitigating the effect of impact of counterargument; *rhetorical expression*: argumentative expression should be flexible rhetoric by dynamically incorporating contextual factors (e.g., target audience and environment) for persuasiveness.

To this end, we propose a novel AEG framework termed DARE (dynamical planning and rhetorical expression), designed to write essays that are *perspective divergent*, *logically consistent*, and *rhetorically* under the plan-then-write paradigm. Specifically, we propose a *depth-then-breadth* strategy to construct an argumentative outline tree using a **host agent**, which posits how to organize the opinion into a hierarchical and branching structure. Each node indicates the argument or supporting evidence. Initiated by the conversational topic, this agent dynamically assesses opinion divergence and progressively decides whether to expand the outline tree vertically (depth) or horizontally (breadth) node-by-node, ensuring a well-structured and multi-perspective outline. Complementing this, inspired by prior multi-agent debate works (Tsao, 2023; Li et al., 2024), each node is associated with **debate agents** to adaptively engage in discussion, raise critique, and iteratively refine their arguments, thus fostering logical coherence through simulating the human debate process. Further, to enhance rhetorical efficacy, we integrate Bitzer's rhetorical situation theory—specifically its three elements of *Exigence*, *Audience*, and *Constraints*—to dynamically guide the selection of rhetorical technique, ensuring the alignment of essays with their intended audience.

Based on the extensive experiments across four English and Chinese AEG datasets, our framework reveals the superior performance over various LLM-based baselines in both automatic and human evaluations. Our analysis indicates that DARE produces argumentative essays with well-divergent depth and breadth, strong logical flow, and high persuasive impact. Additionally, our rhetorical distribution analysis reveals that the integration of Bitzer's rhetorical situation theory significantly enhances persuasive effectiveness, validating its value in rhetorical expression.

2 Related Works

Argument Mining Argument Mining (AM) aims to identify argumentative components and their logical relations from natural language texts. Early work relied on frameworks like the Toulmin model (Stab and Gurevych, 2014; Habernal and Gurevych, 2017)or Rhetorical Structure Theory(RST) (Das and Stede, 2018) with manual annotations. Later, neural models (e.g., RNNs, CNNs, BiLSTMs) enabled end-to-end approaches (Eger et al., 2017; Li et al., 2017; Cao, 2023) and subdivided more downstream tasks such as evidence retrieval (Hua and Wang, 2018) and argument quality assessment (Skitalinskaya et al., 2021).

Argument Generation Argument Generation (AG) extends Argument Mining by focusing on producing coherent and persuasive arguments for a given topic. Unlike AM, AG is more generative and linguistically complex. With the advancement of pre-trained models, AG has garnered significant attention, including areas such as counter-argument generation (Alshomary et al., 2021; Alshomary and Wachsmuth, 2023; Jo et al., 2021; Lin et al., 2023) and controlled argument generation (Schiller et al., 2021; Saha and Srihari, 2023).

Recent work has moved beyond generating isolated short arguments toward long-form argumentative text generation (Guan et al., 2021; Ji and Huang, 2021; Tan et al., 2021). This shift led to the Argumentative Essay Generation (AEG) task, which emphasizes not only high-quality arguments but also global coherence and structure. To support this task, Bao et al. (2022) released the ArgEssay dataset and proposed a dual-decoder Transformer model, while He et al. (2024) introduced argumentative planning for prompting LLMs. Nevertheless, existing AEG methods largely overlook the internal complexity of arguments and underexplore rhetorical strategies that strengthen persuasiveness and expressiveness.

Multi-Agent Debate Multi-agent debate has emerged as an effective approach to enhance the reasoning and generation abilities of large language models in tasks such as question-answering (Khan et al., 2024; Rasal, 2024; Wang et al., 2025) and collaborative problem-solving (Li et al., 2023; Tsao, 2023). Engaging multiple agents in argumentative dialogue has led to significant performance improvements. Zhou et al. (2024) shows that multi-turn interactions introduce complexity, inspired by human argumentation, allowing agents to explore, challenge, and refine arguments from diverse perspectives (Liang et al., 2024a). Building on this, our work integrates iterative multi-agent debate into argumentative essay generation. Unlike approaches that focus on consensus (Abdelnabi et al., 2024; Hu et al., 2025), our framework maintains adversarial dynamics to elicit counterarguments and uncover new perspectives essential for deep reasoning.

3 Methodology

The argumentative essay generation can be modeled as $p(y|x)$, where x is the prompt containing the conversational topic, y denotes the long-form essay presenting logically coherent and subjective arguments. As shown in Fig.2, we propose a novel AEG framework under the plan-then-write paradigm, which consists of two steps: (1) debate-guided outline tree planning: where the agent dynamically assesses the opinion divergence and iteratively expands the outline tree following the depth-then-breadth strategy, thus forming the hierarchical outline tree. In this process, a multi-agent debate is injected to collaboratively engage in debate to raise rebuttals for each argument, ensuring the logical coherence and forming the debate record; (2) rhetorical argumentative writing: which synthesizes the arguments by referring to the outline tree and incorporating Bizter's rhetorical situation theory. All of our prompts are listed in Appendix B.

3.1 Root Argument Initialization

To generate the outline that posits the hierarchical structure in an argumentation essay, the initial step is to determine the central opinion. To this end, we generate and assign a unique **Host** agent using CoT prompt (Wei et al., 2022), which serves as the manager in dynamically and progressively constructing the hierarchical outline tree $\mathcal{T}$ node-by-node. Initially, faced with a proposition on a controversial topic, the Host agent first analyzes the topic in prompt x to identify its scope, implicit assumptions, and core controversy. This guides the generation of the focused initial argument $core$, which serves as the root node in $\mathcal{T}$.

3.2 Debate-Guided Outline Tree Planning

In real-world scenarios, the arguments within high-quality argumentative essays present the divergence property, forming a hierarchical and layered structure in depth and breadth, where depth ensures a thoughtful logic about the opinion, and breadth reflects the diversity of perspective. Therefore, we propose the *depth-then-breadth* strategy to dynamically generate the outline tree, starting from the root node $core$. In addition, inspired by the recent works with multi-agent collaboration, we inject a persona-based multi-agent debate process to ensure the logical coherence for each node by introducing multi-turn interaction that rebuttal, defends, and refines arguments.

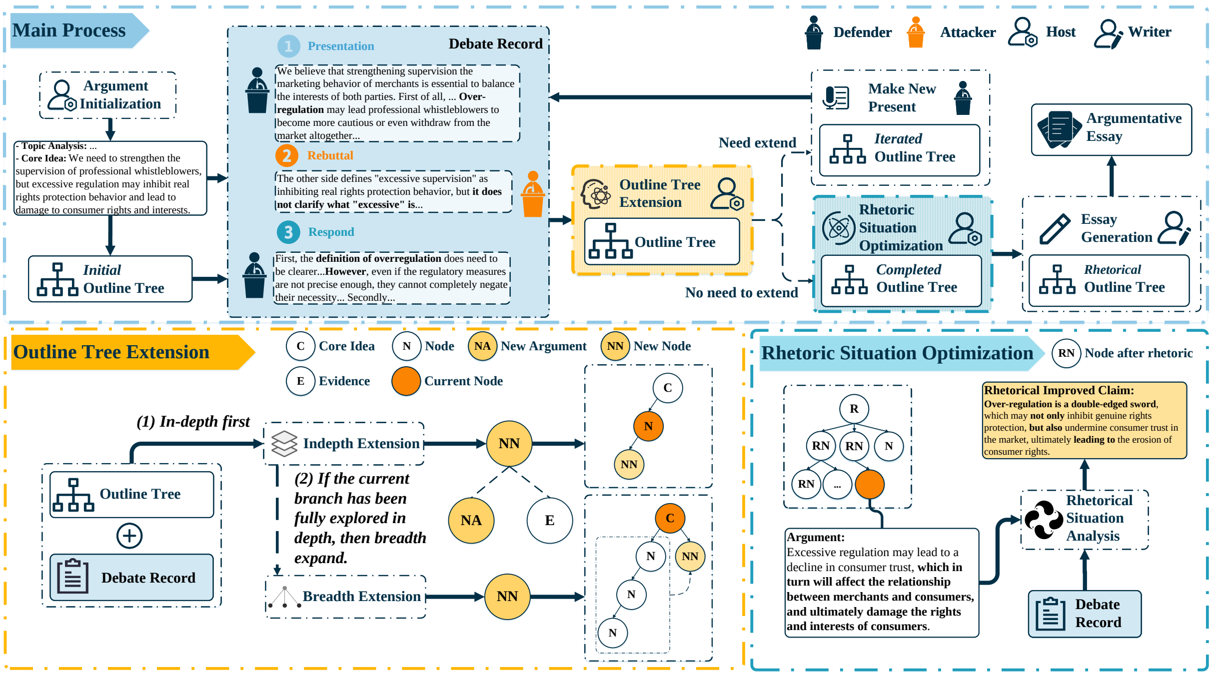

Figure 2: An overview of our framework. It builds an outline tree containing arguments and evidences based on a given position through iterative debates in cycles. Then, organize an argumentative essay based on the outline tree.

