Plan–Write–Refine: An Iterative Framework for Argumentative Essay Generation

Zhijun Zhang[1,2], Qianlong Wang[1,*], Jianzhu Bao[3], Keyang Ding[4], Genan Dai[1], Bowen Zhang[1], Ruifeng Xu[4,5], Yongsheng Liang[1,2,4]

[1] School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China
[2] College of Applied Science, Shenzhen University, Shenzhen, China
[3]Nanyang Technological University, Singapore
[4]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[5]Pengcheng Laboratory, Shenzhen, China
2510263003@mails.szu.edu.cn, wangqianlong@sztu.edu.cn

Abstract

Argumentative essay generation (AEG) aims to produce coherent and persuasive essays on controversial claims. Most existing methods follow a *plan-then-write* paradigm, where a single plan is generated and used to guide long-form writing. However, the quality of the generated essay heavily depends on the initial plan, and errors or omissions in planning often propagate to the final output. To address this limitation, we propose the *Plan–Write–Refine* framework, an iterative approach for AEG. It first generates an initial plan and a draft essay, then abstracts the essay to obtain condensed feedback, and refines the plan accordingly. Compared with the prior static planning methods, our framework enables iterative plan improvement and better structural organization of the essay by closing the loop between planning and writing. We conduct extensive experiments on four AEG benchmarks[1]. Both automatic and human evaluations show that our framework achieves consistent improvements over strong baselines.

1 Introduction

Argumentative essay generation (AEG) is an important task in computational argumentation, aiming to automatically produce long-form essays with coherent structure and persuasive reasoning on claims (Carenini and Moore, 2006; Bao et al., 2022b).

Most existing approaches to AEG are built upon large language models (LLMs) and follow the *plan-then-write* paradigm (Yao et al., 2019; Liang et al., 2024; Hu et al., 2025b). Under this paradigm, LLMs first generate a high-level content plan for a given claim, and then produce the full argumentative essay by expanding this plan into an essay. However, the generated plan in these methods is fixed and remains unchanged throughout the entire writing procedure (He et al., 2024; Hu et al., 2025b), which could potentially lead to suboptimal outcomes. Since the plan serves as the core blueprint of AEG, any omission or logical weakness in the plan will be propagated to the final essay (Hu et al., 2025a). Such limitations may result in incomplete arguments, weak logical connections, or insufficient support for claims, and they cannot be corrected once the writing stage begins.

Human writers rarely follow an initial plan rigidly (Flower and Hayes, 1981). During writing, they often revisit and revise the plan after observing its abstract (Hayes and Flower, 1980). The abstract provides a condensed view of the argument structure and may reveal issues such as missing links or weak claims. By refining the plan based on this abstraction, writers can better reorganize arguments and progressively improve coherence and persuasiveness.

Inspired by this, we revisit argumentative writing from a new perspective and propose an *iterative Plan–Write–Refine framework* for AEG. Specifically, the framework comprises three steps. First, given an input claim, we use an LLM to generate an initial plan. Second, conditioned on the initial plan, the LLM generates a draft argumentative essay. Third, we prompt the LLM to generate an abstract from the draft essay and use this abstract to refine the plan. Also, these three steps can be executed iteratively, progressively strengthening the plan. Crucially, unlike prior *plan-then-write* (Hu et al., 2025b) or *self-refinement* methods (Wan et al., 2025) that revise drafts directly at the surface text level, we introduce abstraction-based feedback for plan refinement that enables refinement to operate at the argumentative structure, effectively closing the loop between planning and AEG. Moreover, compared with a single static plan, the refined plan in our framework is more informative as it incorporates feedback from the essay itself. As a result, the quality of the generated argumentative essay is consistently improved in coherence and persuasive

* Corresponding author.
[1]Code is available at https://github.com/ZhangZJ-GH/PWR-for-AEG

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1440–1460
November 6-10, 2026 ©2026 Association for Computational Linguistics

effectiveness.

We conduct extensive experiments on four public AEG datasets, namely ArgEssay, NYT-Editorial in English, and CHE-Essay, CHE-Editorial in Chinese. The results show that, compared with strong LLM-based baselines, our framework achieves superior overall performance.

Our contributions can be summarized in three aspects:

- We propose a novel *Plan–Write–Refine* paradigm for AEG, which models argumentative writing as an iterative process of planning and generation. Unlike prior *plan-then-write* approaches that rely on a fixed plan, our framework enables dynamic interaction between planning and writing.
- We introduce an abstraction-based feedback mechanism that refines plans using essay-level summaries. This mechanism operates at the level of argumentative structure rather than surface text, allowing the model to improve coherence more effectively.
- Extensive experiments on four benchmarks demonstrate the effectiveness of our framework. Compared with strong baselines, it achieves consistent improvements across datasets, with average gains of 1.00%, 1.72%, 2.76%, and 2.43% in automatic evaluation metrics, validating its effectiveness.

2 Related Works

2.1 Argument Mining

Argument Mining (AM) aims to identify argumentative components and the logical relations between them in text, making it a multifaceted task in natural language processing (Lippi and Torroni, 2015; Bao et al., 2022a). Early studies proposed annotation schemes to support manual labeling of argumentative structures (Das and Stede, 2018). Subsequently, with the advancement of neural network technologies, researchers began exploring end-to-end argument mining approaches (Eger et al., 2017; Bao et al., 2021) and investigated simplifying subtasks to improve overall performance (Morio et al., 2020). In recent years, some studies have integrated various subtasks into a general AM model to enhance its generalization ability (Farzam et al., 2024; Bao et al., 2025).

2.2 Argument Generation

Argument Generation (AG) refers to the automatic generation of high-quality argumentative texts that meet specific requirements based on given information, such as topic, stance, and context (Park et al., 2019). It works in conjunction with tasks like argument mining to satisfy a broader range of practical applications. Depending on the specific scenario, the target texts of argument generation can be categorized into two types: *argument generation in dialogue scenarios and in monologue scenarios*.

Argumentation analysis in dialogue scenarios aims to identify and analyze the relations between the arguments of the two parties. (Cheng et al., 2020). Existing studies primarily approach this by decomposing the task into subtasks, such as argument extraction and relation identification (Cheng et al., 2021). In recent years, some researchers have attempted to use LLMs combined with chain-of-thought techniques to accomplish the task (de Wynter and Yuan, 2023).

Within monologue scenarios, Argumentative Essay Generation (AEG) serves as the most representative task. An AEG system is required to generate structurally complete argumentative essays that include a clear central thesis. Some earlier generation methods attempted to directly learn the mapping from input prompts to full texts (Park et al., 2019; Chang et al., 2024). However, the generated texts often lacked consistency and stability. In recent years, significant progress has also been made in AEG. For example, graph structures have been employed to model semantic relations in long documents to improve long-text quality (Wu et al., 2021; Sun et al., 2026). Moreover, as LLMs have demonstrated increasingly strong performance on generation tasks, numerous studies have begun exploring their use for long-text generation and attempting to integrate them with the *plan-then-write* paradigm to further enhance performance (Fierro et al., 2024; Xiong et al., 2025; Wang et al., 2025).

While yielding promising results, existing methods for generating argumentative essays under the plan-then-write paradigm often overlook the intrinsic characteristics of argumentation. Besides, their intermediate steps rely on fixed planning content, limiting the potential for boosting argument quality.

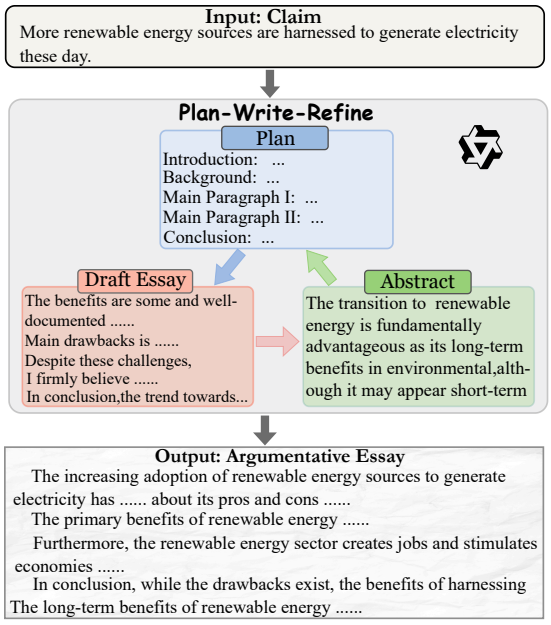


Figure 1: An overview of our framework. It illustrates the three-step iterative process: generating an initial plan, producing a draft argumentative essay, and refining the plan based on feedback from the essay's abstract.

