Decomposing Argumentative Essay Generation via Dialectical Planning of Complex Reasoning

Yuhang He[1,5], **Jianzhu Bao**[1,5], **Yang Sun**[1,5], **Bin Liang**[1,4], **Min Yang**[3], **Bing Qin**[1], **Ruifeng Xu**[1,2,5*]

[1] Harbin Institute of Technology, Shenzhen, China [2] Peng Cheng Laboratory, Shenzhen, China
[3] SIAT, Chinese Academy of Sciences, Shenzhen, China [4] The Chinese University of Hong Kong
[5] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

yuhang.he.hitsz@outlook.com, jianzhubao@gmail.com, yang.sun@stu.hit.edu.cn, bin.liang@cuhk.edu.hk, min.yang@siat.ac.cn
qinb@ir.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Argumentative Essay Generation (AEG) is a challenging task in computational argumentation, where detailed logical reasoning and effective rhetorical skills are essential. Previous methods on argument generation typically involve planning prior to generation. However, the planning strategies in these methods overlook the exploration of the logical reasoning process. Inspired by argument structure-related theories, we propose an argumentative planning strategy for prompting large language models (LLMs) to generate high-quality essays. This strategy comprises two stages: (1) Sketch planning, which creates a rough outline of the essay, and (2) Dialectical planning, which refines the outline through critical self-reflection. Such a planning strategy enables LLMs to write argumentative essays that are more logical, diverse, and persuasive. Furthermore, due to the scarcity of existing AEG datasets, we construct three new datasets. These datasets are from two domains: exam essays and news editorials, covering both Chinese and English. Automatic and manual evaluation on four datasets show that our method can generate more dialectical and persuasive essays with higher diversity compared to several strong baselines[1].

1 Introduction

Recent years have seen a surge of interest in argument generation tasks, such as counter-argument generation (Alshomary et al., 2021b; Alshomary and Wachsmuth, 2023), argumentative essay generation (Carenini and Moore, 2006; Bao et al., 2022c), and controlled argument generation (Alshomary et al., 2021a; Al Khatib et al., 2021). Automatic argument generation poses a significant challenge, as it requires not only coherent and logical sentences (Asher and Lascarides, 2005), but also the ability to think critically and deeply (Fisher, 2001; Freeman, 2011). As a prominent task in the field of argument generation, argumentative essay generation (AEG) aims at generating argumentative essays on controversial issues (Carenini and Moore, 2006; Bao et al., 2022c), which highlights the organization of convincing arguments to fully articulate viewpoints (Zukerman et al., 2000; Mann and Thompson, 1988). AEG is applicable in various domains, such as competitive debating (Bar-Haim et al., 2021), legal argumentation (Elaraby et al., 2023), academic essay writing (Bao et al., 2022c), and news editorials (Syed et al., 2020).

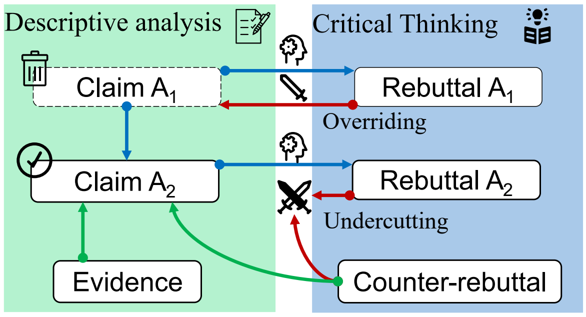

Figure 1: Dialectical planning involves two aspects. The primary content of descriptive analysis (claims and evidence), along with the complementary integration of critical thinking (rebuttals and counter-rebuttals), are key determinants of argument quality.

Previous research on argument generation has mainly focused on generating individual and comparatively concise arguments, typically involving a brief claim and its corresponding reasoning (El Baff et al., 2019; Jo et al., 2021). Only a limited number of studies have explored the generation of long-form argumentative essays, which contain multiple arguments across various aspects (Falk and Lapesa, 2023; Bao et al., 2022c). To better generate long-form text, prior work typically adopts the "plan-and-write" paradigm, generating intermediate planning first (e.g. keywords or relational

* Corresponding Authors

[1]Code and data are available at https://github.com/HITSZ-HLT/AEG_DPE.

Findings of the Association for Computational Linguistics: ACL 2024, pages 12305–12322
August 11-16, 2024 ©2024 Association for Computational Linguistics

triplets) to guide the final output (Hua and Wang, 2020; Hu et al., 2022). Although this approach has been proven beneficial for AEG, it only skims the surface in generating content-rich argumentative essays, overlooking the complex reasoning structure and dialectical nature of argumentative essays.

With the advancement of pre-training techniques, Large Language Models (LLMs) have shown remarkable capabilities in generation tasks (OpenAI, 2023; Ouyang et al., 2022). However, simple task instructions often fall short in steering LLMs to effectively handle complex tasks that need multi-step decomposition (Qin et al., 2023). Previous research has demonstrated that the use of Chain-of-Thought (CoT) (Wei et al., 2022) can significantly enhance the performance of LLMs on complex reasoning tasks by prompting the model to break down the task into a series of intermediate results. Likewise, as the AEG task also necessitates rigorous logical reasoning to generate persuasive essays, simple task instructions are inadequate to handle AEG well. Therefore, we aim to devise effective planning strategies to prompt LLMs to produce better-quality argumentative essays.

Our dialectical planning strategy is inspired by the argument theories proposed by Freeman (Freeman, 2011), which reflects what constitutes a high-quality argument as illustrated in Figure 1. According to Freeman's theory, a comprehensive argument should include claims, evidence, overriding rebuttals, and undercutting rebuttals. The claims and evidences are crucial aspects of descriptive analysis. Overriding rebuttals function as feedback countering the claims to guide the refinement of claims. Simultaneously, undercutting rebuttals and counter-rebuttals collaborate to enhance persuasiveness. Following the aforementioned theories, both humans and language models can develop better persuasive essays with in-depth analysis and critical thinking. However, current methods fail to consider these four determinants of argument quality thoroughly, nor do they address different types of rebuttals (Hu et al., 2022; Bao et al., 2022c).

To fill this gap, we propose a novel framework, DPE (Dialectical Planning of Essays), which utilizes task decomposition to generate high-quality argumentative essays with a fine-grained planning strategy inspired by Freeman's theory. During the planning stage, the framework arranges all vital components of the argumentative essay, which are then aggregated into a well-structured argumentative essay. Specifically, given a writing prompt on a controversial issue, the framework first makes a **sketch planning**, drafting the major claim and relevant claims, which serve as the foundational points of the essay. Subsequently, to foster profound and critical thinking, the framework performs **dialectical planning**, which involves: (1) Overriding Rebuttal that refines claims in sketch planning by identifying weakness, and (2) Undercutting Rebuttal that introduces an assistant counter-rebuttal to defend the claims. Lastly, following the planning stage, the framework implements **essay generation** to systematize the intermediate results and compose a final argumentative essay.

A commonly used dataset in existing AEG work is ArgEssay (Bao et al., 2022c), which is sourced from online forums. In this work, we additionally construct an English dataset and two Chinese datasets. Based on the aforementioned four datasets, we build a benchmark for AEG. The benchmark covers two common AEG scenarios: exam essay writing and journalism writing, both requiring opinion arguing on controversial topics. We conduct extensive experiments with both automatic and human evaluations. Our proposed DPE achieves overall better performance than several strong baselines. The results on three significant aspects of argument quality show that our method is able to generate more persuasive and dialectical essays. We summarize our contributions as follows:

- We propose an effective argumentative essay generation framework, DPE, which imitates the reasoning process proposed by Freeman's theory. DPE can generate argumentative essays incorporating in-depth reasoning and critical thinking.
- We collect three high-quality datasets and construct a benchmark for argumentative essay generation.
- Using both automatic and human evaluations, we demonstrate that our framework can generate more dialectical and persuasive argumentative essays and outperform several baselines.