3.2.1 Multi-Agent Debate

This module aims to confirm and refine the arguments in $\mathcal{T}$ through simulating the **attacker-defender** debate process. Specifically, we assign a unique persona to each agent with a brief description, fixing their beliefs and forming the agents of **Defender** and **Attacker**.

Defender Based on the existing outline tree $\mathcal{T}_i$, root node $core$, and the topic in x, the **defender** needs to generate the argumentative passage p_i by considering their opinion for the i-th turn.

Attacker Refutation is crucial in critical thinking for making rigorous and convincing arguments (He et al., 2024; Toulmin, 2003). Based on the generated passage p_i, the topic, the **attacker** is instructed to challenge and criticize the argument of the **defender** through providing the counterargument, opposing view, and negative evidence, thus forming the rebuttal reb_i to negate the effectiveness of the **defender**.

Response Since the **attacker**'s rebuttal may not always be effective, or the existing outline tree $\mathcal{T}_i$ may be robust enough without this refutation, the **defender** agent is required to provide an analysis response $resp_i$ to rebut the **attacker**. This involves evaluating the validity of rebuttal reb_i, determining its reasonableness, and pointing out its flaws by integrating the information from the **defender**. Then, the **defender** agent further refines and reinforces their arguments with the **host** agent's subsequent analysis of the debate record. This step ensures the logical continuity and prevents topic drift, forcing the thorough examination before generating a new node.

Upon completing the multi-turn debate, the full debate record DR_i is formed by combining the **defender**'s argumentative passage, the **attacker**'s rebuttal, and the **defender**'s response. This record serves as the foundational material for expanding the outline tree and crafting the rhetorical argumentation.

3.2.2 Outline Tree Extension

Based on the divergent perspective during the debate process, to generate an outline tree of an argumentative essay to depict its overall structure, we propose the *depth-then-breadth* strategy to dynamically extend the outline tree node-by-node.

In-depth Extension The goal of in-depth extension is to strengthen the divergence of each argument in-depth by elaborating its reasoning, addressing objections, and reinforcing support, following a top-to-bottom way. Specifically, by summarizing the information of existing outline tree $\mathcal{T}_i$, the debate record DR_i, and the controversial topic, the **host** agent is instructed to recognize the unresolved

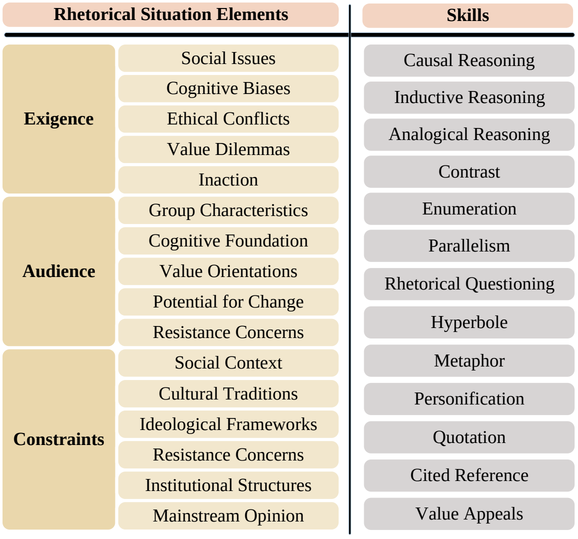

Figure 3: Factors and rhetorical devices to consider in rhetoric optimization.

assumptions around the parent node c_i, and further generate a new argument node c_{i+1}, supporting evidence e_i , and logical relation r_i. Newly generated argument c_{i+1}, evidence e_i will be added as extented nodes to the outline tree, and r_i will become an edge connecting c_i and c_{i+1}, depicting its support or inference relations. Details are in Appendix A.1.

Breadth Extension Once the **host** agent deems a reasoning branch sufficiently developed, it returns to the parent node of the other branch to initiate breadth extension in a bottom-to-top way. Unlike in-depth extension, which explores deep-branch viewpoints, the breadth extension introduces unexplored and diverse perspectives. In specific, By analyzing existing outline tree $\mathcal{T}_i$, the latest debate record DR_i, the argument of the parent c_i, and the controversial topic, the **host** agent is prompted to generates new arguments c_{i+1} from alternative angles, the evidence e_i, and logical relation r_i which is expanded through the same in-depth extension strategy. Details are in Appendix A.2. Until the **host** concludes that no additional elaboration is required, the outline tree will be considered complete. This potentially indicates that there is enough perspectives in crafting the essay and no need to be further divergent.

3.3 Rhetorical Argumentative Writing

To enhance the persuasion of argumentation, unlike prior works using argumentative theory such as Toulmin model (Bao et al., 2022), we introduce the theory of Bitzer's rhetorical situation (Bitzer, 1968) into the argumentative writing, enabling the dynamic selection of appropriate rhetorical strategies in the argumentation generation.

Rhetorical Situation Optimization Bitzer's rhetorical situation theory offers a scenario that each argument should be optimized by its context-specific situation defined by *Exigence* (issues or requirements in the situation), *Audience* (target objective), and *Constraints* (factors influential to speaker or audience), which flexibly guides the choice of rhetorical strategies. In this context, as shown in Fig.3 we summarize the relevant factors (e.g., Social Issues, Value Dilemmas) surrounding the three elements, by referring Mercier and Sperber (2011); Perloff (1993); Foss (2017). In addition, by referring to the rhetorical strategies in Liu et al. (2024), we selected 13 techniques that are particularly relevant to argumentative essay generation in Skills part. For each node in outline tree $\mathcal{T}$, we prompt the **host** agent to dynamically determine the appropriate rhetorical strategies of the argument by analyzing the debate record DR, the controversial topic, and the structured outline tree. Finally, each argument within outline tree $\mathcal{T}$ will be rhetorical optimized accordingly.

Essay Generation To generate a well-organized argumentative essay, based on all elements within the outline tree $\mathcal{T}$, we first instruct the LLM to summarize and synthesize into the introduction and conclusion of the argumentation, which are further transformed into a cohesive argumentative essay. We show the final output examples in Appendix C.

4 Experiments

4.1 Experimental Setting

Datasets. We evaluated our model on four datasets: ArgEssay(Bao et al., 2022), CHE-Essay, NYT-Editorial, and CHN-Editorial(He et al., 2024). These datasets cover two types of argumentative writing—exam essays and news editorials—and span both English and Chinese. ArgEssay and CHE-Essay represent exam-style writing from English proficiency tests and the Chinese Gaokao, respectively. NYT-Editorial and CHN-Editorial consist of professional editorials from the New York Times and Chinese news media. Further dataset details are provided in Appendix D.1.

Baselines. We compare our method with six strong baselines: (1) **DD-KW**: Dual-decoder Trans-

DataSet	Method	Automatic Evaluation						Human Evaluation				
		Relevance	Logic	Complexity	Persuasiveness	Rhetoric	Overall	Relevance	Logic	Complexity	Persuasiveness	Rhetoric
ArgEssay	DD-KW	2.26	1.796	1.62	1.612	1.66	1.83	2.3	2	1.65	1.7	1.5
	E2E	4.426	4.218	3.784	3.946	3.908	4.164	4.55	4.404	4.318	4.41	4.31
	CoT	4.5	4.142	3.69	3.872	3.786	4.036	4.56	4.384	4.286	4.38	4.276
	ToT	4.486	4.228	3.842	4.144	4.074	4.258	4.6	4.46	4.45	4.42	4.34
	DPE	4.488	4.15	4.132	4.148	3.98	4.19	4.486	4.326	4.424	4.377	4.386
	D2W	4.518	4.14	3.986	3.876	4.056	4.122	4.5	4.16	4.13	3.92	4.05
	our DARE	**4.6306**	**4.3408**	**4.3286**	**4.3**	**4.249**	**4.3531**	**4.64**	**4.465**	**4.517**	**4.496**	**4.458**
NYT-Editorial	DD-KW	2.184	1.764	1.742	1.618	1.558	1.798	2.21	1.83	1.73	1.645	1.43
	E2E	4.522	4.222	3.834	3.938	3.93	4.148	4.24	3.934	3.786	3.806	3.86
	CoT	4.68	4.246	3.762	3.93	3.808	4.136	4.14	3.94	3.776	3.791	3.775
	ToT	4.624	4.232	3.888	4.034	3.982	4.2	4.251	4.022	3.882	3.928	3.897
	DPE	4.612	3.914	4.036	3.872	3.722	3.958	4.414	4.199	4.107	4.092	4.002
	D2W	4.588	4.12	3.908	3.916	4.066	4.124	4.52	4.131	4.12	4.02	4.1
	our DARE	**4.674**	**4.252**	**4.2**	**4.1**	**4.1**	**4.258**	**4.548**	**4.328**	**4.418**	**4.21**	**4.316**
CHE-Essay	DD-KW	2.176	1.706	1.614	1.486	1.688	1.774	2.23	1.81	1.642	1.53	1.74
	E2E	4.69	4.308	3.822	3.96	4.05	4.194	4.196	4.006	3.844	3.804	3.838
	CoT	4.7	4.244	3.696	3.882	3.924	4.128	4.191	3.979	3.79	3.745	3.807
	ToT	4.686	4.244	3.772	4	3.966	4.168	4.256	4.096	3.953	3.94	3.924
	DPE	4.666	4.256	4.104	4.106	4.042	4.253	4.35	4.265	4.201	4.254	4.002
	D2W	4.604	4.232	3.968	3.962	3.79	4.118	4.34	4.252	4.186	4.23	3.95
	our DARE	**4.716**	**4.27**	**4.146**	**4.228**	**4.232**	**4.31**	**4.376**	**4.322**	**4.322**	**4.354**	**4.242**
CHE-Editorial	DD-KW	2.144	1.678	1.728	1.55	1.63	1.77	2.12	1.712	1.81	1.53	1.64
	E2E	**4.804**	**4.348**	3.798	3.98	3.8	4.18	4.337	4.081	3.943	3.918	3.993
	CoT	4.764	4.222	3.678	3.866	3.692	4.086	4.328	4.059	3.890	3.893	3.893
	ToT	4.74	4.298	3.766	3.998	3.822	4.158	4.428	4.121	4.012	4.024	4.046
	DPE	4.708	4.24	4.136	4.114	3.848	4.234	4.584	4.331	4.328	4.24	4.231
	D2W	4.7	4.282	4.056	3.954	3.848	4.1892	4.53	4.36	4.284	4.03	4.26
	our DARE	4.774	4.346	**4.148**	**4.28**	**4.188**	**4.35**	**4.687**	**4.393**	**4.478**	**4.375**	**4.462**