3 Method

3.1 Overview

Given an input claim X, our framework generates an argumentative essay Y with an iterative framework, following the *Plan–Write–Refine* paradigm. Specifically, it involves three key steps: generating an initial plan based on the claim, using the plan to produce a draft argumentative essay, and refining the plan using the essay's abstract. The overall framework is shown in Figure 1.

3.2 Initial Plan Generation

This step involves generating an initial argumentative plan from a given input claim X. Here, we use an LLM to identify key content in each part (*e.g.*, *Introduction* and *Background*) related to the claim. LLM is prompted to organize these contents in a coherent and logical hierarchy, ensuring that the plan provides a clear outline. The plan P_0 is:

$$P_0 = LLM(X, plan_prompt) \quad (1)$$

3.3 Argumentative Essay Generation

After generating the initial plan P_0, this step is to use this plan to produce a draft argumentative essay. We employ an LLM to generate the essay Y_0 by using the initial plan P_0 as the primary prompt. The model is guided to expand upon each point in the plan. The AEG can be expressed as follows:

$$Y_0 = LLM(P_0, essay_prompt) \quad (2)$$

3.4 Plan Refinement with Abstract

After generating the initial argumentative essay Y_0, we refine the plan based on the essay's abstract. The motivation for this refinement lies in the fact that human writers often revisit and revise their plans after seeing their work's abstract. Similarly, in our framework, the abstract of the essay provides a condensed view of the argumentative essay, showing missing evidence or invalid points.

To generate the abstract, we prompt the LLM to extract the key points and core content from the draft essay Y_0. This abstract is denoted as A_0:

$$A_0 = LLM(Y_0, abstract_prompt) \quad (3)$$

Then, the plan P_0 is revised based on the feedback derived from the abstract, addressing identified weaknesses and enhancing the overall structure. The refined plan P_1 can be represented as:

$$P_1 = Plan_Refinement(P_0, A_0) \quad (4)$$

The key benefit of this refinement step is that it ensures the plan evolves dynamically, adapting to the issues revealed in the essay's abstract.

3.5 Iterative Process

After each cycle, the refined plan P_i is used to generate a new draft essay Y_i, which is then summarized into an abstract A_i. The abstract serves as feedback to further refine the plan, addressing any remaining gaps or weaknesses. This loop continues until a predefined number of iterations is reached. The final argumentative essay Y_n is then produced based on the last refined plan P_n, ensuring a high-quality output with a well-structured, logically sound, and persuasive argument.

4 Experiments

4.1 Experimental Setting

Datasets. We conduct experiments on four widely used argumentative essay datasets: ArgEssay (Bao et al., 2022b), NYT-Editorial, CHE-Essay, and CHE-Editorial (He et al., 2024). These datasets focus on two types of argumentative essay-writing, namely exam essays and news editorials, cover both English and Chinese.

Among them, ArgEssay and CHE-Essay consist of essays collected from English standardized examinations and China's National College Entrance Examination. CHE-Editorial and NYT-Editorial are collected from the comment sections of Chinese news websites and the New York Times in

the United States, respectively. Additional details about dataset construction and statistics are provided in Appendix A.1.

Implementation Details. We conducted experiments with Llama3.1-8B-instruct and Qwen2.5-7B-Instruct for English and Chinese, respectively. In both models, we set the iteration number and top-p to 3 and 0.9, respectively. The number of iterations is set to 3. And they were deployed using vLLM (Kolluru, 2025; Kwon et al., 2023) for its efficient inference acceleration. More implementation details are in Appendix A.2.

Baselines. We benchmark our framework against several competitive baselines to assess its performance. Their descriptions are as follows: **E2E** (Gretz et al., 2020) directly prompts LLM to generate an essay. **DD-KW** (Bao et al., 2022b) employs a dual-decoder architecture for planning and generation. **CoT** (Wei et al., 2022) conducts a brief planning before generating the article. **ToT** (Yao et al., 2023) explores multiple reasoning trajectories and selects the optimal plan. **DPE** (He et al., 2024) uses a two-stage hierarchical planning strategy. **D2W** (Hu et al., 2025b) conducts the multi-agent debates for essay generation. **DraftRef** (Hu et al., 2024) iteratively evaluates and refines the generated essay draft. **DARE** (Zhao et al., 2025) introduces dynamic rhetorical expression.

Additionally, we defined a self-refine method **ReEssay**, which directly uses the generated essay itself as the refinement subject, and a feedback refinement method **SumRefine**, which replaces the abstract with an essay summary during plan refinement. Further details are given in Appendix A.3.

4.2 Evaluation Metrics

Automatic Evaluation Automatically evaluating the quality of argumentative essays is a challenging task, as traditional string-matching metrics like BLEU and ROUGE are ineffective in measuring argumentative depth, logical coherence, and overall discourse quality (Celikyilmaz et al., 2020; Hu et al., 2024). To overcome this limitation, we follow the recent paradigm of LLM-based evaluation. It adopts an automated assessment framework centered on GPT-4o, aiming to obtain more consistent and human-aligned evaluation results (Zheng et al., 2023; Bavaresco et al., 2025). Specifically, we quantitatively evaluate argumentative essays along the four dimensions:

- **Relevance (Rel.)**: examining the alignment between the essay content and the given claim, as well as the consistency of viewpoints.
- **Logic (Log.)**: assessing the clarity of the discourse structure, the rigor of the reasoning chain, and the strength of evidence supporting the claims.
- **Complexity (Comp.)**: focusing on the depth of argumentation layers, the diversity of argument development, and the coverage of different stances and perspectives.
- **Persuasiveness (Pers.)**: measuring the effectiveness of the discourse in influencing readers and achieving argumentative goals within a specific context.

For the GPT-4o-based automatic evaluation, each dimension is scored on a scale from 1 to 5, with scores reported to three decimal place. To enhance the reliability and interpretability of the evaluation, GPT-4o generates detailed written analyses before outputting scores. All automatic evaluation experiments were conducted independently three times, and the final reported score is the average of the three evaluation results. Figure 7 in the Appendix shows the prompt for automatic evaluation.

Human Evaluation To conduct a more comprehensive analysis, we performed a human evaluation to complement the automatic assessment. Three graduate students with backgrounds in linguistics or related fields were recruited as volunteers to independently score the generated essays. They followed the same four evaluation dimensions used in the automatic assessment. Each evaluator was provided with detailed scoring guidelines to ensure consistency and reliability.

Additionally, we assess agreement among human annotators and between human annotators and the LLM judge to further validate the reliability of our evaluation results. Detailed results are provided in Appendix B.6.

4.3 Main Results

We present the main results in Table 1. Overall, our framework achieves consistent improvements across both automatic and human evaluations on four datasets.

(1) Compared to the SOTA baseline DARE, the average scores on automatic evaluation increased by 1.00%, 1.72%, 2.76%, and 2.43% across the four datasets.