2 Related Works

2.1 Argumentative Essay Analysis

Extensive research has been conducted on the analysis of argumentative essays since an early stage (Madnani et al., 2012; Beigman Klebanov and Flor, 2013). Traditional argumentative essay analysis work mainly studies the argument structure by mining the argumentative components and relations within essays (Stab and Gurevych, 2017; Potash et al., 2017; Sun et al., 2023; Bao et al., 2021a, 2022a; Sun et al., 2022). These studies contribute to a more comprehensive understanding of the argument process (Bao et al., 2021b, 2022b; Guo et al., 2023). Moreover, some other studies focus on analyzing the quality of arguments in writing essays. Lauscher et al. (2020) introduce a theory-based method of annotating and evaluating the quality of essays. Stede (2016) analyzes the depth of argument structure in the genre of news editorials.

These studies are highly relevant to our work, as the analysis of argument structure and quality can guide us in designing better mechanisms for argumentative essay generation.

2.2 Argument Generation

Early approaches to generating argumentative texts involve much manual effort, such as the construction of argument knowledge bases and the design of argumentation strategies (Reed, 1999; Zukerman et al., 2000; Carenini and Moore, 2000).

With the remarkable achievements of pre-trained generative models in recent years, argument generation has received much attention. Many argument generation-related tasks have been extensively studied, such as argument summarization (Fabbri et al., 2021; Syed et al., 2020; Elaraby et al., 2023), counter-argument generation (Alshomary et al., 2021b; Alshomary and Wachsmuth, 2023) and controlled arguments generation (Saha and Srihari, 2023; Schiller et al., 2021; Al Khatib et al., 2021). Hua et al. (2019) decompose counter-argument generation into content planning and realization to produce informative paragraphs. Fabbri et al. (2021) build a benchmark for argument summarization and incorporate argumentative essay analysis to enhance summarization. These argument generation studies mainly focus on generating individual and relatively short arguments.

Recently, some researchers have recognized the significance of generating long and coherent arguments. Bao et al. (2022c) collect the ArgEssay dataset for AEG and propose a dual-decoder model with content planning to generate content-rich argumentative essays. However, they fail to consider the argument structure and dialectical tier of argumentative essays. We decompose AEG into planning and realization using argument analysis to produce better results.

2.3 Planning and LLMs

Standard prompting follows a left-to-right generation process at the token level, which performs poorly for complex tasks (Bubeck et al., 2023). Subsequent works have emerged to involve complicated reasoning steps such as CoT (Wei et al., 2022), self-consistency (Wang et al., 2023b), and Tree-of-Thoughts (ToT) (Yao et al., 2023a). To improve performance on complex problems, recent works decompose the reasoning process and plan intermediate thoughts. (Zhang et al., 2023; Yao et al., 2023b; Huang et al., 2022). Hao et al. (2023) propose a reasoning solver based on LLMs to perform deliberate planning akin to human brains. Zhang et al. (2023) make use of a planning algorithm during decoding to generate more accurate results. In this paper, we explore using planning strategies specifically tailored for AEG to prompt LLMs for better reasoning and generation, thereby composing higher-quality argumentative essays.

3 Dataset Creation

In total, we use four datasets to evaluate our proposed method. Besides the ArgEssay dataset introduced by Bao et al. (2022c), we further collect three datasets for AEG. These three datasets encompass two types of argumentative essay-writing scenarios: exam essays and news editorials, covering both English and Chinese.

Exam Essays. Many examinations, such as IELTS, TOEFL, and the Chinese national college entrance examination, require students to write argumentative essays on controversial topics, which is a common task for testing language proficiency and critical thinking skills. Therefore, exam essays are a crucial data source for AEG. ArgEssay (Bao et al., 2022c), for instance, is collected from an online forum dedicated to revising English exam argumentative essays. However, apart from English data, there is a scarcity of exam essay data in other languages for AEG. Therefore, we propose a Chinese exam argumentative essay (CHE-Essay) dataset, which includes prompt-essay pairs from the Chinese national college entrance examination. In accordance with previous work, we search and download data from real and mock Chinese examinations available on doc-sharing websites (Tan et al., 2021).

News Editorials. News editorials are opinion articles written by editors or columnists of a news organization. Its primary purpose is to provide anal-

Dataset	language	Task	Avg.P	Avg.E	Size
Hua et al. (2019)	English	LTG	19.4	116.6	56,500
Hua and Wang (2020)	English	LTG	9.00	198.2	57,600
ArgEssay	English	AEG	38.59	341.52	11,282
NYT-Editorial	English	AEG	114.17	419.32	9,178
CHE-Essay	Chinese	AEG	192.60	961.03	3,750
CHN-Editorial	Chinese	AEG	171.06	1063.09	2,998

Table 1: Comparison of our dataset with existing argument generation datasets. Avg.P/Avg.E indicates the average number of tokens in the prompts/essays. Size indicates the total number of samples in the dataset. LTG is an abbreviation for long-form text generation.

ysis and insights into social events. Unlike standard news reporting, editorials express the author's subjective views, stances, and comments, making them more argumentative. For the AEG task, we gather two news editorial datasets, one in English and the other in Chinese. The English dataset, NYT-Editorial, is sourced from the "Room for Debate" section of New York Times[2], where a description of an event is provided as a prompt, followed by several professional writers presenting argumentative essays to express their views on that prompt. The Chinese dataset, CHN-Editorial, is collected from the news editorial section of two Chinese news websites, namely PengPai[3] and GuangMing-Wang[4].

We collect prompt-essay pairs from the aforementioned data sources, and then process them as follows:

- Manually excluding non-argumentative essays like propaganda essays, narrative essays, and news reports.
- Separating essays and prompts. For samples where the author does not list the prompt separately, we review and separate them manually.
- Cleaning up irrelevant text such as genre notifications, plagiarism warnings, author names, special characters, and timestamps, etc. We combine rule-based cleaning and manual review to ensure the quality of argumentative essays.
- To avoid ethical issues, we manually review and remove all essays that contain discriminatory, biased, or other inappropriate content.

[2]https://www.nytimes.com/roomfordebate
[3]https://www.thepaper.cn/channel_-24
[4]https://guancha.gmw.cn/node_11273.htm

We illustrate how the data processing works in Appendix G. We compare our datasets with existing argument generation datasets in Table 1. The datasets we propose contain argumentative essays that are longer and cover a wide range of topics. Also, the writing prompts within these datasets are much longer requiring models to have a deeper level of understanding.

4 Method

AEG task can be formulated as follows: given a writing prompt $X = [x_1, x_2, ..., x_k]$, an AEG system needs to generate an argumentative essay $Y = [y_1, y_2, ..., y_l]$.

Figure 2 illustrates the overall framework of DPE. It follows a writing paradigm that prioritizes planning prior to generation. To guide LLMs towards improved critical thinking capabilities, we devise an argumentative planning strategy consisting of two key elements: sketch planning and dialectical planning. Sketch planning prepares a preliminary outline of the essay, including a series of claims to be discussed. Based on the outlined claims, dialectical planning drives the model to make expansive and self-critique thinking, akin to reflective practices in human writing. Finally, LLMs integrate all the above planning results to produce a high-quality argumentative essay.

4.1 Sketch Planning

A high-quality argumentative essay typically centers on a major claim, which is strengthened by several supporting claims. These claims can be regarded as the skeleton of the essay. Thus, we first prompt LLMs to draft a major claim with multiple supporting claims in two consecutive steps:

$$\mathcal{M} : X \rightarrow \widetilde{c}^m \rightarrow (\widetilde{c}_1, \widetilde{c}_2, ..., \widetilde{c}_n) \tag{1}$$

where $\mathcal{M}$ is parameterized by LLMs, $\rightarrow$ indicates prompting LLMs with a specific prompt to generate the desired responses[5], $\widetilde{c}^m$ and $\widetilde{c}_i$ denote the draft of major claim and n supporting claims respectively.

4.2 Dialectical Planning

It is known that adopting critical thinking and ingenious argument strategies can enhance the persuasiveness of an essay (Toulmin, 2008; Wolfe et al., 2009; Musi, 2018; Freeman, 2011). Therefore,

[5]All specific prompts used in our method can be found in Appendix D.