Table 1: The results of the comparison of baselines on automatic and human evaluation metrics. **Bold** numbers denote the best performance among all methods on each dataset.

former with explicit content planning (Bao et al., 2022). (2) **E2E**: Direct essay generation via LLM prompts. (3) **CoT**: Generates a brief plan before writing (Wei et al., 2022). (4) **ToT**: Produces and selects from multiple plans before writing (Yao et al., 2023). (5) **DPE**: Two-stage planning with critical self-reflection (He et al., 2024). (6) **D2W**: Simulates agent debates for collaborative essay generation (Hu et al., 2025). Further details of all baselines are provided in the Appendix D.2.

Implementation Details. We conducted experiments using Llama-3.1-8B-Instruct, bart-base, Qwen2.5-7B-Instruct and bart-base-chinese. Both LLM models were deployed using VLLM to ensure efficient execution. The temperature for both models was set to 0.8 to strike a balance between creativity and coherence in the generated outputs. In addition, we set the maximum depth of the tree to 3 and the maximum width of each node to 2. More implementation details are in Appendix D.3.

4.2 Evaluation Metrics

Automatic Evaluation. Evaluating argumentative essays automatically is challenging, as traditional metrics like BLEU fail to capture depth, logical coherence, and overall quality (Celikyilmaz et al., 2020). Following recent LLM-based evaluation works (Wang et al., 2023; Fagbohun et al., 2024; Xiao et al., 2024; He et al., 2024), we adopt a GPT-4o-based evaluation framework for more consistent, human-aligned assessments. Essays are evaluated across five dimensions: (1) *Relevance*: Topical alignment and internal consistency; (2) *Logical Coherence*: Structural clarity, reasoning quality, and evidential support; (3) *Argument Complexity*: Depth of reasoning and engagement with diverse perspectives; (4) *Persuasiveness*: Effectiveness of argument across contexts; (5) *Rhetoric*: Clarity, precision, and rhetorical strength. A final *Holistic Score* aggregates these dimensions to reflect overall essay quality. To enhance reliability, GPT-4o generates detailed feedback before scoring, improving alignment with human judgments. All evaluations are repeated three times and averaged. Prompts are included in Appendix E.1.

Human Evaluation. For a more comprehensive analysis, we conduct a human evaluation to complement the automatic assessment. Three well-educated master's students with backgrounds in linguistics or related fields are recruited to independently score the generated essays. They follow the same five evaluation dimensions used in the automatic evaluation, excluding the overall score. Each evaluator is provided with detailed scoring guidelines to ensure consistency and reliability. More details are in Appendix E.2.

Model	Relevance	Logic	Complexity	Persuasiveness	Rhetoric	Overall
our DARE	**4.6306**	**4.3408**	**4.3286**	**4.3**	**4.249**	**4.3531**
w/o DG	4.612	4.2	4.2061	4.182	4.282	4.268
w/o DR	4.538	4.322	4.22	4.228	4.326	4.294
w/o WR	4.622	4.32	4.268	4.27	4.06	4.31
w/o DG+DR	4.604	4.18	4.16	4.1	4.136	4.162

Table 2: Ablation study result.

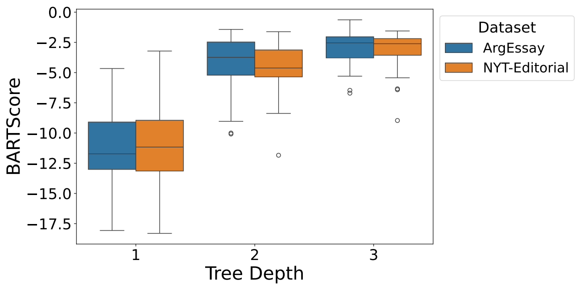

Figure 4: BARTScore of parent and child claim with argument depth in English datasets.

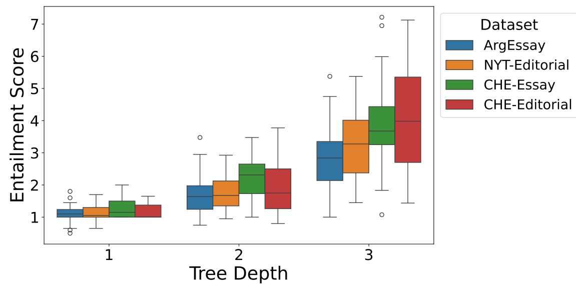

Figure 6: Variation of NLI non-contradiction score with argument depth.

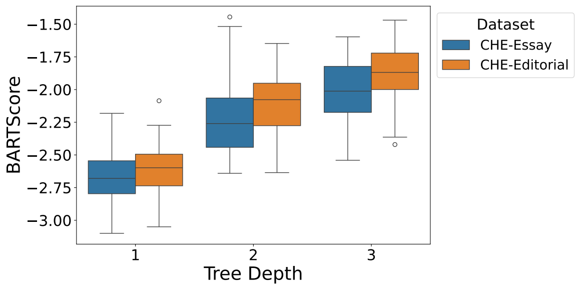

Figure 5: BARTScore of parent and child claim with argument depth in Chinese datasets.

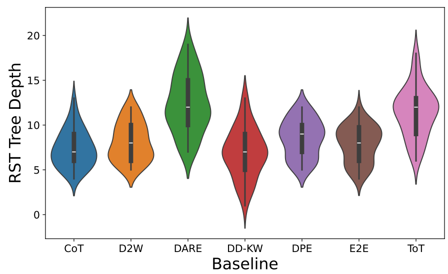

Figure 7: Distribution of RST tree depth of argumentative essays.

5 Results and Analysis

5.1 Main Results

Performance on Automatic Evaluation. As shown in Table 1, DARE achieves the highest or near-highest scores across all five metrics and four datasets. Compared to strong baselines such as ToT, D2W, and DPE, our method exhibits notable improvements, especially in *Logic*, *Argument Complexity* and *Rhetoric*, indicating the effectiveness of our tree-based argumentative reasoning in structuring and deepening content. The improvements on Rhetoric further demonstrate the benefit of incorporating rhetorical situation in generation. Notably, on the CHE-Essay dataset, DARE improves Persuasiveness by +0.12 over DPE and +0.228 over ToT, reflecting its superior capability in constructing convincing arguments.

Performance on Human Evaluation. Human evaluation results align with the automatic metrics, confirming DARE's superiority in generating high-quality arguments. It outperforms all baselines across every human-assessed dimension. Notably, DARE shows significant gains in Rhetoric and Persuasiveness, with average improvements exceeding +0.2 across datasets, underscoring the effectiveness of our rhetorical optimization strategies. It also maintains strong coherence and logical consistency, achieving the highest Logic scores on all datasets. These results indicate that DARE enhances not only surface-level fluency but also the deeper argumentative quality valued by human evaluators.

5.2 The Effect of Outline Tree Depth

To evaluate outline tree depth and progression, we use BARTScore(Yuan et al., 2021) to measure the probability of generating a sub-argument from its parent, and NLI-based non-contradiction probabil-

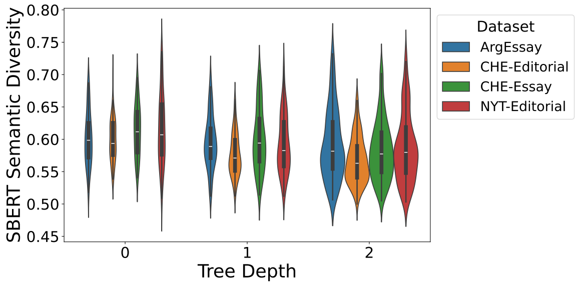

Figure 8: Semantic score of parent and child claim at each layer.