DataSet	Method	Automatic Evaluation					Human Evaluation				
		Rel.	**Log.**	**Comp.**	**Pers.**	**Avg.**	**Rel.**	**Log.**	**Comp.**	**Pers.**	**Avg.**
ArgEssay	E2E	4.426	4.218	3.784	3.946	4.094	4.550	4.404	3.318	4.410	4.171
	DD-KW	2.260	1.796	1.620	1.612	1.822	2.300	2.000	1.650	1.700	1.913
	CoT	4.500	4.142	3.690	3.872	4.051	4.560	4.384	4.286	4.380	4.403
	ToT	4.486	4.228	3.842	4.144	4.175	4.600	4.460	4.450	4.420	4.483
	DPE	4.488	4.150	4.132	4.148	4.230	4.486	4.326	4.424	4.377	4.403
	D2W	4.518	4.140	3.986	3.876	4.130	4.500	4.160	4.130	3.920	4.178
	DraftRef	4.587	4.269	4.258	4.087	4.300	4.566	4.302	4.371	4.180	4.355
	DARE	4.631	4.341	**4.329**	4.300	4.400	4.640	4.456	4.517	4.496	4.527
	ReEssay	4.622	4.286	4.214	4.190	4.328	4.619	4.424	4.433	4.398	4.469
	SumRefine	4.632	4.254	4.204	4.219	4.327	4.653	4.451	4.403	4.475	4.496
	Ours	**4.711**	**4.400**	4.312	**4.351**	**4.444**	**4.732**	**4.603**	**4.539**	**4.605**	**4.620**
NYT-Editorial	E2E	4.522	4.222	3.834	3.938	4.129	4.240	3.934	3.786	3.806	3.942
	DD-KW	2.184	1.764	1.742	1.618	1.827	2.210	1.830	1.730	1.645	1.854
	CoT	4.680	4.246	3.762	3.930	4.155	4.140	3.940	3.776	3.791	3.912
	ToT	4.624	4.232	3.888	4.034	4.195	4.251	4.022	3.882	3.928	4.021
	DPE	4.612	3.914	4.036	3.872	4.109	4.414	4.199	4.107	4.092	4.203
	D2W	4.588	4.120	3.908	3.916	4.133	4.520	4.131	4.120	4.020	4.198
	DraftRef	4.623	4.185	4.132	3.884	4.206	4.536	4.214	4.303	4.152	4.301
	DARE	4.674	4.252	**4.200**	4.100	4.307	4.548	4.328	4.418	4.210	4.376
	ReEssay	4.643	4.279	4.095	4.148	4.291	4.553	4.242	4.464	4.347	4.402
	SumRefine	4.653	4.247	4.083	4.177	4.290	4.577	4.269	4.434	4.390	4.418
	Ours	**4.732**	**4.392**	4.191	**4.207**	**4.381**	**4.688**	**4.404**	**4.502**	**4.443**	**4.509**
CHE-Essay	E2E	4.690	4.308	3.822	3.960	4.195	4.196	4.006	3.844	3.804	3.963
	DD-KW	2.176	1.706	1.614	1.486	1.746	2.230	1.810	1.642	1.530	1.803
	CoT	4.700	4.244	3.696	3.882	4.131	4.191	3.979	3.790	3.745	3.926
	ToT	4.686	4.244	3.772	4.000	4.176	4.256	4.096	3.953	3.940	4.061
	DPE	4.666	4.256	4.104	4.106	4.283	4.350	4.265	4.201	4.254	4.268
	D2W	4.604	4.232	3.968	3.962	4.192	4.340	4.252	4.186	4.230	4.252
	DraftRef	4.659	4.237	4.064	4.151	4.278	4.322	4.286	4.261	4.137	4.252
	DARE	4.716	4.270	4.146	4.228	4.340	4.376	4.322	4.322	4.354	4.344
	ReEssay	4.664	4.247	4.224	4.123	4.315	4.499	4.253	4.460	4.347	4.390
	SumRefine	4.776	4.304	4.272	4.206	4.390	4.551	4.319	4.502	4.499	4.468
	Ours	**4.811**	**4.388**	**4.356**	**4.284**	**4.460**	**4.619**	**4.412**	**4.557**	**4.578**	**4.542**
CHE-Editorial	E2E	4.804	4.348	3.798	3.980	4.233	4.337	4.081	3.943	3.918	4.070
	DD-KW	2.144	1.678	1.728	1.550	1.775	2.120	1.712	1.810	1.530	1.793
	CoT	4.764	4.222	3.678	3.866	4.133	4.328	4.059	3.890	3.893	4.043
	ToT	4.740	4.298	3.766	3.998	4.201	4.428	4.121	4.012	4.024	4.146
	DPE	4.708	4.240	4.136	4.114	4.300	4.584	4.331	4.328	4.240	4.371
	D2W	4.700	4.282	4.056	3.954	4.248	4.530	4.360	4.284	4.030	4.301
	DraftRef	4.733	4.315	3.922	4.172	4.286	4.624	4.379	4.351	4.041	4.349
	DARE	4.774	4.346	4.148	4.280	4.387	4.687	4.393	4.478	4.375	4.483
	ReEssay	4.688	4.362	4.296	4.295	4.410	4.679	4.330	4.551	4.358	4.480
	SumRefine	4.740	4.360	4.344	4.278	4.431	4.705	4.399	4.592	4.407	4.526
	Ours	**4.810**	**4.436**	**4.408**	**4.362**	**4.504**	**4.786**	**4.441**	**4.638**	**4.477**	**4.586**

Table 1: Main results. **Avg.** represents the average of scores across the four metrics. **Bold** numbers denote the best performance among all methods on each dataset. The significance test results indicate that all improvements are statistically significant with $p < 0.05$.

(2) The improvements are especially pronounced on the CHE-Editorial dataset. Our framework surpasses prior SOTA methods across all metrics, indicating that it generalizes well beyond English.

(3) Our framework consistently achieves performance gains over DraftRef, ReEssay and SumRefine. This indicates that abstract-level feedback provides a more effective and structurally informative signal for plan refinement than direct essay-level revision, draft-level iterative refinement, or summary-based plan refinement. These results further show that the effectiveness of our framework comes not merely from iterative refining a certain step, but from refining the plan with abstract-based

DataSet	Model	Automatic Evaluation					Human Evaluation				
		Rel.	Log.	Comp.	Pers.	Avg.	Rel.	Log.	Comp.	Pers.	Avg.
ArgEssay	**Ours**	**4.711**	**4.400**	**4.312**	**4.351**	**4.444**	**4.732**	**4.603**	**4.539**	**4.605**	**4.620**
	w/o Abstract	4.644	4.290	4.214	4.196	4.336	4.654	4.397	4.409	4.343	4.451
	w/o Refine	4.552	4.176	3.962	4.052	4.186	4.573	4.275	4.026	4.183	4.264
NYT-Editorial	**Ours**	**4.732**	**4.392**	**4.191**	**4.207**	**4.381**	**4.688**	**4.404**	**4.502**	**4.443**	**4.509**
	w/o Abstract	4.688	4.302	4.078	4.154	4.306	4.682	4.297	4.411	4.405	4.449
	w/o Refine	4.591	4.196	3.822	3.920	4.132	4.514	4.059	4.128	4.208	4.227
CHE-Essay	**Ours**	**4.811**	**4.388**	**4.356**	**4.284**	**4.460**	**4.619**	**4.412**	**4.557**	**4.578**	**4.542**
	w/o Abstract	4.766	4.212	4.186	4.190	4.339	4.522	4.213	4.391	4.304	4.358
	w/o Refine	4.558	4.108	3.826	3.972	4.116	4.401	4.055	3.973	4.116	4.136
CHE-Editorial	**Ours**	**4.810**	**4.436**	**4.408**	**4.362**	**4.504**	**4.786**	**4.441**	**4.638**	**4.477**	**4.586**
	w/o Abstract	4.792	4.351	4.317	4.272	4.433	4.618	4.200	4.422	4.353	4.398
	w/o Refine	4.744	4.170	4.031	3.852	4.199	4.401	4.098	4.002	3.996	4.124

Table 2: Ablation study. The significance test indicates that all differences are statistically significant with $p < 0.05$.

feedback.

(4) The conclusions from human evaluation align with those of automatic evaluation. Here, annotators consistently prefer essays from our framework, particularly in persuasiveness, with relative gains of up to 2.86% over strong baselines. This shows that by refining the plan using feedback abstracted from the essay, our framework enables more persuasive discourse, validating its effectiveness.

(5) We performed significance tests on the average scores of all four evaluation metrics, comparing the proposed framework with the strongest baseline (DARE) under both automatic and human evaluation settings. And in all cases we found that the improvements are statistically significant with $p < 0.05$. This indicates that the performance gains of our framework reflect consistent and reliable improvements over strong baselines.