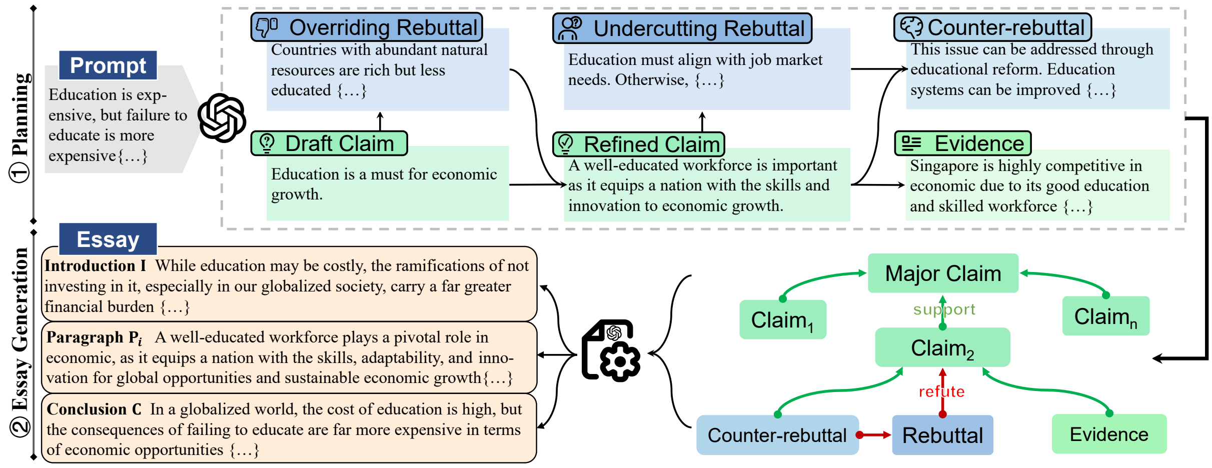

Figure 2: Overview of our framework. It meticulously plans all significant arguments in human-like reasoning steps. Then, LLMs organize arguments and underlying relations to yield an argumentative essay.

grounded in sketch planning, we present dialectical planning, which prepares arguments step by step. First, drawing from Freeman's theory, we propose overriding rebuttal and undercutting rebuttal to make LLMs to mimic deep thinking in argumentative reasoning. LLMs are required to address these two types of rebuttals differently. Concretely, LLMs must acknowledge overriding rebuttal to formulate refined claims accordingly. When faced with undercutting rebuttals, LLMs should put forward counter-rebuttals to negate such rebuttals. Second, we generate evidence in support of the refined claims. Lastly, the major claim will be updated conditioned on newly refined claims.

Overriding rebuttal. Directly instructing $\mathcal{M}$ to refine claims can be inappropriate, as there is no criteria to guide the revision process. Thus, we first prompt LLMs to challenge the claims $\widetilde{c}_i$ generated in sketch planning by producing an overriding rebuttal. Then, conditioned on the rebuttal, LLMs are required to optimize $\widetilde{c}_i$:

$$\mathcal{M} : \widetilde{c}_i \rightarrow r_i^o \tag{2}$$
$$\mathcal{M} : (\widetilde{c}_i, r_i^o) \rightarrow c_i \tag{3}$$

where r_i^o is the generated overriding rebuttal and c_i is the refined claim. The overriding rebuttal here functions as feedback, countering the input claim and pointing out its inherent weaknesses. Such an approach of improving claims by attacking the weakness has also been proven effective in other studies (Alshomary et al., 2021b).

Undercutting rebuttal. In argumentative writing, presenting potential rebuttals and then counter-rebutting them is a widely recognized argumentation strategy (Gao et al., 2019; Orbach et al., 2019). Although the refined claim rectifies the weakness, rebuttal still exists to undercut the claim in terms of feasibility or risks. In dealing with such a rebuttal, LLMs should address the rebuttal via a counter-rebuttal to defend the claim. Specifically, we prompt LLMs to first generate an undercutting rebuttal followed by its counter-rebuttal:

$$\mathcal{M} : c_i \rightarrow r_i^u \tag{4}$$
$$\mathcal{M} : (c_i, r_i^u) \rightarrow r_i^c \tag{5}$$

where r_i^u and r_i^c denote undercutting rebuttal and counter-rebuttal respectively.

Evidence. As evidence is crucial to support a claim, we further prompt $\mathcal{M}$ to produce k relevant evidence for each claim:

$$\mathcal{M} : c_i \rightarrow e_i \tag{6}$$

where c_i is the claim and e_i denotes the corresponding evidence.

Major claim. Since each claim is revised, the draft major claim should also be modified accordingly. Thus, we implement a bottom-to-top update of the major claim with refined claims c_i.

$$\mathcal{M} : (c_1, c_2, ..., c_n) \rightarrow c^m \tag{7}$$

4.3 Essay Generation

During the generation phase, we aggregate the planning content, including c_i, e_i, r_i^c, and r_i^u, into the introduction, conclusion, and body paragraphs. Then, these parts are transformed into a cohesive argumentative essay.

Leveraging the writing prompt and claims, we instruct the model to generate the introduction and conclusion:

$$\mathcal{M} : (X, c^m) \rightarrow \mathbf{I} \quad (8)$$

$$\mathcal{M} : (c^m, c_1, c_2, ..., c_n) \rightarrow \mathbf{C} \quad (9)$$

where **I** and **C** represent the introduction and conclusion respectively.

In paragraph generation, we explicitly incorporate argument relations (refute or support) into the instructions to prompt $\mathcal{M}$ with underlying rules (Zhu et al., 2023). For each refined claim c_i and its related arguments, we generate paragraph $\mathbf{P}_i$ as follows:

$$\mathcal{M} : (c_i, e_i, r_i^u, r_i^c) \rightarrow \mathbf{P}_i \quad (10)$$

Thus far, we have generated all parts of an argumentative essay. Finally, we generate an argumentative essay based on these parts:

$$\mathcal{M} : (\mathbf{I}, \{\mathbf{P}_i\}_l, \mathbf{C}) \rightarrow Y \quad (11)$$

where $\{\mathbf{P}_i\}_l$ denotes a set of l paragraphs and Y is the output essay.

5 Experiment Setup

5.1 Baselines and Model Implementations

We evaluate AEG in a zero-shot setting, where LLMs generate an argumentative essay based on a writing prompt without any demonstrations. Our method is evaluated across four datasets, two in Chinese and two in English. From each dataset, we randomly select 50 writing prompts for both automatic and human evaluation.

To verify the effectiveness of our framework, we compare it with the following zero-shot baselines that adopt different planning strategies by prompting LLMs. (1) **E2E**: End-to-end generation that directly produces a target essay ; (2) **CoT**: Chain-of-Thought generation (Wei et al., 2022) that first generates a brief plan as an intermediate guideline, and then generates an argumentative essay in the same response. (3) **ToT**: Tree-of-Thought generation (Yao et al., 2023a) that first derives multiple writing plans, evaluates the quality of each plan, and finally selects the best plan as the input prompt to produce an essay. We compare different LLMs including `Baichuan-7B` and `ChatGPT` for Chinese and `LLaMA2-7B` and `ChatGPT` for English. More implementation details are in Appendix A.

5.2 Evaluation Metrics

Automatic Evaluation. We find that most existing automatic metrics such as BLEU can not correspond closely with human judgments, as observed by Sellam et al. (2020). Motivated by recent developments in automatic evaluation where LLMs serve as judges and demonstrate good alignment with human evaluations (Wang et al., 2023a), we employ this approach for assessing the quality of generated essays in the following three aspects (Lauscher et al., 2020): (1) `Cog`: Cogency signifies whether an essay's premises are acceptable and sufficient to support the conclusion. (2) `Per`: Persuasiveness reflects whether an essay is well-organized, contextually appropriate, and emotionally appealing. (3) `Rea`: Reasonableness refers to the essay's ability to address counterarguments adequately, including acceptability and sufficiency in resolving the issue.

We leverage GPT-4 to score essays on a scale of 1 (worst) to 5 (best) (Liu et al., 2023a). We notice that recent work Zheng et al. (2023) emphasize the effects and solutions of self-enhancement and reasoning bias in scoring responses. Following them, we use CoT and reference-guided strategy to alleviate the bias. The evaluation templates and more details can be found in Appendix B. To reduce randomness, all experiments are performed three times, and the evaluation scores are averaged.

Human Evaluation. For a more comprehensive analysis, we conduct a human evaluation. We hire three well-educated master students to score the output quality following the three aspects in the automatic evaluation.