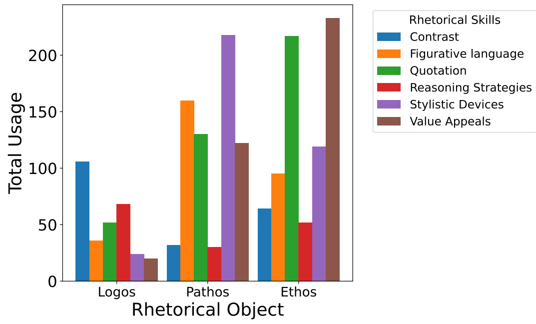

Figure 9: Rhetoric distribution analysis results.

ity(Dušek and Kasner, 2020; Nie et al., 2020) to assess logical support. The results visualized in box plots by tree depth, show consistent upward trends. As shown in Fig.4 and Fig.5, higher BARTScores at greater depths indicate improved semantic coherence, while rising non-contradiction scores in Fig.6 suggest increasingly consistent logical reasoning, reflecting the deeper support relation between different arguments.

5.3 The Effect of Outline Tree Breadth

To assess the breadth and diversity of viewpoints in the generated essays, we parsed all generated essays into Rhetorical Structure Theory (RST) trees. A higher number of arguments (EDUs) and more complex relationships result in deeper RST trees (Stede, 2016). As shown in Fig.7, DARE outputs have the most EDUs and the deepest structure, indicating richer content and more varied discourse relations. In contrast, outputs from other models are generally shallower and more linear.

In addition, we further analyzed the balance trade-off to reveal how structural depth affects argument diversity, by computing the semantic similarity among sibling nodes at each depth. As shown in Fig.8, diversity remains stable across depths in both English and Chinese datasets. In English datasets, semantic diversity scores range from 0.60 to 0.75. Similar trends are observed in Chinese datasets. Based on the observed results, this consistency suggests our framework preserves content diversity as depth increases, avoiding redundancy and maintaining engagement across deeper argument layers.

5.4 Manual Analysis of Rhetoric

To examine the distribution of rhetorical strategies, we conducted a manual analysis on 30 randomly sampled essays from each of the four datasets. Related rhetorical skills were grouped into six categories: *Contrast*, *Value Appeals*, *Figurative Language* (e.g., metaphor), *Quotation* (e.g., citation, enumeration), *Reasoning Strategies* (e.g., causal, analogical), and *Stylistic Devices* (e.g., parallelism, rhetorical questioning). We manually analyzed and counted the use of each category in improving the logic (*Logos*), emotional resonance (*Pathos*), and credibility (*Ethos*) of argumentations.

As shown in Fig.9, *Logos* are used least frequently overall. Among them, *Contrast* and *Reasoning Strategies* dominate, reflecting an emphasis on logical clarity and argumentative rigor. In contrast, rhetorical appeals to *Pathos* and *Ethos* are more prevalent. For *Pathos*, *Stylistic Devices* account for the largest share, aiming to evoke emotional resonance through expressive language. For *Ethos*, *Value Appeals*, and *Quotation* are most prominent, helping establish credibility through shared values and authoritative references.

5.5 Ablation Study

We evaluated 50 samples to assess the contribution of each module in our framework. The results are presented in Table 2. Directly prompted to generate sub-arguments without debate (*w/o* DG), we observe that the *Logic*, *Argument Complexity*, and *Persuasiveness* are significantly reduced. Notably, directly prompting for rhetorical optimization without considering the rhetorical situation (*w/o* DR) yields the best performance in *Rhetoric*, but underperforms DARE on other metrics. Manual analysis reveals that this method often overuses rhetorical techniques, undermining the rigor and effectiveness of argumentative writing. After removing DG and DR at the same time(*w/o* DG+DR), the performance of all aspects has dropped significantly. We also removed rhetorical optimization entirely (*w/o* WR); while this led to the poorest performance in *Rhetoric*, the overall performance was second only to DARE. This suggests that appropriate use of

{Parent Claim}: Risk-taking promotes personal growth.

{Defender}: Risk-taking allows individuals to step outside their comfort zones and gain new experiences...

{Attacker}: But not all risk-taking is beneficial—look at how governments encourage risky investment behaviors that led to financial crises. These risks don't promote growth; they create harm.

{Sub-claim proposed by the host agent after debate}: Government-endorsed financial risk-taking often leads to large-scale economic harm rather than personal development.

Table 3: An example of debates caused topic drift.

rhetoric enhances overall quality, whereas excessive use diminishes argumentative strength.

6 Error Analysis

6.1 Topic Drift Induced by Debates

We observed that even with explicit prompts, topic drift may still occur, especially when agents introduce arguments from adjacent but semantically distinct domains. For example, in Table 3, while the sub-claim is logically coherent, it deviates from the parent claim's focus on individual psychological growth. The debate had subtly shifted the frame from personal development to macroeconomic policy ethics, resulting in a semantic mismatch between the claim layers.

6.2 Rhetorical Strategies Undermining Coherence

In some cases, the rhetorical optimization process unintentionally disrupts the logical flow between parent and sub-claims, making it harder to preserve structural coherence in the final argument. For example, in Table 4, while our method explicitly models inter-claim logical relations, highly abstract or figurative rewrites can obscure these relations in the final essay. As a result, even with correct relation labels, the generated text may present logically related claims as disjointed.

7 Conclusion

In this paper, we propose DARE, a debate-driven framework for argumentative writing that integrates perspective divergence, logical coherence, and rhetorical expression. By simulating human debate, DARE iteratively expands outline trees to enrich both depth and breadth of perspectives, while rhetorical situation theory guides context-sensitive optimization. Experiments show that DARE significantly enhances the logical, rhetorical, and overall quality of generated arguments.

{Original Parent Claim}: We should reduce the use of social media to improve our attention span.

{Original Sub-Claim}: Social media platforms are designed to encourage short-term engagement, which weakens our ability to focus on long tasks."

{Relation}: This sub-argument is derived from the causal reasoning chain: 'from platform design → short-term engagement → reduced attention' of the parent argument.

{Optimized Parent Claim}: We should reduce the use of social media, the fast food of the mind, to protect our cognitive health.

{Optimized Sub-Claim}: Like junk food that trains our bodies to crave sugar, social media rewires our brains to crave novelty and distraction.

{Paragraph in final essay}: In an age of digital saturation, social media has become the fast food of the mind — quick, addictive, and cognitively corrosive. Just as we regulate our diets to protect our physical health, we must also regulate our information consumption to safeguard our cognitive well-being. Like junk food that conditions our bodies to crave sugar, these platforms rewire our neural circuits to seek novelty and instant rewards, eroding our capacity for deep, sustained focus.

Table 4: An example of rhetoric hurt coherence.

Limitations

A key challenge of our framework lies in helping LLMs capture the hierarchical structure of the outline tree. Although sub-claims are iteratively generated in debates, logical relations remain only textually annotated. More effective ways to implicitly convey hierarchy—via discourse markers, rhetorical cues, or prompting—are needed. Moreover, as outline depth and breadth grow, arguments become longer and harder to process. Handling such long contexts while preserving core logic and rhetorical coherence may require hierarchical summarization, selective filtering, or memory-augmented mechanisms.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.62402258, No.62376130, No.62502249), Taishan Scholars Program of Shandong Province (No.tsqn202507242), Shandong Provincial Natural Science Foundation (No.ZR2024QF099, No.ZR2024QF054), Program of New Twenty Policies for Universities of Jinan (No.202333008), Program of Innovation Improvement for Small and Medium-sized Enterprises of Shandong (No.2024TSGC0039), and the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2024ZDZX08).

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-Argument Generation by Attacking Weak Premises. In *Proceedings of the 59th Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1816–1827.

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 (EACL)*, pages 957–967.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A Neural Transition-Based Model for Argumentation Mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6354–6364.

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 27th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5134–5148.

Lloyd F Bitzer. 1968. The Rhetorical Situation. *Philosophy & Rhetoric*, 1(1):1–14.

Katarzyna Budzyńska and Magdalena Kacprzak. 2008. A Logic for Reasoning about Persuasion. *Fundamenta Informaticae*, 85(1-4):51–65.

Lang Cao. 2023. AutoAM: An End-To-End Neural Model for Automatic and Universal Argument Mining. In *Proceedings of the 19th International Conference on Advanced Data Mining and Applications (ADMA)*, pages 517–531.

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

Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. 1978. Mind in Society. *Mind in society the development of higher psychological processes. Cambridge, MA: Harvard University Press*.

Debopam Das and Manfred Stede. 2018. Developing the Bangla Rst Discourse Treebank. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (ICLR)*.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG)*, pages 131–137.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11–22.

O Fagbohun, NP Iduwe, M Abdullahi, A Ifaturoti, and OM Nwanna. 2024. Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *Journal of Artifical Intelligence and Machine Learning & Data Science*, 2(1):1–8.

Ralph P Ferretti and Steve Graham. 2019. Argumentative Writing: Theory, Assessment, and Instruction. *Reading and Writing*, 32:1345–1357.