4.4 Ablation Study

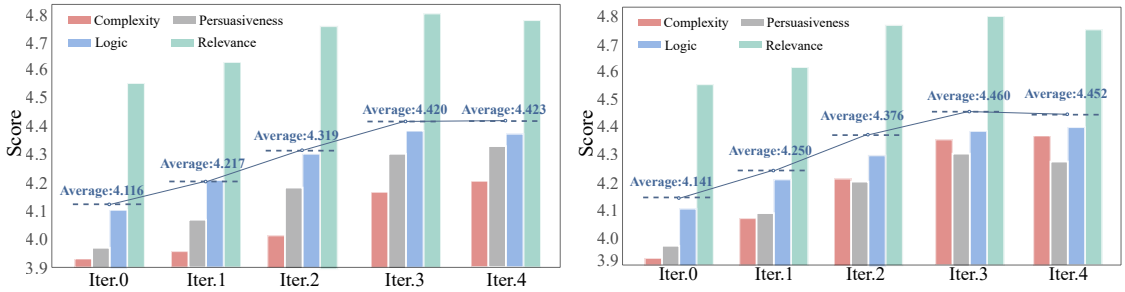
The ablation results are presented in Table 2. First, "*w/o* Abstract" means that we retain the *plan refinement* step but omit the abstract. We observe that LLM incurs a certain degree of performance penalty, particularly on logic and complexity metrics. This indicates that using the abstract as a feedback signal is important, since it helps LLM better identify weaknesses in the plan. Second, "*w/o* Refine" means that LLM generates a plan and directly produces the essay as the final output. Experimental results show that each metric on both datasets declines substantially. Additionally, we found that, compared with the approach that merely excludes the summary from the refinement process, completely abandoning the refinement of the plan leads to a more pronounced decline in performance. This phenomenon is reflected across all metrics in both automatic and human evaluations. These results further validate the claim proposed in this study that refining the plan can substantially improve the quality of generated essay. In addition, compared with the other three metrics, "complexity" is more strongly affected in the *w/o* Refine setting, indicating that improvements in the complexity of generated articles benefit more substantially from the refinement stage.

In this section, we also conducted significance tests between the proposed framework and each ablation variant. The result indicates that all differences are statistically significant with $p < 0.05$.

4.5 Analysis

Effect of Iteration Number on Performance. We investigated how the number of refinement iterations affects performance. Figure 2 illustrates the trends across different iterations on ArgEssay and CHE-Essay datasets. The results shows that in both datasets, as iterations increase, all automatic evaluation metrics improve steadily, validating the effectiveness of iteration refinement. Improvements are largest in early iterations (Iter. 1→Iter.2), suggesting that the procedure rapidly enhances argument coherence. The gains exhibit diminishing returns in later iterations (Iter. 3→Iter.4). In the CHE-Essay dataset, there is even a downward trend. These empirical patterns motivate our choice of three iterations as a practical trade-off between effectiveness and computational efficiency.

Efficiency Analysis. We further analyze the computational efficiency of different methods. Table 4 reports the efficiency results on ArgEssay and CHE-Essay. The **Time** (hours) shown in the table



(a) The results on the ArgEssay dataset.

(b) The results on the CHE-Essay dataset.

Figure 2: Performance comparison of different rounds.

Model	Method	ArgEssay					CHE-Essay				
		Rel.	**Log.**	**Comp.**	**Pers.**	**Avg.**	**Rel.**	**Log.**	**Comp.**	**Pers.**	**Avg.**
Mistral-7B	D2W	4.397	4.026	3.797	3.771	3.998	4.077	3.875	3.992	4.010	3.989
	DARE	4.521	4.322	4.189	4.243	4.319	4.421	4.096	4.031	4.098	4.162
	Ours	**4.690**	**4.401**	**4.294**	**4.256**	**4.410**	**4.582**	**4.233**	**4.016**	**4.137**	**4.242**
LLama3.1-8B	D2W	4.518	4.140	3.986	3.876	4.130	4.502	4.111	4.009	3.993	4.154
	DARE	4.631	4.341	**4.329**	4.300	4.400	4.634	4.377	4.293	4.207	4.378
	Ours	**4.711**	**4.400**	4.312	**4.351**	**4.444**	**4.772**	**4.416**	**4.384**	**4.268**	**4.460**
ChatGLM4-9B	D2W	4.565	4.104	4.005	3.992	4.167	4.607	4.204	4.298	4.131	4.310
	DARE	4.652	4.297	4.083	4.193	4.306	4.683	4.317	4.442	4.396	4.460
	Ours	**4.786**	**4.370**	**4.202**	**4.286**	**4.411**	**4.790**	**4.498**	**4.602**	**4.522**	**4.603**
Qwen3-14B	D2W	4.755	4.314	4.443	4.164	4.419	4.677	4.382	4.413	4.145	4.404
	DARE	4.732	4.403	4.598	4.493	4.557	4.759	4.501	4.505	4.395	4.540
	Ours	**4.856**	**4.516**	**4.616**	**4.586**	**4.644**	**4.878**	**4.584**	**4.666**	**4.568**	**4.674**

Table 3: Robustness analysis across different LLMs. Here, all metrics are obtained from the automatic evaluation.

refers to the total running time over all samples, measured as the wall-clock time elapsed from the start of inference on the first sample to the completion of inference on the last sample. As shown in Table 4, our framework produces the largest number of tokens and achieves the best average performance on both automatic and human evaluations. Importantly, despite its iterative nature, the overall running time remains well within an acceptable range. It is substantially lower than that of more complex plan-then-write-based methods (DPE and DARE). This shows that iterative plan refinement guided by abstract feedback improves argumentative quality without incurring prohibitive computational costs.

DataSet	Method	Time	Avg.Token	Avg.Auto	Avg.Human
ArgEssay	CoT	0.98H	447	4.051	4.403
	D2W	1.67H	579	4.130	4.178
	DPE	2.92H	457	4.230	4.403
	DARE	8.58H	994	4.400	4.527
	Ours	1.62H	1268	**4.444**	**4.620**
CHE-Essay	CoT	0.88H	436	4.131	3.926
	D2W	2.05H	823	4.192	4.252
	DPE	2.57H	965	4.283	4.268
	DARE	4.08H	1313	4.340	4.344
	Ours	1.87H	1324	**4.460**	**4.542**

Table 4: Efficiency analysis. **Time** is the total running time (hours) over all samples, measured as the wall-clock time elapsed from the start of inference on the first sample to the completion of inference on the last sample. **Avg.Token** denotes the average number of generated tokens per essay. **Avg.Auto** and **Avg.Human** are the averaged scores across automatic and human evaluations, respectively.

Controlled Analysis of Computational Budget and Output Length To investigate whether the performance improvements of our method primarily stem from additional model calls or longer generated texts, we count the total number of LLM calls required by each method to produce the final essay, including planning, intermediate feedback generation, and essay generation. Meanwhile, we impose unified target-length constraints on all methods, requiring approximately 600 words for English essays and 1,000 Chinese characters for Chinese essays.

As shown in Table 5, our method requires only 11 model calls to generate a final essay, substantially fewer than DPE and DARE and only slightly more than D2W. Under the unified length con-

DataSet	Method	Model Calls	Avg.Token	Avg.Auto
ArgEssay	D2W	9	759	4.203
	DPE	24	679	4.270
	DARE	33	818	**4.387**
	Ours	11	831	4.369
CHE-Essay	D2W	9	912	4.305
	DPE	24	1063	4.251
	DARE	33	1103	4.298
	Ours	11	1038	**4.415**

Table 5: Length-controlled comparison with model-call statistics. Model Calls denotes the total number of LLM invocations required to produce the final essay.

straints, the actual output lengths of different methods become more comparable. In this setting, our method still achieves competitive performance on both datasets, obtaining the highest average automatic evaluation score on CHE-Essay while performing slightly below DARE on ArgEssay. These results demonstrate that our method achieves competitive generation performance with relatively few model calls and that its improvements cannot be explained solely by increased model calls or longer outputs. Additionally, we report the average number of generated tokens across different iterations in Appendix B.2.

Robustness Analysis. We conduct a robustness analysis across LLMs from different families and scales. Table 3 reports the results on both datasets[2]. The results show that, when combined with any of these base models, our framework consistently outperforms the D2W and the best baseline DARE on average scores. *This indicates that the observed improvements are attributable to the proposed method itself rather than to a particular choice of base model.* Moreover, when combined with the main results in Table 1 using Qwen2.5-7B, our framework maintains strong advantages across models of different scales within the same family. This shows that the proposed framework is robust to variations in backbone LLMs. Moreover, as the capabilities of LLMs continue to improve, the framework can also benefit from these advancements.