6 Results and Analysis

6.1 Automatic Evaluation

We present the evaluation results on English and Chinese datasets in Table 2 and Table 3, respectively. We observe that ChatGPT-based DPE achieves better performance by 0.11 in terms of `Avg` score compared to other ChatGPT-based methods on ArgEssay and NYT-Editorial. Among these LLaMA-based methods, DPE outperforms CoT and ToT and achieves comparable performance to E2E. On Chinese datasets, DPE outperforms other baselines with regard to ChatGPT-based methods while showing minimal drops (i.e., 0.02 drop in terms of `Avg` score on CHN-Editorial). This verifies the effectiveness of our dialectical planning.

We also find that CoT underperforms E2E in

Method	ArgEssay				NYT-Editorial			
	Rea	Cog	Per	Avg	Rea	Cog	Per	Avg
ChatGPT-based								
E2E	‡4.41	3.94	‡4.08	‡4.14	4.16	3.86	**4.00**	‡4.01
CoT	4.30	3.94	4.00	4.08	4.10	3.78	3.92	3.93
ToT	4.32	‡3.96	4.06	4.11	‡4.20	‡3.88	3.94	‡4.01
DPE	**4.44**	**4.20**	**4.12**	**4.25**	**4.30**	**4.08**	‡3.98	**4.12**
LLaMA-based								
E2E	**4.16**	3.91	**4.02**	**4.03**	**4.06**	‡3.64	‡3.88	‡3.86
CoT	3.92	3.70	3.88	3.83	3.84	3.52	3.82	3.73
ToT	3.98	‡3.92	3.88	3.92	3.90	3.66	3.78	3.78
DPE	‡4.02	**3.96**	‡3.92	‡3.97	‡4.00	**3.78**	**3.89**	**3.89**

Table 2: Automatic evaluation on English datasets. Avg is the average score of Rea, Cog, and Per. The best score is in **bold** and ‡ indicates the second-best result. The settings for all subsequent tables are consistent with this format.

Method	CHE-Essay				CHN-Editorial			
	Rea	Cog	Per	Avg	Rea	Cog	Per	Avg
ChatGPT-based								
E2E	4.02	3.84	‡3.60	3.82	4.13	**3.93**	3.54	‡3.87
CoT	4.06	3.78	3.50	3.78	4.12	3.86	3.52	3.83
ToT	‡4.10	**3.94**	3.58	‡3.87	‡4.14	3.85	‡3.62	3.86
DPE	**4.22**	‡3.90	**3.78**	**3.97**	**4.16**	‡3.92	**3.64**	**3.91**
Baichuan-based								
E2E	‡3.94	3.66	‡3.58	‡3.72	3.93	3.56	3.4	3.63
CoT	3.80	‡3.70	3.48	3.66	3.96	3.60	3.34	3.63
ToT	3.86	3.52	3.36	3.58	**4.04**	‡3.66	**3.44**	**3.71**
DPE	**4.06**	**3.76**	**3.66**	**3.82**	‡4.00	**3.70**	‡3.38	‡3.69

Table 3: Automatic evaluation results on Chinese datasets. Avg is the average score of Rea, Cog, and Per.

most metrics, possibly due to LLM's limited reasoning ability in handling complex planning. ChatGPT-based methods obtain significantly better results than LLaMA-based methods. An interesting finding is that DPE's performance exhibits a more pronounced improvement in conjunction with the enhanced capabilities of LLMs. This suggests that our method is particularly suited to benefit from the evolving landscape of LLM development, promising even greater efficacy as these models continue to advance.

Additionally, experiments on output diversity in Appendix C demonstrate that DPE has the best content richness. We think these improvements primarily stem from sound counter-rebuttals, as they constitute a substantial part of the essays, as analyzed in the Section 6.5.

Method	ArgEssay				NYT-Editorial			
	Rea	Cog	Per	Avg	Rea	Cog	Per	Avg
Ref	3.62	3.68	3.50	3.60	3.77	3.94	4.03	3.91
ChatGPT-based								
E2E	3.44	3.69	3.63	3.59	3.64	3.83	3.72	3.73
CoT	3.42	3.55	3.55	3.51	3.42	3.57	3.33	3.44
ToT	3.98	4.09	3.99	4.02	4.05	4.18	3.95	4.06
DPE	**4.26**	**4.48**	**4.33**	**4.35**	**4.32**	**4.43**	**4.26**	**4.34**
LLaMA-based								
E2E	3.39	3.45	3.32	3.39	3.30	3.41	3.24	3.32
CoT	2.91	2.87	2.80	2.86	2.94	3.00	2.98	2.97
ToT	3.54	3.57	3.45	3.52	3.55	3.53	3.45	3.51
DPE	**3.68**	**3.70**	**3.67**	**3.68**	**3.73**	**3.64**	**3.60**	**3.65**

Table 4: Human evaluation results on English datasets. Avg is the average score of Rea, Cog, and Per. Ref denotes ground truth essays. The average Fleiss' kappa is 0.51.

Method	CHE-Essay				CHN-Editorial			
	Rea	Cog	Per	Avg	Rea	Cog	Per	Avg
Ref	3.51	4.13	4.22	3.95	3.97	4.10	3.94	4.00
ChatGPT-based								
E2E	3.30	3.45	3.31	3.35	3.03	2.92	2.97	2.97
CoT	3.13	3.25	3.13	3.17	2.95	2.95	2.92	2.94
ToT	3.42	3.61	3.42	3.48	3.31	3.50	3.25	3.35
DPE	**4.02**	**4.25**	**4.05**	**4.10**	**3.74**	**4.11**	**3.68**	**3.84**
Baichuan-based								
E2E	3.49	3.71	3.44	3.54	3.29	3.36	3.06	3.23
CoT	3.12	3.31	3.07	3.17	3.37	3.52	3.18	2.97
ToT	3.41	3.77	3.40	3.53	3.41	3.65	3.21	3.42
DPE	**4.11**	**4.23**	**3.91**	**4.08**	**3.88**	**4.16**	**3.79**	**3.94**

Table 5: Human evaluation results on Chinese datasets. Avg is the average score of Rea, Cog, and Per. Ref denotes ground truth essays. The average Fleiss' kappa is 0.48.

6.2 Human Evaluation

Human evaluation results on the English and Chinese datasets are presented in Table 4 and Table 5, respectively. Our DPE exhibits considerably better performance in terms of Rea, Cog, Per, and Avg. The ChatGPT-based DPE significantly outperforms other baselines relying on the superior reasoning abilities of ChatGPT, while LLaMA/Baichuan-based DPE show moderate improvements, demonstrating the effectiveness and adaptability of our approach in leveraging the more capable LLMs.

Owing to limited reasoning ability and self-enhancement bias (Zheng et al., 2023) of GPT4, there are discrepancies between automated evaluators and humans, particularly in the evaluation of lengthy essays with complex structures. For example, automatic evaluators prefer the outputs of LLaMA-based E2E, while human evaluators

Model	Rea	Cog	Per	Avg
DPE	**3.61**	**4.21**	**3.97**	**3.93**
w/o OR	3.51	4.06	3.82	3.80
w/o UR	3.16	3.96	3.56	3.56
w/o OR+UR	3.39	4.11	3.84	3.78

Table 6: Ablation Study.

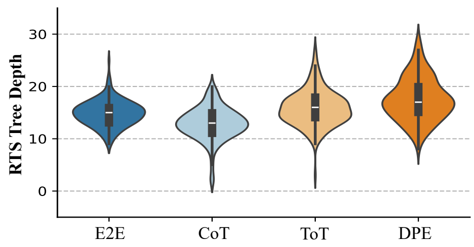

Figure 3: Distribution of RST tree depth of argumentative essays.

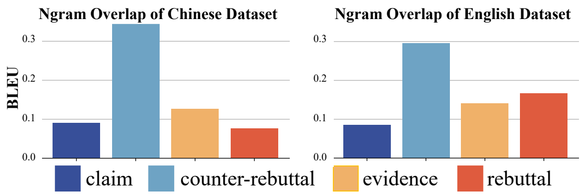

Figure 4: Ngram overlap between planning content and output essays

find DPE better. Furthermore, human evaluators prefer essays of DPE in 68% cases with acceptable pairwise-agreement. We also perform a case study and find that DPE generates argumentative essays with critical thinking by tailoring rebuttal and counter-rebuttal. The experiment results can be found in Appendix E.