Sonja K Foss. 2017. *Rhetorical Criticism: Exploration and Practice*. Waveland Press.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long Text Generation by Modeling Sentence-Level and Discourse-Level Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6379–6393.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Yuhang He, Jianzhu Bao, Yang Sun, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2024. Decomposing Argumentative Essay Generation Via Dialectical Planning of Complex Reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 12305–12322.

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 (COLING)*, pages 4689–4703.

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 (ACL)*, pages 219–230.

Haozhe Ji and Minlie Huang. 2021. DiscoDVT: Generating Long Text with Discourse-Aware Discrete Variational Transformer. In *Proceedings of the 21th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4208–4224.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-Enhanced Evidence Retrieval for Counterargument Generation. In *Proceedings of 21th Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 3074–3094.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 23662–23733.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can LLMs Speak For Diverse People? Tuning LLMs via Debate to Generate Controllable Controversial Statements. In *Proceedings of the 62nd Association for Computational Linguistics (ACL)*, pages 16160–16176.

Minglan Li, Yang Gao, Hui Wen, Yang Du, Haijing Liu, and Hao Wang. 2017. Joint RNN model for argument component boundary detection. In *Proceedings of the 46th IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 57–62. IEEE.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024a. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17889–17904.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and 1 others. 2024b. Integrating Planning into Single-Turn Long-Form Text Generation. *arXiv preprint arXiv:2410.06203*.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuan-Jing Huang, and Zhongyu Wei. 2023. Argue with Me Tersely: Towards Sentence-Level Counter-Argument Generation. In *Proceedings of the 23th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16705–16720.

Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, and Yan Xia. 2024. CERD: A Comprehensive Chinese Rhetoric Dataset for Rhetorical Understanding and Generation in Essays. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6744–6759.

Hugo Mercier and Dan Sperber. 2011. Why Do Humans Reason? Arguments for An Argumentative Theory. *Behavioral and brain sciences*, 34(2):57–74.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4885–4901.

Richard M Perloff. 1993. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. Routledge.

Paul A Prior. 2005. A Sociocultural Theory of Writing. In *The Handbook of Writing Research*, pages 54–66. Guilford Press.

Sumedh Rasal. 2024. Llm Harmony: Multi-Agent Communication for Problem Solving. *arXiv preprint arXiv:2401.01312*.

Sougata Saha and Rohini K Srihari. 2023. ArgU: A Controllable Factual Argument Generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8373–8388.

Morteza Sarafyazd and Mehrdad Jazayeri. 2019. Hierarchical Reasoning by Neural Circuits in the Frontal Cortex. *Science*, 364(6441):eaav8911.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-Controlled Neural Argument Generation. In *Proceedings of the 12th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 380–396.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, pages 1718–1729.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical papers (COLING)*, pages 1501–1510.

Manfred Stede. 2016. Towards Assessing Depth of Argumentation. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING)*, pages 3308–3317.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive Generation of Long Text with Pretrained Language Models. In *Proceedings of the 18th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 4313–4324.

Stephen E Toulmin. 2003. *The Uses of Argument*. Cambridge university press.

Wen-Kwang Tsao. 2023. Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning. In *Proceedings of the 10th International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 365–370.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*. Association for Computational Linguistics.

Rui Wang, Yonghe Chen, Weiyu Zhang, Jiasheng Si, Hongjiao Guan, Xueping Peng, and Wenpeng Lu. 2025. MedConMA: A Confidence-Driven Multi-agent Framework for Medical Q&A. In *Proceedings of the 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 421–433. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Horng Jyh P Wu and Steven L Lytinen. 1990. Coherence Relation Reasoning in persuasive discourse. In *Proceedings of the 12nd Annual Meeting of the Cognitive Science Society (CogSci)*, volume 12.

Ruiyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, and Ting Liu. 2024. Prove Your Point!: Bringing Proof-Enhancement Principles to Argumentative Essay Generation. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18995–19008.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-And-Write: Towards Better Automatic Storytelling. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 7378–7385.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating Generated Text as Text Generation. *Advances in neural information processing systems*, 34:27263–27277.

Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. 2024. Speak Out of Turn: Safety Vulnerability of Large Language Models in Multi-Turn Dialogue. *arXiv preprint arXiv:2402.17262*.

A Examples of extension

We follow the depth-then-breadth strategy in our outline tree generation, here is an elaboration on "Outline Tree Extension" with examples.

A.1 In-depth Extension

It starts from a **Parent-Claim** and generates more specific and targeted sub-arguments divergently, by referring to the generated outline and Debate Record. These sub-arguments directly support its parent argument, digging deeper into its logic. As shown in Table 5, the New Node (NN) is more concrete and supports its parent argument (i.e., Current Node (N)), while surrounding the Core Idea (C).

A.2 Breadth Extension

When a branch is fully explored in its depth, we extend a new argument branch that differs from existing ones, still rooted in the **Parent-Claim**. As shown in Table 6, breadth extension might extend a new branch by linking a New Node (NN) which broadens the scope of discussion and provides diverse opinion perspective.

A.3 Relation

As shown in Table 7, the form of relation is expressed in natural language (e.g., “[Sub-argument] supports [Parent-argument] due to. . . "), explaining the logical relations that why the sub-argument can support its parent-argument. Compared to the explicit markers (e.g., but, because), this way would efficiently help LLMs to understand the logical coherence between different arguments.

{Core Idea (C)}: Online education improves learning efficiency.
{Current Node (N)}: Flexible scheduling benefits learners.
{New Node (NN)}: Flexible scheduling helps working adults arrange study time around job duties, reducing dropout risks.

Table 5: An example of in-depth extension.

{Core Idea (C)}: Online education improves learning efficiency.
{Current Node (N)}: Flexible scheduling benefits learners.
{Existing node (N)}: Flexible scheduling helps working adults arrange study time around job duties, reducing dropout risks.
{New Node (NN)}: Interactive virtual classrooms.

Table 6: An example of breadth extension.

B Prompt

Table 9 presents different system prompts tailored to various roles based on the overall goal of argumentative writing and the specific task of debate. The prompt for generating the core of the essay is shown in Table 10. The prompts for generating *Present*, *Rebuttal*, and *Respond* are detailed in Tables 11, 12, and 13, respectively. The generation of arguments and their rhetorical optimization are addressed in Tables 14, 15, and 16. The introduction and conclusion prompts are found in Table 17, and the final essay generation is covered in Table 18.

{Parent-Claim}: Risk-taking is an essential component of human development and can lead to significant personal growth and self-improvement, outweighing potential risks and negative consequences.
{Sub-Claim}: The crucial concept of "controlled risk-taking." By recognizing the importance of controlled risk-taking and bounded rationality, we can harness the power of risk-taking to drive personal growth while minimizing the risks of failure and negative consequences.
{Relation}: This sub-argument challenges the root cause of the parent argument by highlighting the limitations of risk-taking and the need for careful consideration of risks and consequences.

Table 7: An example of relation between parent-argument and sub-argument.

C Case Study

We show an example of rhetoricization optimization in Table 20 and an example of an outline tree in Table 21. In addition, we present examples of the final generated English and Chinese argumentative essays in Table 22 and Table 23. Given the considerable length of complete argumentative essays, we have made additional supplementary materials publicly accessible for further reference. These materials include more generated essay examples as well as output results from other baseline methods. All these resources are available in our GitHub repository.

D Experimental Details

D.1 Datasets

We evaluated our model using four datasets: ArgEssay (Bao et al., 2021), CHE-Essay, NYT-Editorial, and CHN-Editorial (He et al., 2024). These datasets cover two types of argumentative essay scenarios—exam essays and news editorials—and include both English and Chinese content.

ArgEssay This dataset contains 11,282 writing topic-argumentative text pairs, sourced from international standardized English writing tests such as IELTS and TOEFL. It covers a wide range of con-

Samlpe Dataset	Relevance	Logic	Complexity of Claims	Persuasiveness	Rhetoric
ArgEssay	0.8332	0.8099	0.8185	0.8337	0.8345
NYT-Editorial	0.6477	0.7700	0.8519	0.8037	0.8205
CHE-Essay	0.8399	0.775	0.8418	0.8628	0.7966
CHE-Editorial	0.7764	0.8071	0.8708	0.8970	0.8538

Table 8: Inter-annotator agreement for human evaluation.

Host:	You are an experienced argumentative essay editor, skilled at uncovering arguments based on established positions and claims. You have been invited to a debate competition as the backup for the affirmative team.
Debater:	You are an experienced debater.
Writer:	You are an experienced argumentative essay writer, skilled at organizing multi-layered arguments using the Toulmin model of argumentation and combining rhetorical techniques to craft argumentative essays.

Table 9: Prompt of different agent identity background.

troversial topics, including technological advances, educational approaches, and environmental issues.

NYT-Editorial The NYT-Editorial dataset includes 9,178 topic-argumentative text pairs from the "Room for Debate" section of the New York Times. Each prompt describes a current event or social issue, followed by argumentative perspectives from professional writers. This dataset provides valuable insights into diverse argumentative approaches in editorial writing.