Abstract Quality Evaluation. To validate the effectiveness of abstraction-based feedback in plan refinement, we conduct a dedicated analysis of abstract quality. Table 6 reports the results across four metrics (see Appendix B.5). We compare our method with SumRefine, which produces longer summaries that emphasize local textual content. Our abstracts received higher scores than the summary, except for the informativeness metric (due to the longer content of the summary), indicating their high quality and effectiveness. This improvement stems from explicitly prompting the abstract to capture high-level argumentative structure rather than surface realizations.

[2]since the Mistral-7B model originally employed does not provide native support for Chinese, we instead used its Chinese fine-tuned variant, Chinese-Mistral-7B, for the experiments.

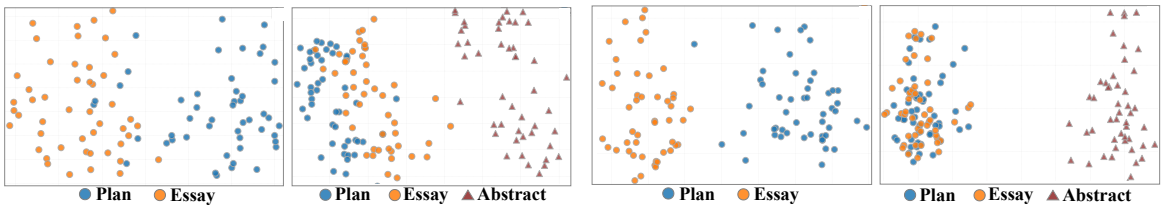
Dataset		Fact.	Info.	Con.	Cohe.
ArgEssay	Summary(SumRefine)	4.71	**4.24**	4.29	4.03
	Abstract(Ours)	**4.78**	3.98	**4.40**	**4.16**
CHE-Essay	Summary(SumRefine)	4.69	**4.18**	4.02	3.91
	Abstract(Ours)	**4.82**	4.05	**4.21**	**4.02**

Table 6: Abstract quality analysis. Four metrics are Factuality (Fact.), Informativeness (Info.), Contribution (Con.), and Coherence (Cohe.).

Visualization Analysis. We conduct a visualization analysis of the semantic relationships between plans, generated essays, and intermediate abstracts. Figure 3a and 3b present the visualization results on dataset CHE-Essay and ArgEssay, respectively, where representations are derived using BERTScore-based embeddings (Zhang et al., 2019). For baseline, following the *plan–write* paradigm, the plan and essay form two separate clusters. This suggests that a single static plan is insufficient to reliably guide long-form argumentative generation. In contrast, the refined plan and the essay in our framework are closely clustered. Meanwhile, the abstract forms a distinct cluster, reflecting its role as a higher-level structural representation. These observations show that iterative plan refinement effectively closes the gap between planning and generation, validating our motivation.

Argumentative Structure Analysis. To evaluate the framework's ability to optimize argumentative structure, we analyze essays in terms of both argumentative depth and breadth, where breadth reflects the number of distinct arguments and depth captures the relational complexity among them. Specifically, we annotate essays using a pre-trained argument mining (AM) model (Morio et al., 2022), and compute these structural metrics based on identified components and their logical relations.

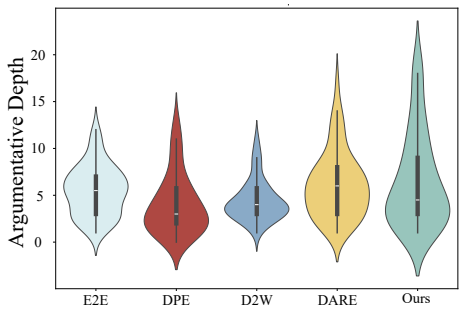
Figure 4 shows that our framework consistently achieves higher depth and breadth than all baselines on ArgEssay, indicating more comprehensive and better-connected argument structures. Com-



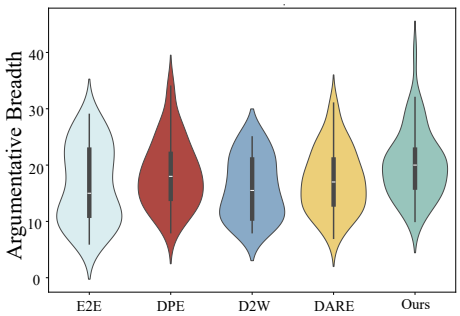
(a) Visualization on the CHE-Essay dataset.

(b) Visualization on the ArgEssay dataset.

Figure 3: Visualization Analysis



(a) Distribution of argument depth.



(b) Distribution of argument breadth.

Figure 4: Argumentative Structure Analysis.

pared with prior methods that rely on a fixed plan throughout generation, our framework refines the plan through abstraction-based feedback that operates at the argumentative structure rather than the surface text level. This allows the model to progressively enrich both the number of arguments and the connections among them, resulting in structurally stronger and more coherent argumentative essays.

Similarity to Human-Written Essays To further evaluate the similarity between generated essays and human-written essays, we employ MAUVE (Pillutla et al., 2021) to measure the distributional differences between essays generated by different methods and human-written essays. MAUVE measures the similarity between generated and human-written texts at the distributional level. We compare our method with CoT, D2W, and DARE across all four datasets.

As shown in Table 7, our method achieves the highest average MAUVE score across the four datasets, outperforming the compared methods on all datasets except NYT-Editorial, where it performs slightly below DARE. These results indicate that the essays generated by our method exhibit high distributional similarity to human-written essays. Furthermore, considering the differences in essay length and textual characteristics across datasets, we primarily use MAUVE to compare generation methods within each dataset rather than directly comparing absolute scores across datasets.

Method	ArgEssay	NYT-Editorial	CHE-Essay	CHE-Editorial	Avg.
CoT	0.612	0.586	0.684	0.661	0.636
D2W	0.628	0.598	0.701	0.676	0.651
DARE	0.647	**0.621**	0.719	0.695	0.671
Ours	**0.659**	0.615	**0.742**	**0.716**	**0.683**

Table 7: MAUVE comparison between generated andhuman-written essays across four datasets.Higher scores indicate greater distributionalsimilarity to human-written essays.**Avg.** denotes the average MAUVE score-across the four datasets.

5 Conclusion

In this paper, we propose *Plan–Write–Refine*, an iterative framework for AEG. Compared with prior *plan-then-write* methods, our framework refines the plan based on abstracting from the generated essay. Experimental results on four AEG datasets indicate the effectiveness of our framework. Both automatic and human evaluations show consistent improvements over competitive baselines.

Limitations

Despite the superior performance achieved by the proposed *Plan–Write–Refine* framework, two limitations remain.

- First, *the iterative process inevitably introduces additional computational overhead*, as multiple rounds of planning, writing, and abstraction are required. This extra cost is accompanied by consistent improvements in

argumentative structure and coherence. As shown in Table 4, although the additional computational overhead is acceptable, we suspect there is still room for reducing the overhead. In future work, we plan to explore more efficient refinement strategies, such as lightweight feedback mechanisms, to reduce computation cost without sacrificing performance.

- Second, although our framework demonstrates promising gains across different languages in AEG, *it does not explicitly account for cultural differences during either plan initialization or plan refinement*. Argumentation styles can vary substantially across cultural contexts, and ignoring these differences may limit the effectiveness of generated essays. We believe that incorporating culture-specific debating strategies into the planning stage could further enhance argumentative quality. This direction remains underexplored and represents an important avenue for future research.

Ethics Statement

Our work builds on LLMs, which may reflect biases and inaccuracies present in their training corpora and can produce misleading or fabricated content. To reduce such harms, we use only publicly available AEG datasets (ArgEssay, NYT-Editorial, CHE-Essay and CHE-Editorial) and comply with the original data and model licenses. We document dataset provenance and remove items judged to contain defamation, hate, or other inappropriate material.

Nonetheless, generated outputs can still be erroneous or harmful, so we require human review and fact-checking for any downstream use and do not advocate unattended deployment in safety-critical settings. We also encourage developers to apply technical and organizational safeguards to mitigate misuse, and to be transparent about limitations when releasing models or systems.

For this manuscript, AI tools were used only for grammar checking and text polishing.