6.3 Ablation Study

We evaluate 20 samples to reveal the effect of each module in our model. To analyze the impact of different modules in our method, we conduct ablation studies in terms of removing overriding rebuttal (*w/o* OR), removing undercutting rebuttal (*w/o* UR), and removing both them (*w/o* OR+UR). The average performance of four datasets is reported in Table 6. We observe that *w/o* OR decreases the performance, verifying the importance of the refining claims. Removing UR leads to a significant drop in performance, indicating the effectiveness of reasoning is improved by addressing undercutting rebuttals. Surprisingly, removing both OR and UR *w/o* OR+UR) does not result in a catastrophic drop in performance. It even outperforms *w/o* UR. This suggests that using either OR or UR individually might confuse LLMs while leveraging the synergistic effect of both OR and UR can effectively enhance the model's performance.

6.4 Analysis on Discourse Structure

The discourse structure offers insights into the text's high-level organization, making the depth of the Rhetorical Structure Theory (RST) tree a key indicator of text quality (Stede, 2016). Hence, we parse output essays into RST trees and analyze the depth distribution of the RST trees to assess the quality, with the results presented in Figure 3. DPE generates argumentative essays with deeper structures on average. Additionally, it displays a broad depth distribution, encompassing a wider range than ToT, CoT, and E2E methods. This indicates that our method effectively contributes to creating more diverse and intricate structures.

6.5 Analysis on Planning Content

During the planning phase, we generate the essential planning content, including claims and evidence for crafting a well-structured essay. We conduct experiments to assess how various planning content influences the output essay. To do this, we compute the n-gram overlap between the output essay and each type of argumentative planning content using BLUE score (Papineni et al., 2002). We present the results in Figure 4.

We find that the output essay incorporates planning content from every reasoning step in the dialectical reasoning process. Notably, counter-rebuttal achieves the highest scores, further indicating the importance of critical thinking. As an example in Figure 19 illustrates, an effective counter-rebuttal not only addresses the potential counterarguments but also strengthens the entire reasoning in support of the claim.

7 Conclusion

In this work, we propose a theory-based framework that decomposes AEG by dialectically planning intermediate reasoning steps. The planning strategy imitates human critical thinking to generate more logical and persuasive essays. To boost further study on AEG, we build a benchmark in Chinese and English where three datasets are newly collected. The experiment results on the benchmark demonstrate the superiority of our framework.

Limitations

We mainly study two common scenarios for AEG, that is, exam essays and news editorials. Other scenarios may require different genres, such as speech and legal argumentation. Consequently, for future research, we aim to explore a more intelligent framework capable of automatically devising reasoning steps for various topics and scenarios.

In our method, we do not design mechanisms to guarantee faculty. Therefore, future research could explore enhancing content reliability through the verification and editing of intermediate arguments. Additionally, the editing of arguments can double as a plugin for controllable AEG.

Ethics Statement

We strictly follow the licenses and policies of released LLMs and publicly available datasets. Our datasets are created from publicly available sources without any personal identity information. When crawling data from "New York Times", "PengPai", and "GuangMingWang", we follow the privacy policy and terms of use of these platforms. We have annotated the source of each data to comply with the copyright regulations. To address ethical concerns, we review and remove any essays that are considered to be libelous, racist, or otherwise inappropriate.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of GuangDong 2023A1515012922, the Shenzhen Foundational Research Funding JCYJ20220818102415032 (JCYJ20210324115614039), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

References

Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument 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 4744–4754, Online. Association for Computational Linguistics.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021a. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

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

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

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Studies in natural language processing. Cambridge University Press.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021a. 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 (Volume 1: Long Papers)*, pages 6354–6364, Online. Association for Computational Linguistics.

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022a. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. 2021b. Argument pair extraction with mutual guidance and inter-sentence relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3923–3934, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianzhu Bao, Jingyi Sun, Qinglin Zhu, and Ruifeng Xu. 2022b. Have my arguments been replied to? argument pair extraction as machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 29–35, Dublin, Ireland. Association for Computational Linguistics.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022c. 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.

Roy Bar-Haim, Liat Ein-Dor, Matan Orbach, Elad Venezian, and Noam Slonim. 2021. Advances in debating technologies: Building AI that can debate humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Giuseppe Carenini and Johanna D. Moore. 2000. A strategy for generating evaluative arguments. In *INLG 2000 - Proceedings of the First International Natural Language Generation Conference, June 12-16, 2000, Mitzpe Ramon, Israel*, pages 47–54. The Association for Computer Linguistics.

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

Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is GPT-4 a good data analyst? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9496–9514, Singapore. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Alec Fisher. 2001. Critical thinking: An introduction. Cambridge University Press.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. 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 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Jia Guo, Liying Cheng, Wenxuan Zhang, Stanley Kok, Xin Li, and Lidong Bing. 2023. AQE: Argument quadruplet extraction via a quad-tagging augmented generative approach. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 932–946, Toronto, Canada. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. 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 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.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on 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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Nitin Madnani, Michael Heilman, Joel R. Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 20–28. The Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Elena Musi. 2018. How did you change my view? a corpus-based study of concessions' argumentative role. *Discourse Studies*, 20:270 – 288.

OpenAI. 2023. Gpt-4 technical report.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. 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 5591–5601, Hong Kong, China. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Chris Reed. 1999. The role of saliency in generating natural language arguments. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 876–883. Morgan Kaufmann.

Sougata Saha and Rohini Srihari. 2023. ArgU: A controllable factual argument generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8373–8388, Toronto, Canada. Association for Computational Linguistics.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguistics*, 43(3):619–659.

Manfred Stede. 2016. Towards assessing depth of argumentation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3308–3317, Osaka, Japan. The COLING 2016 Organizing Committee.

Yang Sun, Bin Liang, Jianzhu Bao, Min Yang, and Ruifeng Xu. 2022. Probing structural knowledge from pre-trained language model for argumentation relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3605–3615, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Sun, Bin Liang, Jianzhu Bao, Yice Zhang, Geng Tu, Min Yang, and Ruifeng Xu. 2023. Probing graph decomposition for argument pair extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13075–13088, Toronto, Canada. Association for Computational Linguistics.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5384–5396, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online. Association for Computational Linguistics.

Stephen E. Toulmin. 2008. *The Uses of Argument, Updated Edition*. Cambridge University Press.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. 2023. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules. *ArXiv*, abs/2310.07064.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG 2000 - Proceedings of the First International Natural Language Generation Conference, June 12-16, 2000, Mitzpe Ramon, Israel*, pages 55–62. The Association for Computer Linguistics.

A More Implementation Details

In our experiments, all ChatGPT-based models are implemented with `gpt-3.5-turbo-1106` API. For `Baichuan-7B`[6] and `LLaMA-7B`[7] experiments, we strictly follow the commands and procedures to recover the weights of these models. We set the temperature as 0.8 and top-k as 50 for generation in all models. For claim generation in sketch planning, we set the number of claims to be generated as $m = 3$. We generate $k = 2$ pieces of evidence for each claim. In overriding rebuttal generation, we prompt the LLMs in one iteration to revise the draft claims. We generate $l = 3$ paragraphs for each argumentative essay. Additionally, we show the prompt used in E2E, CoT, and ToT in Figure 6 to 8.

B Automatic Evaluation Setting

The automatic evaluator is based on `GPT-4` (`get-4-0613`). The evaluator is required to first give feedback and then predict the scores in a CoT manner. The prompts are reference-guided with a comprehensive list of criteria of Reasonableness, Cogency, and Persuasiveness. The evaluation prompts are shown in Figure 5.