CHE-Essay The CHE-Essay dataset consists of 3,750 prompt-essay pairs from the Chinese national college entrance examination (Gaokao). The data is sourced from both real and mock exams available on doc-sharing websites, offering a unique perspective on argumentative writing within the context of Chinese education.

CHN-Editorial The CHN-Editorial dataset contains 2,998 argumentative essays compiled from the editorial sections of two Chinese news websites: PengPai and GuangMingWang. It provides a rich resource for analyzing argumentative writing within Chinese media discourse, covering a variety of topics related to current events.

These four datasets provide a comprehensive foundation for evaluating argumentative essay generation models in both exam-based and editorial contexts, with coverage across both English and Chinese.

D.2 Baselines

DD-KW DD-KW employs a dual-decoder Transformer architecture tailored for long-form argumentative writing. The model employs two decoders: a planning decoder (PD) to generate an explicit content plan in the form of structured sequences, and a writing decoder (WD) that conditions on the generated plan to produce the final essay.

E2E E2E represents a direct prompting approach, where a large language model (LLM) generates an argumentative essay in a single step conditioned solely on the given topic. Unlike planning-based methods, E2E does not incorporate explicit intermediate structures or reasoning strategies, serving as a straightforward end-to-end generation paradigm.

CoT CoT follows the Chain-of-Thought (CoT) (Wei et al., 2022) prompting paradigm, where the model first produces a brief reasoning or planning sequence before generating the essay. The intermediate reasoning acts as an implicit content plan that guides the subsequent writing process, often improving coherence and logical progression compared to direct end-to-end generation.

ToT ToT extends the CoT paradigm into a more structured search framework, known as Tree of Thoughts (ToT) (Yao et al., 2023). Instead of relying on a single linear reasoning path, ToT explores multiple possible reasoning branches (i.e., "thoughts") and performs self-evaluation, look ahead, and backtracking to select the most promising path before proceeding. Originally proposed for problem-solving tasks such as arithmetic puzzles and creative writing, ToT provides a stronger deliberation mechanism that can be adapted for argumentative essay generation by allowing the model to generate, evaluate, and refine multiple essay plans before writing.

Now you need to designate an initial debate direction for your team based on the given topic: {topic}. Your tasks are as follows: 1. Analyze the given topic to identify its core discussion issues. 2. Establish the core argument for your team. Please strictly follow the output format below and do not add any irrelevant content, symbols, or punctuation marks: **Topic Analysis:** Briefly analyze the core scope of the topic, including explanations of key concepts and the direction of discussion. **Core Idea:** Propose the core of the argument.

Table 10: Prompt for generating core idea of essay.

You are participating in a debate competition, and the **topic of the debate** is: {topic}. Your team's **stance** is: {stance}. **Arguments and Evidence:**{claims} **Task Requirements:** Please organize the above argument outline into a high-quality debate speech according to the standards of clear logic, complete structure, and concise power.

Table 11: Prompt for generating Present.

DPE DPE introduces a two-stage planning strategy for argumentative essay generation, inspired by theories of argument structure. In the first stage, the model performs sketch planning to produce a rough outline of the essay. In the second stage, it engages in dialectical planning, where the outline is critically refined through self-reflection to improve logical rigor and persuasiveness. This planning framework explicitly emphasizes the exploration of reasoning processes, enabling large language models to generate essays that are more dialectical, diverse, and persuasive.

D2W D2W proposes a persona-driven multi-agent framework for argumentative essay generation, inspired by human debate. Multiple agents are assigned distinct personas, each reflecting unique high-level beliefs and perspectives on the topic. Through a structured debate and interaction process, these agents collaboratively exchange, challenge, and refine ideas to construct an overall plan for essay writing. This debate-to-write paradigm enables more diverse, coherent, and persuasive arguments by allowing nonlinear idea development and integration of multiple viewpoints.

D.3 Model Details

All LLM-based methods in our experiments—including our proposed framework and the other baselines—use the same base model for fair comparison. In specific, we use Qwen2.5-7B-Instruct, bart-base-chinese for the Chinese datasets and Llama-3.1-8B-Instruct, bart-base for the English datasets. The only exception is DD-KW, which is BART-based. To ensure its applicability to our Chinese dataset, we adapted it by replacing the base model with bart-base-chinese and substituting its original keyword extraction method with jieba for Chinese word segmentation.

E Evaluation Details

E.1 Automatic Evaluation Prompts

To perform automatic evaluation based on GPT-4o, we have designed the following evaluation dimensions and scoring criteria in this study. The scoring range for each dimension is from 1 to 5, with scores being precise to one decimal place. When scoring, the model is required to first provide brief feedback, followed by specific scores for each dimension. This design draws on the Chain-of-Thought (CoT) method, aiming to make the scoring process more rational and reduce randomness and variance in the evaluation.The evaluation prompts are shown in Table 19.

E.2 Human Evaluation details

Dimension of evaluation Due to inherent cognitive biases, human annotators may place uneven emphasis on certain dimensions depending on the topic or stance, which can lead to inconsistency in subjective overall judgments. Therefore, to ensure fairness and objectivity, we made the deliberate choice to exclude the overall human evaluation score and focus on five fine-grained dimensions.

Annotation training We recruited 5 graduate students with backgrounds in NLP and conducted a three-days training to ensure a clear understanding

You are in the midst of a debate on the **topic** of: {topic}.
Below is the statement made by your opponent in the previous round:
{statement}
Now, you need to challenge the following argument made by your opponent:
{currentClaim}
Please analyze the underlying logical flaws or unproven assumptions behind this argument, **identify its deeper issues**, and launch a strong counterattack from a broader perspective or a dimension not previously mentioned. **Your rebuttal should be profound, revealing the limitations or contradictions in their argument.**
Please proceed with your response directly, ensuring that it is logically sound, sharp in language, and concise.

Table 12: Prompt for generating Rebuttal.

You are currently engaged in a debate on the **topic** of: {topic}.
Below is your **previous statement**:
<Your Previous Statement>
{present}
</Your Previous Statement>
Your **opponent's attack statement** is as follows:
<Opponent's Attack Statement>
{rebuttal}
</Opponent's Attack Statement>
Now, you need to respond to their challenge by analyzing the root of the issue or its underlying logic, and revealing the deeper causes or implications of the argument. Consider the following approaches:
- **If the opponent's challenge is valid**, acknowledge its merit. Through further logical reasoning, explain why your argument still holds true, or re-examine your argument from a different perspective to reveal its broader applicability or significance.
- **If the challenge is not valid**, dissect the unproven assumptions in their challenge, point out its weaknesses, and introduce new perspectives or dimensions to demonstrate the rationality of your argument in a broader context.
Your response should be concise and forceful, avoiding verbosity or digression from the topic.

Table 13: Prompt for generating Respond.

of all evaluation dimensions. During this period, annotators received detailed scoring guidelines and examples. After the training stage, we conducted trial annotations. Only three qualified annotators pass the exam and were selected for the final evaluation.

Sample selection Due to the considerable length of the generated essays, it was infeasible to conduct human evaluations on all automatically evaluated instances. Thus, we comprimized selected 50 samples per method for manual evaluation.

Inter-annotator agreement We also calculate the inter-annotator agreement (IAA) using Krippendorff's Alpha (interval scale) across all dimensions of evaluation. The results are shown in Table 8. These score indicates that the manual results are acceptable (>0.6477).

The **topic** of the debate is: {topic}
Your team (affirmative) holds the position: {stance}
The argument framework you have constructed for your team is:
{claims}
The current debate record is as follows:
<Current Debate Record>
{debateRecord}
</Current Debate Record>
The argument you need to focus on (parent argument) is: {currentClaim}
Now, based on the debate record, you need to conduct an in-depth exploration of the parent argument by following these steps:
1. Referencing the debate record, identify the deeper contradiction or fundamental issue behind the parent argument using the reasoning chain: "Phenomenon → Direct Cause → Root Cause → Potential Impact."
2. Propose a deeper-level sub-argument. This sub-argument may be:
- A new argument that further touches the core issue of the parent argument.
- A contrast argument to the parent argument.
- A response to potential rebuttals.
Ensure the sub-argument does not overlap with existing ones.
3. Provide strong supporting evidence for the sub-argument (specific real-world examples, data, etc.). Make sure to provide new evidence that has not been used in your team's previous arguments.
If a deeper sub-argument can be identified, output the new sub-argument, its logical relationship to the parent argument, and supporting evidence in the format below:
<subClaim> This is the content of the sub-argument </subClaim>
<relation> This explains the logical relationship to the parent argument </relation>
<evidence> This is the supporting evidence for the sub-argument </evidence>
If no such sub-argument can be found, output:
<None>[Specific reason]</None>

Table 14: Prompt for in-depth expansion.