Acknowledgments

This research was supported by the Natural Science Foundation for Top Talents of SZTU (Nos. GDRC202518 and GDRC202320).

References

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 ACL*, pages 6354–6364.

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 EMNLP*, pages 10437–10449.

Jianzhu Bao, Mohan Jing, Kuicai Dong, Aixin Sun, Yang Sun, and Ruifeng Xu. 2025. Uniasa: A unified generative framework for argument structure analysis. *Computational Linguistics*, 51(3):739–784.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022b. Aeg: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of EMNLP*, pages 5134–5148.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of ACL*, pages 238–255.

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

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

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of EMNLP*, pages 7000–7011.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of ACL*, pages 6341–6353.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. Instructeval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of SCALE-LLM*, pages 35–64.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Debopam Das and Manfred Stede. 2018. Developing the bangla rst discourse treebank. In *Proceedings of LREC*.

Adrian de Wynter and Tangming Yuan. 2023. " i'd like to have an argument, please": Argumentative reasoning in large language models. *arXiv preprint arXiv:2309.16938*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL*, pages 11–22.

Amirhossein Farzam, Shashank Shekhar, Isaac Mehlhaff, and Marco Morucci. 2024. Multi-task learning improves performance in deep argument mining models. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 46–58.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of ACL*, pages 11397–11417.

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

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of EMNLP*, pages 528–544.

John R Hayes and Linda S Flower. 1980. Identifying the organization of writing processes. *Cognitive processes in writing*, pages 3–30.

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 *Findings of ACL*, pages 12305–12322.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025a. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of ACL*, pages 32779–32798.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025b. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of COLING*, pages 4689–4703.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. Americano: Argument generation with discourse-driven decomposition and agent interaction. In *Proceedings of COLING*, pages 82–102.

Saicharan Kolluru. 2025. Comparative analysis of large language model inference serving systems: A performance study of vllm and huggingface tgi. *arXiv preprint arXiv:2511.17593*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of SOSP*, pages 611–626.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, and Wang. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *International Workshop on Theorie and Applications of Formal Argumentation*, pages 163–176. Springer.

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

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of ACL*, pages 3259–3266.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

ChaeHun Park, Wonsuk Yang, and Jong Park. 2019. Generating sentential arguments from diverse perspectives on controversial topic. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 56–65.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Proceedings of NeurIPS*, 34:4816–4828.

Matthew Renze, Erhan Guven, and 1 others. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of EMNLP*, pages 7346–7356.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.

Yang Sun, Yuqi Huang, Bin Liang, Min Yang, and Ruifeng Xu. 2026. Enhancing debate dialogue generation via dual-dimensional reflection and refinement. In *Proceedings of ICASSP*, pages 18117–18121. IEEE.

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. A cognitive writing perspective for constrained long-form text generation. *arXiv preprint arXiv:2502.12568*.

Chenglong Wang, Hang Zhou, Kaiyan Chang, Tongran Liu, Chunliang Zhang, Murun Yang, Quan Du, Tong Xiao, Yue Zhang, and Jingbo Zhu. 2025. Learning evaluation models from large language models for sequence generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2902–2917.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*, pages 24824–24837.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Ziqiang Cao, Sujian Li, Hua Wu, and Haifeng Wang. 2021. Bass: Boosting abstractive summarization with unified semantic graph. In *Proceedings of ACL*, pages 6052–6067.

Ruibin Xiong, Yimeng Chen, Dmitrii Khizbullin, Mingchen Zhuge, and Jürgen Schmidhuber. 2025. Beyond outlining: Heterogeneous recursive planning for adaptive long-form writing with language models. In *Proceedings of EMNLP*, pages 24689–24725.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of AAAI*, 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. In *Proceedings of NeurIPS*, pages 11809–11822.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xueguan Zhao, Wenpeng Lu, Chaoqun Zheng, Weiyu Zhang, Jiasheng Si, and Deyu Zhou. 2025. Plan dynamically, express rhetorically: A debate-driven rhetorical framework for argumentative writing. In *Proceedings of EMNLP*, pages 9562–9584.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, and Li. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of NeurIPS*, pages 46595–46623.

A Experimental Setting

A.1 Datasets

We conduct experiments using four argumentative essay datasets: **ArgEssay** (Bao et al., 2022b), **CHE-Essay**, **NYT-Editorial** and **CHE-Editoral** (He et al., 2024). These datasets are derived from two common genre of argumentative writing, namely exam essays and news editorials, and include both English and Chinese data. Table 8 provides statistical information about the datasets. Below is a detailed description of the four datasets.

ArgEssay The dataset is sourced from the online community Essay Forum, which contains argumentative essays for English exams such as IELTS and TOEFL, along with their corresponding prompts. Through a multi-step filtering and cleaning process, we ultimately constructed a high-quality English dataset comprising 11,282 prompt-essay pairs. The essays in this dataset are formally written and structurally complete, with an average length significantly longer than existing argument generation datasets, suitable for research and evaluation of long-text argument generation tasks.

CHE-Essay A Chinese exam argumentative essay dataset, whose data originates from the writing prompts and exemplary essays of China's National College Entrance Examination. This dataset contains 3,750 prompt-essay pairs, with an average essay length of 786.03 tokens and an average prompt length of 255.20 tokens, covering controversial topics such as society, education, culture, and technology. As the first long-text argument generation dataset tailored for the Chinese exam scenario, CHE-Essay fills the gap in non-English argumentative essay data.

NYT-Editorial An argumentative text generation dataset designed for the English news commentary scenario. It is sourced from the "Room for Debate" opinion section of The New York Times, containing 9,178 "prompt-essay" pairs. The essays have an average length of 419.32 tokens, with prompts averaging 114.17 tokens. The language is formal, the argumentation is in-depth, and the topics span diverse domains including society, politics, and culture. This dataset is authored by professional commentators, making it suitable for training and evaluating long-text generation models tailored to news commentary scenarios that require handling complex logical structures and multi-perspective

argumentation.

CHE-Editorial CHE-Editorial is an argumentative text generation dataset designed for the Chinese news commentary scenario. It is sourced from the editorial sections of The Paper and Guangming Online, containing 2,998 "prompt-essay" pairs. The essays have an average length of 1,063.09 tokens, with prompts averaging 171.06 tokens, making it the longest among the four datasets in terms of text length. This dataset focuses on social problems and controversial topics. The essays, written by professional commentators, feature formal language and in-depth argumentation. It is thus suitable for training and evaluating argument generation models tailored to news commentary scenarios that require handling complex logical structures and long-text generation capabilities.

A.2 Implementation Details

In our experiments, **to ensure a fair comparison, our framework and all baselines employ the same base model**. Specifically, we use Llama3.1-8B-instruct[3] for the English dataset and Qwen2.5-7B-instruct[4] for the Chinese dataset . Here, we leveraged vLLM (Kolluru, 2025; Kwon et al., 2023) for its efficient inference acceleration and set the temperature and top-p hyper-parameters to 0.8 and 0.9, respectively (Renze et al., 2024). This configuration balances creativity and coherence in the model's generation.

Following (Shahriar et al., 2024; Liu et al., 2023), we uniformly employed GPT-4o for automatic evaluation, which has been demonstrated to excel in text evaluation (Chia et al., 2024).

For human evaluation, we recruited three human annotators to assess the model outputs across four metrics. We ensured that all annotators were proficient in English and Chinese, held at least a bachelor's degree, and were all graduate students. Here, we randomly selected 40 inputs and presented the outputs from different models anonymously for each input. We asked the human annotators to score each essay on all metrics, and finally computed the average of all scores as the final evaluation result.

A.3 Baselines

The general information of the baseline model is as follows.

[3]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

E2E (Gretz et al., 2020) is an end-to-end generation approach, which bypasses intermediate processes like plan-then-write, directly prompting the large language model to generate an argumentative essay on a given proposition in a single step. The quality of the generated essay depends entirely on the model's inherent text generation capability.

DD-KW (Bao et al., 2022b) is an AEG model that employs a dual-decoder architecture, whose core methodology follows a "plan-and-write" paradigm. First, a structured content plan is generated by the planning decoder; then, the writing decoder, integrated with a plan-aware attention mechanism, dynamically focuses on the planning information during essay generation to ensure content consistency and richness.