Recent researchers use GPT-4 as an NLG evaluator to score a single response on a scale of 1 to 100. The GPT-4 only received ambiguous evaluation criteria like "a score of zero means disfluency and score of one hundred means perfect fluency" (Liu et al., 2023a). However, such evaluation introduces two biases (Zheng et al., 2023): (1) self-enhancement bias where LLM evaluators may favor the responses generated by themselves or other models. (2) reasoning bias where LLM evaluators make incorrect judgments due to their limited capability in scoring reasoning questions. To make the evaluator score essays fairly, we prompt GPT-4 to score by giving reference-guided scores ranging from 1 to 5. Additionally, the generated feedback can enable GPT-4 to evaluate more reasonably. Our early evaluation didn't adopt these techniques and suffers from bias (Zheng et al., 2023; Wang et al., 2023a). In practice, we observe that the feedback can cite the input reference to give stable and reasonable scores. An example is shown as follows:

{"Feedback": "This essay presents a clear argument against the consumption of sugar-based drinks, with a particular emphasis on how aggressive marketing and advertising contributes to this problem. It offers solid evidence to back up these claims, most notably citing various studies and their results. Although it addresses the potential counterargument that hectic lifestyles might contribute to the consumption of such drinks, it quickly debunks it with evidence. The essay could improve its persuasiveness by using more emotional appeals, perhaps by describing the health dangers associated with excessive sugar consumption on a more personal level.", "Reasonableness": 5, "Cogency": 4, "Persuasiveness": 4}

Yet we also observe that the GPT-4 evaluator occasionally outputs positive feedback but relatively low scores when the input essays are complex in

Reasonableness:
1-The authors do not conclude by analysis. The authors do not contribute to the resolution of the issue and do not address counterarguments.
2-The authors provide little information to support the claims and thus neither contribute to the resolution nor draw a valid conclusion. Counterarguments are not addressed.
3-The authors analyze and present some information to support the claims. The authors mention counterarguments but barely rebut them.
4-The readers may not accept the conclusion because it makes broad generalizations or lacks enough supportive information. The authors address counterarguments but may not adequately rebut them.
5-The readers would accept the conclusion and consider it in the larger discussion. The authors rebut and adequately address counterarguments.

Cogency:
1-The authors provide no evidence for their claims.
2-The evidence for the author's claim may not be believable. The authors do not provide support to draw a conclusion.
3-The authors provide some relevant and acceptable evidence (the evidence can be more specific) for their claims but not enough to draw a conclusion.
4-The essay includes acceptable and relevant evidence (basic reasons or examples) that may or may not provide enough support to draw a conclusion.
5-The essay contains acceptable and believable evidence for the author's claims. The evidence (effective facts or statistics) is relevant to the author's point and is sufficient for drawing a conclusion.

Persuasiveness:
1-The way the essay is written and organized is hard to follow. The content is not presented in a sensical order.
2-The authors evoke emotions that make the readers less likely to agree. The authors do not use clear language or organization.
3-Although the argumentative essay is appropriate, the authors do not evoke an emotional response. The organization of the argument could be improved.
4-The authors use clear and appropriate language and structure their argument in a way that makes sense. The essay is lacking in emotional appeal.
5-The authors present their argument using clear organization and language. The authors demonstrate their opinion with supporting points, which is an effective organizational structure. The author evokes emotions that make the argument more agreeable.

Instructions:
You are provided with an argumentative essay and some descriptions about Reasonableness, Cogency, and Persuasiveness. You should first provide brief feedback of the essay's quality. Then, please select the most appropriate description of the essay ranging from 1 to 5. You need to output your options in JSON format.
An example is as follows:
"Feedback": "Feedback should cover pros, cons, suggestions, etc.", "Reasonableness": 3, "Cogency":3, "Persuasiveness":3

Figure 5: The prompts we used for the GPT-4 evaluator.

[6]https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat
[7]https://huggingface.co/meta-llama/Llama-2-7b-hf

disclosure structure and logic. Instead, GPT-4 may prefer shorter essays with straightforward expression. This partly explains why E2E sometimes gets higher scores than ToT and DPE while humans think otherwise.

Writing prompt: {input}
Write an argumentative essay according to the above prompt. Only output an essay between the tags <essay> and </essay>.

Figure 6: The prompts for E2E.

Writing prompt: {input}
Write an argumentative essay following the writing prompt.
Make a brief plan then write. Your output should be in the following format:

<plan>Your plan here</plan>
<essay>Your essay here</essay>

Figure 7: The prompts for CoT.

Writing prompt: {input}
Original writing plan :{plan}
Make or improve the original plan for writing an argumentative essay. Your output should be in the following format:
<plan>Your plan here</plan>

Writing plan: {plan}
Analyze the quality of the writing plan for an argumentative essay. Score it from 0 to 10, where score 10 means the plan is coherent and logical.
Provide reasoning for your scoring and place the score numbers between tags <score> and </score>.

Writing prompt: {input}
Writing plan: {plan}
Write a coherent and persuasive argumentative essay following the writing prompt. You should produce the essay based on the plan. Your output should be in the following format:
<essay>Your essay here</essay>

Figure 8: The prompts for ToT.

C Analysis on Diversity

We analyze output diversity by averaging the number of distinct bigrams (Li et al., 2016). We compute average distinct bigrams for Chinese and English datasets in Figure 9.

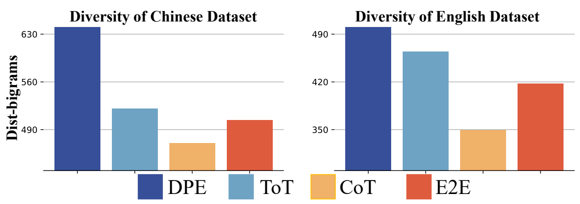

Figure 9: Distribution of Dist-bigrams of argumentative essays.

Our method generates more diverse content compared to others, especially in the Chinese dataset. On the contrary, CoT yields the least diverse outputs. We hypothesize that thinking step by step is more suitable to achieve an assertive answer instead of a diverse reasoning result. E2E outperforms CoT as it follows the instructions well to output essays naturally and diversely. ToT benefits from producing more detailed plans and selecting the best one. Decomposing AEG with our method further improves diversity since we introduce claims, evidence, and rebuttals of different perspectives, leading to high content richness.

D Prompts for DPE

To better illustrate how generations of argumentative components work, we provided detailed prompts in the dialectical planning stage. Concretely, the prompt we used for sketch planning is in Figure 10 and 11. The prompt for generating overriding and undercutting rebuttals is in Figure 12. The prompt for generating claims, counter-rebuttal, and final major claim are presented in Figure 13, 14, and 15, respectively. The prompts we used in the essay generation stage are in Figure 16 to 18.

Writing prompt: {input}
Write a concise, contentious, and coherent Thesis Statement (major claim) given the writing prompt.

Figure 10: The prompt for planning a draft of major claim.

Major Claim: {input}
To support the major claim, please further derive {num_branches} effective claims in one sentence. Think about the claims from different perspectives.
Please Note that each claim must end with token <sep>.

Figure 11: The prompt for planning draft of claims.

E More Human Evaluation

E.1 Human Ranking Result

we investigate human preferences where human annotators are required to rank output essays based on identical prompts. Table 7 illustrates the pairwise pairwise-agreement among the three annotators, measured by Kendall's τ, yielding an average of 0.41. We observe that Annotator 3 and Annotator 2 agree notably more with each other than with Annotator 1.

Table 8 reports the avg-rank and topmost percentage achieved by each approach. Our method achieved 1.5 avg-rank indicating that essays generated with self-critique perform best. As for the top-p, our method ranks first in 68% cases. Compared

Claim: {claim}
Evaluate the claim and refute it. Only output {num_branches} pieces of rebuttal. Please Note that each rebuttal must end with a special token <sep>.

Figure 12: The prompt for planning the overriding and undercutting rebuttals.

Claim: {claim}
Rebuttal: {rebuttal}
Improve the above claim considering the weakness that the rebuttal points out. Directly output an improved claim in one sentence without any supporting evidence or acknowledging the weakness again.

Figure 13: The prompt for planning refined claims.

Claim: {claim}
Rebuttal: {rebuttal}
Carefully review the claim and rebuttal. Please write a brief and persuasive counter-rebuttal to defend your claim or give solutions. Only output the counter-rebuttal.

Figure 14: The prompt for planning counter-rebuttals.

Writing prompt: {input}
Claims: {claim}
Write a concise, contentious, and coherent Thesis Statement (major claim) given the writing prompt. The thesis statement should cover the main points of the listed claims.

Figure 15: The prompt for planning the final major claim.

Writing prompt: {input}
Major claim: {majorclaim}
Write an introduction of an argumentative essay given a writing prompt. The introduction should clearly state the major claim.

Figure 16: The prompt for writing an introduction.

Claims: {claim} Write a conclusion of an argumentative essay referring to the claims above. Only output a concise conclusion.

Figure 17: The prompt for writing a conclusion.

with other methods, DPE stands out by delivering high-quality essays.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	-	0.29	0.41
Annotator 2	0.29	-	0.54
Annotator 3	0.41	0.54	-

Table 7: Pairwise inter-annotator agreement of ranking in terms of Kendall's τ

E.2 Is GPT-4 evaluator qualified?