The topic of the debate is: {topic}
Your team (affirmative) holds the position: {stance}
Here are the arguments you have already provided for your team:
{claims}
The current debate record is as follows:
<Current Debate Record>
{debateRecord}
</Current Debate Record>
The argument you need to focus on now (parent argument) is: {currentClaim}
Existing sub-arguments under his parent argument:
{subclaims}
Now, based on the debate record, you need to further **explore the parent argument from other perspectives** by following the steps below:
1. Referencing the debate record, identify the deeper contradiction or essential issue behind the parent argument using the reasoning chain of "Phenomenon $\rightarrow$ Direct Cause $\rightarrow$ Root Cause $\rightarrow$ Potential Impact."
2. Propose a deeper-level sub-argument. This sub-argument may be:
- A new argument that further touches the core of the parent argument's issue.
- A contrast argument to the parent argument.
- A response to potential rebuttals.
Make sure the new sub-argument does not overlap with existing ones.
3. Provide strong supporting evidence for the sub-argument (specific real-world examples, data, etc.). Ensure that the evidence is new and not reused from your team's previous arguments.
If you are able to extract a deeper sub-argument, output the new sub-argument, its logical relation to the parent argument, and the supporting evidence in the following format:
<subClaim> This is the content of the sub-argument </subClaim>
<relation> This explains the logical relationship to the parent argument </relation>
<evidence> This is the supporting evidence for the sub-argument </evidence>
If you are not able to identify a deeper sub-argument, output:
<None>[Specific reason]</None>

Table 15: Prompt for breadth expansion.

Your team is on the affirmative side. Below is the current debate record between your team and your opponent:
<Current Debate Record>
{debateRecord}
</Current Debate Record>
Existing arguments:
{argumentTree}
You have now proposed a new claim for your team: {currentClaim}
Its corresponding parent claim is: {parentClaim}
The logical relationship with the parent argument is: {relation}
You need to eliminate redundancy between the new claim and its parent, and improve its rhetorical effectiveness.
First, analyze the rhetorical situation referring to the following three elements:
1.Exigence: Social Issues / Cognitive Biases / Ethical Conflicts / Value Dilemmas / Inaction;
2.Audience: Group Characteristics / Cognitive Foundation / Value Orientations / Potential for Change / Resistance Concerns;
3.Constraints: Social Context / Cultural Traditions / Ideological Frameworks / Resistance Concerns / Institutional Structures / Mainstream Opinion;
Then, according to the identified rhetorical situation, select one or more rhetorical devices to enhance the claim in terms of logical clarity, emotional resonance, and credibility building.
Available rhetorical strategies include:
Causal Reasoning / Contrast / Analogical Reasoning / Enumeration / Inductive Reasoning / Parallelism / Rhetorical Questioning / Metaphor / Personification / Quotation / Hyperbole / Cited Reference / Value Appeals.
Please output in the following format. Note: Only optimize the claim rhetorically, do not over-expand the arguments, only give the optimized claim, do not give too much content.
<explain> Briefly explain how to perform rhetorical optimization </explain>
<improvedClaim> This is the optimized argument </improvedClaim>

Table 16: Prompt for rhetorical optimization.

The **topic** of the argumentative essay is: {topic}
The **main theme of the argumentative essay** you are to write is: {stance}
The following are the **main arguments** of the essay:
{mainClaims}
Please write an **introduction and a conclusion** for the argumentative essay(Do not provide the main body of the article), which should echo each other. Please strictly follow the format below for output(Do not omit the tags </Introduction> and </Conclusion>):
<Introduction> Here is the content of the introduction </Introduction>
<Conclusion> Here is the content of the conclusion </Conclusion>

Table 17: Prompt for writing introduction and conclusion of essay.

The **topic** of the argumentative essay is: {topic}
The **central thesis** of the essay is: {stance}
The **introduction** of the essay is: {introduction}
The **arguments, their logical relationships, and supporting evidence** are as follows:
{argumentTree}
The **conclusion** of the essay is: {conclusion}
Note: Do not mechanically stack arguments. Pay close attention to the logical relationships between them. The writing should be coherent, logically rigorous, and transition naturally between ideas.
Based on the information above, please provide a complete argumentative essay between the tags ‘<essay>‘ and ‘</essay>‘.

Table 18: Prompt for writing essay.

Topic: {Topic}

Argumentative Essay: {Argumentative Essay}

Please evaluate the argumentative essay above based on the following dimensions:

Relevance (1-5 points): Assess the relevance of the essay to the topic. Evaluation criteria include: whether all claims in the article are related to the theme or central argument. Whether the arguments are sufficiently diverse and mutually supportive.

Logic (1-5 points): Assess the logic of the argumentative essay. Evaluation criteria include: whether the article's arguments are clear, whether the evidence is sufficient, whether the argument structure is rigorous, and whether there is a certain level and progression between different arguments.

Complexity of Claims(1-5 points): Assess the complexity of the claims in the essay. Evaluation criteria include: the claims should cover multiple perspectives or domains; Arguments should be based on multi-layered chains of reasoning or sub-arguments that support each other.

Persuasiveness(1-5 points): Assess the overall persuasiveness of the essay. Evaluation criteria include: whether the essay presents clear claims; whether it uses diverse evidence (such as facts, data, theories, etc.) and in-depth analysis to reveal the essence of the issue; whether the claims have a certain degree of universality, allowing the arguments to apply to broader contexts.

Rhetoric(1-5 points): Assess the language style of the essay. Evaluation criteria include: whether the essay uses clear and precise language to express ideas; whether rhetorical devices (such as metaphors, rhetorical questions, parallelism, etc.) are effectively employed to enhance persuasiveness; whether language and structure engage readers and make the essay more compelling.

Overall(1-5 points): Provide an overall evaluation of the essay by considering all the dimensions above.

Scores for each dimension should range from 1 to 5, with decimal precision allowed.

First, provide brief feedback on the quality of the essay, followed by specific scores for each dimension.

Output must strictly follow the format below, without adding any irrelevant content or symbols:

<feedback> Feedback here </feedback>

<Relevance> Relevance score </Relevance>

<Logic> Logic score </Logic>

<Complexity of Claims> Complexity of Claims score </Complexity of Claims>

<Persuasiveness> Persuasiveness score </Persuasiveness>

<Rhetoric> Rhetoric score </Rhetoric>

<Overall> Overall score </Overall>

Table 19: Prompt for evaluating essay.

Rhetoric optimization before:

The increasing popularity of advertising is a natural response to the complexities and challenges of modern society, as it provides a platform for businesses to adapt to changing consumer behaviors, technological advancements, shifting market trends.

After rhetoric optimization:

The increasing popularity of advertising is a natural response to the complexities and challenges of modern society, as it provides a dynamic platform for businesses to adapt to changing consumer behaviors, technological advancements, and shifting market trends, much like a resilient ecosystem that evolves to thrive in a changing environment. By reflecting our society's values and desires, advertising serves as a mirror that reveals our strengths and weaknesses, and by embracing this reflection, we can create a more responsible and sustainable advertising industry that promotes social responsibility, environmental sustainability, and consumer protection.

Table 20: An example of rhetorical optimization, with different rhetorical techniques marked with different colors: metaphor, causal reasoning.

Core: Celebrities' involvement in supporting international aid organizations can have a positive impact, as their influence can effectively raise awareness about global issues, inspire public engagement, and ultimately drive donations and support for these causes.

- **Claim**: Celebrities' involvement in supporting international aid organizations can have a positive impact because it leverages their influence to shape public discourse and create a sense of social norms around global issues, thereby fostering a more engaged and supportive community.

- **Evidence**: For instance, a study by the University of California, Berkeley, found that when a celebrity endorses a social cause, it can increase public discussion and engagement around that issue by 20-30%.

    - **Claim**: The influence of celebrities can be a direct driver of social change, without considering the potential for their involvement to create a "spectacle effect," where the focus shifts from the underlying issue to the celebrity's personal life or actions.

    - **Evidence**: ...

    - **Claim**: Celebrities' involvement in supporting international aid organizations should be complemented by a focus on amplifying and supporting grassroots initiatives and community-led projects, rather than relying solely on celebrity endorsements. This approach can help to create a more equitable and sustainable impact on global issues.

    - **Evidence**: ...

- **Claim**: Celebrities can leverage their influence to amplify marginalized voices, but this amplification can also perpetuate existing power dynamics, potentially silencing or tokenizing marginalized individuals and communities.

- **Evidence**: ...

    - **Claim**: Celebrities' involvement in amplifying marginalized voices can perpetuate a "silent majority" effect, not only because of the increased visibility and attention brought by celebrity involvement, but also because of the inherent power imbalance and lack of accountability in celebrity-led campaigns, which can create unrealistic expectations and pressure on marginalized individuals and communities to conform to the celebrity's message or image. This power imbalance and lack of accountability can lead to a situation where marginalized voices are not truly amplified, but rather tokenized or co-opted for the benefit of the celebrity's personal brand and image.

    - **Evidence**: ...

    - **Claim**: The romanticization of celebrity involvement in social justice movements can create a culture of performative activism, where celebrities prioritize their own brand and image over the authentic voices and stories of marginalized individuals and communities, ultimately undermining the potential for meaningful change.

    - **Evidence**: ...

Table 21: An example of an outline tree.