CoT (Wei et al., 2022) use Chain-of-Thought prompting method, models are instructed to generate intermediate reasoning steps or a preliminary outline before composing the full essay. This step-by-step reasoning serves as an underlying structural guide for the writing phase, which tends to yield more coherent and logically organized text compared to approaches that generate the output directly without such intermediate planning.

ToT (Yao et al., 2023) (Tree of Thoughts) builds upon Chain-of-Thought prompting by introducing a structured, multi-path reasoning process. Rather than following a single sequence of reasoning, ToT enables a model to explore multiple alternative reasoning trajectories—each representing a potential "thought"—and systematically evaluates, compares, and even backtracks across these branches before committing to a final direction. Initially applied to tasks like puzzle solving and creative writing, this approach offers a more deliberate planning capability that can be leveraged for argumentative essay generation, allowing the model to develop, assess, and iteratively improve several essay outlines prior to text production.

DPE (He et al., 2024) employs a two-stage hierarchical planning strategy for generating argumentative essays. The process begins with sketch planning to draft a preliminary outline, followed by a dialectical planning stage where the outline is iteratively refined through critical self-evaluation. This framework incorporates explicit reasoning mechanisms, enhancing the ability of LLMs to produce essays with greater argument depth, content diversity, and persuasive strength.

D2W (Hu et al., 2025b) introduces a persona-based multi-agent framework for argumentative essay generation, inspired by the dynamics of human debate. Each agent is assigned a distinct persona, embodying specific high-level convictions and perspectives on the topic. Through structured rounds of debate and interaction, these agents collaboratively exchange, challenge, and synthesize ideas to form a comprehensive essay plan. This debate-to-writing paradigm facilitates the non-linear evolution and integration of diverse viewpoints, thereby producing arguments with greater diversity, coherence, and persuasive impact.

DraftRef (Hu et al., 2024) This method follows the evaluator-refiner interaction paradigm for argument generation. It first generates an initial essay draft, and then iteratively evaluates and refines the draft. Specifically, the evaluator assesses the current draft in terms of argumentative quality, coherence, and persuasiveness, while the refiner revises the draft according to the feedback. The final refined draft is used as the generated essay. Unlike our framework, which uses the abstract of the generated essay to refine the intermediate plan, *DraftRef directly optimizes the essay draft itself without updating the plan*. This baseline is designed to examine whether the performance gains mainly come from iterative draft-level revision rather than abstract-guided plan refinement.

DARE (Zhao et al., 2025) proposes a debate-driven rhetorical framework for AEG, which integrates: a depth-then-breadth dynamic outline tree construction strategy, where a host agent recursively generates hierarchical argument nodes to enhance perspective diversity; a multi-agent debate mechanism that simulates attacker-defender interactions; and the incorporation of Bitzer's rhetorical situation theory to dynamically select rhetorical devices based on contextual factors, thereby improving expressiveness and persuasive impact. By structuring content planning through formalized debate and adapting rhetoric to context, the framework significantly enhances the depth, logical coherence, and rhetorical effectiveness of generated long-form argumentative texts.

ReEssay This method inherits the core idea of improving essay quality through iterative self-refinement. Unlike our framework, which refines the intermediate plan content, *this method directly uses the generated essay itself as the subject of refinement*. This enables LLMs to directly analyze the essay content and make immediate modifications at the surface text level based on the analysis, thereby bypassing the traditional stepwise "plan-then-write" paradigm.

Dataset	Language	Size	Avg.C	Avg.E
ArgEssay	English	11,282	38.59	333.80
NYT-Editorial	English	9,178	114.17	419.32
CHE-Essay	Chinese	3,750	255.20	786.03
CHE-Editorial	Chinese	2,998	171.06	1063.09

Table 8: Datasets Information. Size indicates the total number of samples in the dataset. Avg.C/Avg.E indicates the average number of tokens in the claim/essays.

SumRefine This method follows our *Plan-Write-Refine* framework but replaces the abstract with an essay summary during plan refinement. Unlike an abstract—which distills the essay's key information at the argumentative structure—a summary offers a detailed overview at the surface text level, typically covering a wider range of content.

A.4 Algorithm

Below is the pseudo-code algorithm for the proposed framework.

Algorithm 1 Plan-Write-Refine Algorithm

Require: Prompt p including task instruction, Claim X

function LLMs(p)
  $P_0 \leftarrow$ GeneratePlan(X) (Eq. 1)
  $Y_0 \leftarrow$ WriteEssay(X, P_0) (Eq. 2)
  for $i = 1, 2, ...n$ **do** // i denote iteration
    $A_i \leftarrow$ GenerateAbstract (X, Y_{i-1}) (Eq. 3)
    $P_i \leftarrow$ RefinePlan (X, A_i, P_{i-1}) (Eq. 4)
    $Y_i \leftarrow$ WriteEssay (X, P_i) (Eq. 2)
  end for
  return Y_n
end function

B Additional Analysis and Discussion

B.1 Discussion on Novelty and Design Rationale

Prior work on AEG, such as DPE (He et al., 2024) and DARE (Zhao et al., 2025), largely follows the *plan-then-write paradigm*: an LLM first generates a single plan via prompting (e.g., dialectical planning or dynamic planning), and the final essay is then produced conditioned on this plan and

prompts. In this paradigm, the plan remains unchanged throughout generations, which means that any omission, imbalance, or logical weakness in the plan is directly propagated to the final essay, often resulting in weak logical connections or insufficient support for claims.

In contrast, this work introduces *a new paradigm, Plan–Write–Refine*, which fundamentally differs from plan-then-write in how planning and generation interact. **The key innovation is that**, throughout the generation process, the plan is refined through abstraction-based feedback that operates at the argumentative structure rather than at the surface text level, whereas prior work relies on a fixed plan throughout generation. By refining the plan using the essay's abstract, the model can reorganize arguments, identify missing or weak components, and progressively improve coherence and persuasiveness—mirroring how human writers revise plans after reviewing an abstract of their draft.

Moreover, we *extend this new paradigm into an iterative framework*, where plan refinement and essay generation form a closed loop. As shown in Figure 2, iterative refinement consistently improves performance, *providing empirical evidence that this abstract-driven structural feedback mechanism, rather than a prompt engineering tweak, is responsible for the observed gains*. Importantly, compared with the current SOTA model DARE, which remains within the plan-then-write paradigm, the average scores of our framework on automatic evaluation increased by 1.00%, 1.72%, 2.76%, and 2.43% across the four datasets. (Table 1).

We acknowledge that each step is implemented via prompts. However, the novelty of our work lies in the conceptual reframing of AEG as an abstraction-guided, iterative plan refinement process during essay generation, which departs from prior work where the plan remains static once writing begins.

B.2 Iteration-wise Output Length Analysis

To further investigate the effect of iterative refinement on output length, we report the average number of generated tokens at each iteration in Table 9. Notably, our method regenerates the complete essay based on the refined plan at each iteration rather than continuously appending new text to the previous draft. As the number of iterations increases, the average output length increases slightly on both datasets, but the overall growth remains relatively limited. These observations indicate that the iterative process does not simply extend essays by appending additional text; instead, it reorganizes and regenerates essays based on progressively refined plans.

Dataset	Iter. 0	Iter. 1	Iter. 2	Iter. 3
ArgEssay	786	808	824	831
CHE-Essay	997	1019	1031	1038

Table 9: Iteration-wise output length analysis. We report the average number of generated tokens per essay at each iteration. **Iter. 0** denotes the initial essay generated before plan refinement, while **Iter. 1–3** denote the essays regenerated after successive rounds of plan refinement.

B.3 Single-Agent vs. Multi-Agent Design

In fact, as a preliminary probing experiment, we implemented a multi-agent variant of the *Plan–Write–Refine* framework, where the planning, writing, and abstraction/refinement steps are handled by separate agents. On both ArgEssay and CHE-Essay, this multi-agent realization yields modest but consistent improvements of around 0.1 on average across the four metrics in both automatic and human evaluations.

However, these gains come at a substantially higher computational cost. Compared to our single-agent implementation—where all three stages are executed sequentially within a single agent—the multi-agent setting requires loading and maintaining multiple agents and their interactions, resulting in approximately three times higher computation and communication overhead. We therefore found that the performance gains are relatively small compared to the significantly increased resource consumption, making the multi-agent design less cost-effective in our current setting. For this reason, we adopt a single-agent framework in the final implementation to achieve a better trade-off between performance and efficiency.