Although the GPT-4 diverges from humans (Shen et al., 2023), the GPT-4 is a qualified evaluator giving reasonable scores. Using GPT-4 to evaluate is much more reliable than other automatic metrics (Rouge, BLEU). Other works (Liu et al., 2023b; Wang et al., 2023a) also prove GPT-4 to be a qualified evaluator. It have demonstrated its ability in various analytical tasks (Cheng et al., 2023; Ding et al., 2023). Following these works, we attempt to use GPT-4 socre on more complex criteria and we compute the Spearman correlation to be around 0.33. How to make GPT-4 as a human-level evaluator is left for future work.

Claim: {claim}
<Hints>
Evidence: {evidence}
Rebuttal: {evidence}
Counter_rebuttal: {counter_rebuttal}
</Hints>
The primary incorporation of claim and evidence coupled with the complementary integration of rebuttal and counter-rebuttal determine the quality of the argument. Evidence is important to support Claim. You should adequately rebut and address the rebuttal according to Counter_rebuttal to further support the claim. Consider the main points of the Hints. Write a coherent and persuasive body paragraph in favor of the claim.

Figure 18: The prompt for writing a paragraph.

Method	avg-rank ↓	top-p (%)
E2E	2.90	8%
CoT	3.26	5%
ToT	2.30	19%
DPE	**1.50**	**68%**

Table 8: Human ranking: *avg-rank* signifies average ranks across all samples within four datasets; *top-p* indicates the percentage of instances where the model achieved topmost.

E.3 Qualitative Comparison with Previous Work

We conduct a qualitative analysis to compare DPE with previous SOTA Bao et al. (2022c). We sample 50 essays from the ArgEssay dataset and manually compare the cogency, persuasiveness, and reasonableness. The annotators consistently agree that essays generated by DPE are better considering these three aspects.

E.4 Case Study

We present an example output highlighting intermediate argumentative components with different colors in Figure 19. Due to limited space, we only show an introduction and the following paragraph. Our approach incorporates the prompt and major claim during the essay generation stage. As a result, the introduction smoothly transitions from the writing prompt to the thesis statement. Moreover, the essay articulates a profound perspective, facilitated by the feedback of overriding rebuttal. To persuade the readers, the essay adequately refutes po-

Prompt: Many of this summer's action-filled blockbusters feature destruction and mayhem, in keeping with the trend of PG-13 movies becoming increasingly violent. But much of that violence is cartoonishly sterilized and bloodless to ensure the movies maintain a rating that can attract young ticket buyers. Some argue this shows how meaningless the rating system is, but others ask whether this "sugarcoating of violence" in PG-13 films makes for a more enjoyable movie-going experience. Is it irresponsible to portray conflict without its gory consequences?

The increasing violence in PG-13 movies has become a topic of debate, with some arguing that the sugarcoated portrayal of conflict allows for a wider audience to enjoy the movie experience **while** others raise concerns about its potential impact on younger viewers. In this essay, we will examine the major claim that depicting conflict without excessive gore in these films creates a more accessible narrative and storytelling opportunity for filmmakers. **However**, we will also explore the potential consequences of excluding gory aftermath, particularly in terms of how it may influence the perceptions and understanding of conflict among younger viewers. **By** delving into both perspectives, we can gain a comprehensive understanding of whether the sugarcoating of violence in PG-13 blockbusters is irresponsible or a necessary compromise.

Portraying conflict without excessive gore in PG-13 films allows for a wider audience to engage with the characters and their experiences, while also allowing filmmakers to focus on more nuanced storytelling elements. **For example**, films like "The Hunger Games" series and "The Dark Knight" showcase intense conflict and explore complex themes without relying on excessive violence or gore. **While** some may argue that this approach limits storytelling and diminishes the impact of the narrative, it is possible to create a compelling and thought-provoking story by focusing on other elements such as character development, dialogue, and pacing. **By** emphasizing the emotional and psychological aspects of conflict, filmmakers can tap into universal themes that resonate with viewers of all ages. **Moreover**, a more restrained approach to conflict can encourage filmmakers to explore creative and innovative ways to depict tension and suspense. **Therefore**, portraying conflict without excessive gore in PG-13 films not only widens the audience but also enhances the overall cinematic experience. {...}

Figure 19: Example output of our method with the intermediate argument components in different colors: claim, evidence, rebuttal, counter-rebuttal. Discourse markers are emphasized in **bold**.

tential counterarguments and proposes valid solutions. This implies that organizing the reasoning by decomposing the essay into argumentative components enhances overall persuasiveness. Discourse markers sensibly embody the logic between individual sentences. By explicitly incorporating argumentative relations, the essay avoids self-contradiction. More example outputs and comparisons can be found in Figure 22 and Figure 21. We present the introduction and a body paragraph for each approach.

F Cost for OpenAI API

Specifically, the cost per essay of OpenAI API and the average tokens number of generated essays are shown in the following table:

	E2E	CoT	ToT	DPE
average cost	0.001708 $	0.002370 $	0.014390 $	0.021890 $
average tokens	854	1185	7159	10945

G Data Process

It is believed that an argumentative essay should have a clear claim and line of reasoning. We review the first few paragraphs to exclude the propaganda essays,narrative essays, and news reports. As for propaganda essays, such essays typically include political speeches and advertisements. We defined propaganda essays in the following criteria: 1) essays that use selective facts, exaggerations, or distortions of truth to persuade the reader and 2) essays that repeat slogans, key messages, and concepts. The annotators are require to review the data in four steps as shown in Figure 20.

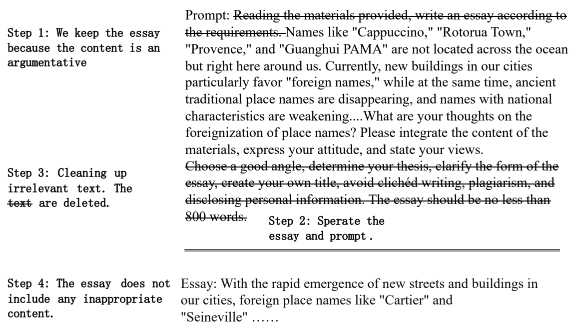

Figure 20

Prompt: Technological developments lead to lots of environmental issues. Some people think a simple lifestyle can preserve the environment. While others argue technology itself can solve it. Discuss both these views and give your own opinion.

E2E:

Technology has undoubtedly revolutionized our lives, making it more convenient and efficient. However, it has also brought about numerous environmental issues that cannot be ignored. While some individuals advocate for a simpler lifestyle as a means to preserve the environment, others argue that technology itself can provide the solutions needed. In this essay, I will discuss both viewpoints and present my own opinion on this matter.

{...} In my opinion, both viewpoints have their merits, and a combination of approaches is necessary to address the environmental challenges we face. While a simpler lifestyle can indeed contribute to reducing environmental impact, it may not be feasible or practical for everyone. Technology, on the other hand, has the potential to provide innovative solutions that can be implemented on a larger scale. However, it is crucial to ensure that these technological advancements are developed and utilized in a sustainable manner. This means considering the entire lifecycle of technology, including the extraction of raw materials, manufacturing processes, and disposal methods. Additionally, it is essential to prioritize research and development in environmentally friendly technologies.{...}

CoT:

In our modern world, technological developments have brought about numerous environmental issues. While some believe that adopting a simple lifestyle is the key to preserving the environment, others argue that technology itself can provide the solutions. This essay will discuss both views and present a balanced opinion.

{...}It is crucial to recognize that both a simple lifestyle and technology have their merits. Rather than adopting an extreme stance, a balanced approach is necessary. Individuals can embrace elements of a simple lifestyle by reducing excessive consumption, embracing sustainable practices, and prioritizing the use of renewable resources. Simultaneously, advancements in technology should be channeled towards sustainable solutions, such as investing in research and development of clean energy technologies, creating eco-friendly manufacturing processes, and promoting circular economy practices.{...}

ToT:

Technological advancements have undoubtedly brought numerous benefits to society, but they have also caused significant environmental issues. While some argue that adopting a simple lifestyle can effectively preserve the environment, others believe that technology itself holds the key to solving these issues. In this essay, we will explore both viewpoints and ultimately argue for a balanced approach that combines elements of both a simple lifestyle and technological advancements.