Topic:

Advertising is now becoming more popular in modern society. While some people believe this is a positive development, others see it as a negative one. Discuss both these views and give your opinion.

Argument:

In today's fast-paced, consumer-driven world, advertising has become an integral part of modern society, shaping the way we think, behave, and interact with one another. With its increasing influence, the role of advertising in modern society has sparked a heated debate, with some viewing it as a positive force that promotes economic growth, consumer awareness, and social progress, while others see it as a negative development that threatens cultural identity, exacerbates social inequality, and perpetuates a culture of disposability. As we navigate this complex landscape, it is essential to examine both sides of the argument and consider the multifaceted implications of advertising's growing popularity. In this essay, I will argue that advertising is a vital component of modern society that has become increasingly influential, and its growing popularity is a positive development that can be harnessed to promote economic growth, consumer awareness, and social progress.

One of the primary concerns surrounding advertising is its potential impact on cultural diversity. Some argue that the growing influence of advertising can lead to a homogenization of cultural values, where unique cultural identities are suppressed in favor of a dominant, consumerist culture. A study published in the Journal of Consumer Research found that exposure to advertising can lead to a decline in cultural identity and a shift towards a more homogenized, consumerist culture. For example, in Papua New Guinea, the introduction of Western consumer culture has led to the decline of traditional tribal practices and the erosion of cultural identity. Similarly, in the United States, the proliferation of advertising has contributed to the homogenization of cultural values, where unique cultural identities are suppressed in favor of a dominant, consumerist culture.

However, it is essential to note that advertising can also promote cultural exchange and understanding, allowing different cultures to share their values and traditions with a broader audience. By showcasing diverse cultural practices and traditions, advertising can foster a sense of inclusivity and appreciation for the richness of human experience. For instance, a study by the Smithsonian Institution found that cultural festivals and events sponsored by companies like Coca-Cola and Pepsi have helped to promote cultural exchange and understanding between different communities.

Another concern surrounding advertising is its potential to perpetuate a culture of disposability, where traditional practices and cultural identities are seen as disposable and inferior to modern, consumerist values. A study by the Ellen MacArthur Foundation found that the global value of discarded products and packaging is estimated to be around $4.5 trillion, highlighting the culture of disposability that has been perpetuated by consumerist values. Furthermore, a survey by the Pew Research Center found that 70% of millennials believe that traditional practices and cultural identities are important for their identity, but are also willing to discard them in favor of modern, consumerist values.

However, it is also worth noting that advertising can promote sustainability and environmental awareness, encouraging consumers to adopt more environmentally friendly practices and reduce their waste. For example, a study by the World Wildlife Fund found that companies like Patagonia and REI have successfully promoted sustainability and environmental awareness through their advertising campaigns, encouraging consumers to adopt more environmentally friendly practices.

In addition to its potential impact on cultural diversity and sustainability, advertising also plays a significant role in promoting economic growth and consumer awareness. The growing popularity of advertising has led to the creation of new industries and job opportunities, stimulating economic growth and development. For instance, a study by the World Advertising Research Center found that the global advertising market is projected to reach $860 billion by 2025, with the majority of this revenue being generated by companies that target high-income consumers.

However, it is essential to acknowledge that the growing wealth gap in modern society is a major concern, with those who have access to advertising and consumer goods gaining a significant advantage over those who do not. According to a study by the Organization for Economic Cooperation and Development (OECD), the top 10% of households in the United States hold 77% of the country's wealth, while the bottom 50% hold only 1%. This widening wealth gap is closely tied to the increasing influence of advertising, as those who have access to advertising and consumer goods have a significant advantage over those who do not.

Furthermore, the exploitation of consumer data is a major driver of the widening wealth gap, with companies targeting high-income consumers and leaving low-income communities behind. A study by the Electronic Frontier Foundation found that the average American is exposed to over 5,000 advertisements per day, and that this exposure is linked to a significant increase in consumer spending and economic inequality.

In conclusion, the growing influence of advertising on modern society is a double-edged sword that holds both promise and peril. While it is true that advertising can lead to a homogenization of cultural values and exacerbate social inequality, it also has the potential to promote economic growth, consumer awareness, and social progress. By harnessing the power of advertising, we can create a more inclusive and equitable society that values diversity, promotes social mobility, and fosters a culture of sustainability. Ultimately, the key to unlocking the full potential of advertising lies in its responsible use, one that prioritizes people over profit, promotes cultural heritage over consumerism, and seeks to create a more just and equitable world for all.

Table 22: An example of English argument generated by our method.

Topic:

“双十一”“双十二”，各路商家都会忙不迭地展开“短信攻势”，向广大手机用户发送堪称海量的广告信息。当然，用户也可以通过相关渠道退订或者举报未经允许发送的短信。有记者近日暗访发现，在一些互联网等平台，部分职业举报人发布模板化举报教程、宣扬举报获利，教学收费从数百元到上万元不等。

Argument:

在这个数字化时代，电商平台如“双十一”和“双十二”成为商家竞相展示商品、争夺市场份额的重要战场。然而，随着电商营销手段的不断进化，短信广告的泛滥也引发了广泛关注。商家们通过海量短信推送广告，试图吸引消费者的注意，但这种行为是否正当？它又如何影响了消费者与商家之间的关系？本文将探讨商家在电商节日期间通过短信进行广告营销的合理性及其法律界限，并剖析职业举报人在互联网平台上推广举报教程并收取高额费用的行为对市场秩序的冲击。

商家在电商节日期间通过短信进行广告营销虽然合理，但如果过度滥用，将如同病毒般侵蚀消费者对电商的信任，最终导致整个行业的信誉受损，甚至面临消费者流失的风险。首先，过度短信营销会迅速耗尽消费者的注意力资源，导致“信息疲劳”。当消费者每天接收到大量冗余的广告信息时，他们不仅会感到厌烦，更会质疑电商平台的专业性和诚信度。这种持续的信息轰炸不仅削弱了消费者对电商的整体信任感，还可能促使他们在未来的购物决策中倾向于其他平台，从而对整个电商行业造成不可逆的伤害。根据《消费者心理与行为学报》的一项研究显示，消费者每天接收大量信息会导致“信息疲劳”，进而影响其情绪和决策能力。特别是在电商节这样的促销高峰期，消费者面对海量的广告信息，容易感到厌烦和反感，这不仅降低了他们的购物意愿，还可能使他们对整个电商行业产生怀疑。例如，2020年“双十一”期间，某电商平台因过度短信营销导致用户投诉激增，最终使得该平台的用户满意度评分下降了15%，反映出短信营销过度滥用对消费者信任感的负面影响。

其次，过度短信营销不仅会迅速耗尽消费者的注意力资源，还会引发严重的隐私泄露问题，进一步损害消费者的信任。据《信息安全研究杂志》报道，频繁收到无意义短信的消费者中，有超过80%的人担心个人信息安全，其中60%的人曾遭遇过身份盗用或隐私泄露事件。这种隐私泄露不仅让消费者的个人信息处于危险之中，还让他们感到自己的权益被侵犯，从而对电商行业的信任度急剧下降。正如一棵健康的树木需要阳光和雨露滋养，商家的品牌形象也需要消费者的支持和信任来茁壮成长。一旦失去消费者的信任，就像失去了土壤中的养分，树木将无法抵御风雨的侵袭，最终枯萎凋零。商家应当在追求利益的同时，坚守法律底线，尊重消费者的选择权，构建健康的品牌形象。然而，职业举报人在互联网平台上推广举报教程并收取高额费用，这种行为损害了正常市场秩序，应当受到规范和监管。这种行为不仅加剧了消费者与商家之间的矛盾，还破坏了公平竞争的市场环境。据统计，过度短信营销导致的消费者流失率平均高达30%，这不仅直接减少了商家的市场份额，还削弱了品牌的市场竞争力。长期来看，这种短视行为将阻碍商家实现可持续发展，甚至可能导致品牌声誉的永久性损害。

过度短信营销不仅导致当前消费者的流失，还会对未来的新客户获取造成不利影响，形成恶性循环。据统计，过度短信营销导致的消费者流失率高达30%，这意味着商家不仅丧失了现有客户的信任，还失去了重新赢回这部分客户的成本。更糟糕的是，这种行为会进一步削弱商家的品牌吸引力，导致潜在客户望而却步，形成一种恶性循环。例如，一家名为“绿叶科技”的公司，在2020年“双十一”期间进行了大规模的短信营销活动，结果导致其品牌价值评估下降了15%，并在随后的一年内股价下跌了10%。这表明，过度短信营销不仅影响当前的销售业绩，还会对品牌的长期价值造成负面影响。

电商营销的繁荣离不开双方的信任与合作，过度短信营销如同一把双刃剑，既可能带来短期的利益，也可能导致长期的信任崩塌。商家应当在追求利益的同时，坚守法律底线，尊重消费者的选择权，构建健康的品牌形象。而对于职业举报人推广举报教程并收取高额费用的行为，相关部门应加强监管，维护市场的公平竞争环境。只有这样，电商行业才能在良性循环中稳健前行，共同构建一个更加透明、健康的市场生态。

Table 23: An example of Chinese argument generated by our method.