{...}Both viewpoints have their strengths and weaknesses. Embracing a simple lifestyle can certainly make a difference on an individual level. By reducing consumption, conserving energy, and opting for sustainable alternatives, individuals can contribute to the preservation of the environment. However, it is important to acknowledge the limitations of relying solely on a simple lifestyle to address global environmental issues. The scale of the challenges we face requires more comprehensive solutions that go beyond individual actions. Technology, on the other hand, offers immense potential for addressing environmental problems. Its ability to provide alternative energy sources, improve transportation systems, and enhance waste management cannot be overlooked. However, technology alone cannot solve all environmental issues. It requires responsible and sustainable implementation to ensure that the benefits outweigh the potential harm caused by its production and use.{...}

DPE:

In today's rapidly advancing technological era, there is an ongoing debate regarding the best approach to address the critical environmental issues that our planet faces. Some argue that adopting a simple, minimalistic lifestyle is the key to preserving the environment, while others contend that technology itself provides the solutions needed. In my opinion, a combination of both strategies is necessary in order to effectively tackle these complex challenges. By adopting a mindful and sustainable lifestyle, and carefully implementing and regulating technology, we can reduce individual consumption and waste production, minimizing our environmental footprint while still meeting the demands of modern society.

{...}Adopting a more mindful and sustainable lifestyle can indeed contribute significantly to reducing individual consumption and waste production, thereby resulting in a smaller environmental footprint. Evidence shows that making sustainable choices such as reducing meat consumption, using public transportation, and minimizing energy usage can reduce an individual's carbon footprint by up to 25%. Additionally, cities like San Francisco have successfully implemented sustainable initiatives, resulting in diverting 80% of waste from landfills through a comprehensive recycling and composting program. While it is true that individual actions alone cannot fully address the complex nature of our modern society and the scale of global consumption and production, dismissing the impact of adopting a mindful and sustainable lifestyle is not the solution. Every small action counts, and when multiplied by millions of individuals, a significant difference can be made. Moreover, adopting a sustainable lifestyle can inspire others and create a ripple effect, leading to a collective effort towards environmental sustainability. Rather than waiting for structural changes and systemic interventions, individuals can take immediate action and contribute to reducing their environmental footprint. It is not a matter of one replacing the other, but rather a combination of individual actions and systemic changes that can bring about meaningful and lasting impact. By adopting a more mindful and sustainable lifestyle, individuals can play an active role in addressing the environmental challenges we face.{...}

Figure 21: Sample output of different models

Prompt:这两年，走在大街上，尤其是在景区，汉服的“曝光率”可以说是越来越高了。天猫平台数据显示，过去3年，汉服品类销售额增长超过6倍。在B站，2019年国风爱好者达到8347万人，其中83%为24岁以下的年轻人。越来越多的年轻人青睐汉服，这既为传统文化注入了活力，也是一股新的经济力量。你如何看待或思考上述现象?

E2E:

近年来，汉服在大街小巷的曝光率越来越高，尤其是在景区中更是随处可见。这一现象是不可忽视的，因为汉服的迅速普及在传承和发展传统文化的同时，也为经济带来了新的助力。首先，越来越多的年轻人选择穿汉服，为传统文化注入了新的活力。传统文化作为一个国家的瑰宝，承载着历史、艺术和价值观等多个层面。然而在快节奏现代生活的冲击下，传统文化渐渐被边缘化，甚至有些传统仪式、节日逐渐淡出人们的生活。而汉服的流行可以看作是年轻人对传统文化的热爱和渴望的表现，他们通过穿汉服来表演、拍照、参加活动，让传统文化焕发新的生机。年轻人的行为为传统文化注入了年轻的元素，使其在新时代焕发光彩。{...}然而，汉服的迅速流行也带来了一些问题需要引起关注。一些商家盲目跟风，推出质量低劣的产品来满足市场需求，甚至出现了商标侵权等问题。此外，一些年轻人追求时尚、纯粹追求照片的美感，而忽视了对传统文化的深入了解和内化。因此，我们需要在推动汉服流行的同时，加强对品质和文化内涵的把控，以免流行变成了纯粹的商业操作，而忽略了传统文化的核心价值。{...}

CoT:

随着时间的推移，汉服在年轻人中的受欢迎程度越来越高。这一现象不仅给传统文化注入了活力，也成为了一股新的经济力量。对于这个现象，我持积极的态度并愿意进行深入思考。首先，我们需要分析汉服在年轻人中的受欢迎程度原因。汉服以其华丽的外观、独特的设计和对传统文化的致敬而吸引了大批年轻人。在当今快节奏的生活中，年轻人们渴望一种与众不同的体验，而穿汉服就成为了他们展示个性和独特审美的方式。此外，互联网的发展也为汉服的传播提供了更多机会，各种社交媒体平台成为他们了解、分享和交流汉服文化的重要窗口。其次，汉服的活力对传统文化的注入至关重要。年轻人对汉服的热爱促使传统文化焕发出新的活力。传统文化是一个国家和民族的骄傲，通过年轻人的参与和传承，我们能够更好地保护和传承我们的传统文化。汉服的兴起不仅让更多人了解古代服饰的魅力，也推动了传统文化的创新和发展，使其与现代生活相结合，更符合年轻人的审美追求。{...}

ToT:

近年来，年轻人对中国传统文化的关注度明显提升，其中汉服现象备受瞩目。汉服作为中国传统服饰的代表，正逐渐占据年轻人的时尚圈和生活方式。这一现象不仅为传统文化注入了活力，还成为了新的经济力量。对于这一现象，我认为需要从多个方面进行思考和分析。首先，汉服现象背后的复兴源于年轻人对传统文化的热爱。在当今社会，年轻人随着教育水平的提高，开始重新审视中国传统文化的深厚底蕴。汉服作为古代中国的传统服饰，具有独特的历史和文化意义。年轻人通过穿汉服，不仅能够感受到传统文化的独特魅力，还能够表达对中华传统文化的尊重和热爱。此外，汉服的流行也受益于传统文化的宣传和推广，各类文化活动和节日庆典的举办，进一步激发了年轻人对传统文化的兴趣。{...}然而，对于汉服现象也需保持理性思考。一方面，我们应注意传统文化的商业化和变形问题。传统文化应当保持其独特的历史和文化价值，而不是被简单地转化为商品来迎合市场需求。另一方面，我们也要警惕传统文化的过度商业化可能对其本身产生的影响。传统文化应该扎根于精神内涵，而非单纯追求功利性的外在表象。{...}

DPE:

近年来，越来越多的年轻人选择青睐汉服，不仅在大街上穿着展现，而且在社交媒体平台上频繁出现。据天猫平台数据显示，汉服品类销售额在过去3年内增长超过6倍，而B站上国风爱好者数量更是达到了惊人的8347万人。这样庞大的数字不仅说明了汉服热潮的兴起，更是突显了年轻人对传统文化的热爱。青睐汉服的年轻人既促进了汉服产业链的蓬勃发展，也通过穿着和宣传，有效维护了中国传统服饰设计的保护和传承。这种现象不仅推动了相关产业的发展，也有力地促进了传统文化的传播和传承，对经济增长起到了积极的推动作用。因此，汉服热潮的兴起不仅为传统文化注入了活力，而且成为推动文化产业复兴的重要力量。然而有人认为汉服热潮只是暂时的时尚风尚难以对文化产业的复兴产生长期的影响，这样的观点是不全面的。首先，汉服热潮已经持续了相当长的时间，并且仍然在社交媒体和线下活动中受到广大群众的喜爱和参与。事实上，汉服的热潮在中国社会中已形成了较为稳定的文化现象，这是不容忽视的持久性。其次，汉服热潮对文化产业的复兴带来了显著的经济效益。许多企业和设计师推出了与汉服相关的产品，在市场上取得了巨大的成功。这些产品不仅包括服装和饰品等消费品，还涉及到了化妆品、餐饮等多个领域。对于文化产业来说，这是一个重要的发展动力。同时，汉服热潮也引发了文化教育的重视，相关书籍、教育机构和线上线下活动得到了更多的关注和参与，为文化产业的进一步发展提供了机遇。{...}

Figure 22: Sample output of different models