B.4 Analysis of Iterative Behavior and Stability

Concerning potential topic drift when increasing the number of iterations, our results show no evidence of such an issue. As illustrated in Figure 2, performance does not significant degrade in later iterations but instead stabilizes, which indirectly verifies that the iterative refinement does not cause topical drift. This stability further supports the ef-

Dimension	Human–Human Agreement				Human–LLM Agreement			
	A1–A2	A1–A3	A2–A3	Mean	A1–LLM	A2–LLM	A3–LLM	Mean
Relevance	0.721	0.684	0.748	0.718	0.676	0.642	0.701	0.673
Logic	0.662	0.613	0.691	0.655	0.602	0.575	0.631	0.603
Complexity	0.581	0.548	0.624	0.584	0.621	0.528	0.557	0.569
Persuasiveness	0.647	0.602	0.671	0.640	0.584	0.562	0.615	0.587

Table 10: Agreement analysis of human and LLM evaluations using Quadratic Weighted Kappa (QWK). A1, A2, and A3 denote the three human annotators. Mean denotes the average QWK across the three corresponding annotator pairs.

fectiveness of abstraction-guided plan refinement and motivates our selection of three iterations as a practical trade-off between performance gains and computational efficiency.

B.5 Abstract Quality Analysis

We employ the following four evaluation metrics to assess abstract quality. Detailed descriptions are as follows:

Factuality (Fact.): The degree to which the abstract precisely preserves the original argument's stance strength, evidence scope, and critical qualifiers without distortion.

Informativeness (Info.): The completeness of the abstract's coverage of the core controversy, central thesis, essential sub-arguments, and conclusion forming a coherent logical chain.

Contribution (Con.): The contribution of the abstract to plan revisions is measured by the fidelity of the refined writing plan to the abstract's distilled core thesis and complete reasoning path, including the elimination of all non-aligned content.

Coherence (Cohe.): The extent to which revisions eliminate logical discontinuities of the plan by referencing the reasoning path distilled in the abstract.

We also used GPT-4o for abstract quality evaluation; each dimension is scored on a scale from 1 to 5. We conducted three evaluations on both datasets, and the final result is the average of the three evaluations.

B.6 Agreement Analysis of Human and LLM Evaluations

To further assess the reliability of human and LLM-based evaluations, we employ Quadratic Weighted Kappa (QWK) to measure agreement among different evaluators (Cohen, 1968). As shown in Table 10, we calculate pairwise QWK among the three human annotators and between each human annotator and the LLM judge. For the LLM-based evaluation, we take the median of three independent ratings for each essay and evaluation dimension as the final LLM score.

The results show that the mean human–human QWK ranges from 0.584 to 0.718, while the mean human–LLM QWK ranges from 0.569 to 0.673, indicating a reasonable level of agreement across the four evaluation dimensions. Agreement is relatively lower for Complexity, suggesting that assessments of this dimension may involve greater subjectivity.

B.7 Case Study

To qualitatively examine how the proposed framework refines argumentative planning and improves essay quality, we conduct a case study on a representative topic. Figure 5 illustrates how the initial plan is progressively refined across iterations, together with the evolution of the abstract that serves as the feedback signal. Figure 6 presents the corresponding essays generated at different iteration rounds. In both figures, the blue text highlights newly added and modified content compared with the previous iteration.

As shown in Figure 5, the initial plan is coarse and generic. The corresponding abstract summarizes these points at a high level, *e.g., "renewable energy offers significant environmental benefits but faces challenges like intermittency and high initial costs."* This abstract exposes the limited argumentative depth of the initial plan. Guided by this feedback, the refined plan at Iter. 3 introduces more fine-grained and hierarchically organized arguments. For instance, new plan elements like *"evaluating the broader economic and social implications"* and *"addressing the challenges through innovative solutions and policy reforms"* are added, which were absent from the initial outline.

These structural refinements are directly re-

flected in the generated essays (Figure 6). The initial essay largely follows a simple pros-and-cons structure, as seen in statements such as "*renewable energy offers substantial environmental benefits... however, challenges remain*." In contrast, the essay generated at Iter. 3 exhibits richer argumentative development and improved coherence. It explicitly elaborates on infrastructure constraints, grid modernization, and policy coordination, for example, "*existing power grids require modernization to accommodate decentralized renewable systems*". These additions demonstrate a clearer logical progression and more comprehensive coverage of the topic.

Overall, this case study highlights that a single static plan fails to anticipate many structural and argumentative deficiencies that only become apparent after text generation. By using the abstract as an abstraction-level feedback signal, our framework revises the plan at the level of argumentative structure rather than surface text. This iterative *Plan–Write–Refine* process progressively aligns planning with realized content, leading to essays that are more coherent, better structured, and more persuasive.

C Prompt

For clarity and reproducibility, we provide the prompts used in each step of our framework.

Specifically, the prompt for initial plan generation in Equation 1 is shown in Figure 8. The prompt for argumentative essay generation in Equation 2 is presented in Figure 9. The prompt for abstract generation in Equation 3 can be found in Figure 10. Finally, the prompt for plan refinement in Equation 4 is provided in Figure 11.

Claim: More renewable energy sources are harnessed to generate electricity these days.

Plan (Initial):

Introduction:The transition towards harnessing more renewable energy sources to generate electricity is a significant shift with far-reaching implications. ...

Main Paragraph One:Renewable energy sources are essential for sustainable development and reducing carbon emissions.

Main Paragraph Two:Despite the advantages, there are several challenges and limitations associated with relying heavily on renewable energy.

Conclusion:The trend towards harnessing more renewable energy sources to generate electricity is multifaceted, with both advantages and disadvantages. ...

Abstract (Initial):

The trend towards increasing renewable energy sources offers significant environmental benefits but faces challenges like intermittency and high initial costs, necessitating a balanced approach for ...

Plan (Iter.1):

Introduction:The shift towards more renewable energy sources has gained momentum globally as a response to climate change and a move away from fossil fuels.

Main Paragraph One:Renewable energy sources are essential for sustainable development and offer significant environmental benefits.

Main Paragraph Two:The trend faces substantial challenges that limit its effectiveness.

Main Paragraph Three:Both the environmental benefits and the practical challenges must be carefully considered.

Conclusion:This essay has explored the dual nature of the trend towards increased renewable energy use, highlighting both its promising environmental impacts and the practical challenges it presents.

Abstract (Iter.1):

The shift towards more renewable energy sources offers environmental benefits like reduced emissions and improved public health but faces challenges such as intermittency and high initial costs. ...

Plan (Iter.2):

Introduction:The shift towards more renewable energy sources has gained momentum globally as ...

Main Paragraph One:Renewable energy sources are essential for sustainable development and offer significant environmental benefits.

Main Paragraph Two:The trend faces substantial challenges that limit its effectiveness.

Main Paragraph Three:Environmental solutions are a double-edged sword: their benefits and practical difficulties are two sides of the same coin.

Main Paragraph Four:Addressing the challenges through innovative solutions and policy reforms.

Conclusion:This essay has examined the dual nature of the renewable energy transition, weighing its environmental promise against its practical complexities.

Abstract (Iter.2):

The shift towards more renewable energy sources offers environmental benefits but faces challenges like intermittency and high initial costs; a balanced approach integrating these technologies ...

Plan (Iter.3):

Introduction:The global transition to renewable energy has accelerated, driven by the urgency of climate change and the need to move beyond fossil fuels.

Main Paragraph One:Renewable energy sources are essential for sustainable development and offer significant environmental benefits.

Main Paragraph Two:The trajectory of this transition is complicated by considerable impediments that undermine its overall efficacy.

Main Paragraph Three:Environmental solutions are a double-edged sword: their benefits and practical difficulties are two sides of the same coin.

Main Paragraph Four:Addressing the challenges through innovative solutions and policy reforms.

Main Paragraph Five:Evaluating the broader economic and social implications of increased renewable energy adoption.

Conclusion:By interrogating the dual promise and challenges of renewable energy, this essay posits that a strategic approach is imperative to maximize benefits and address limitations.

Figure 5: Examples of iterative plan refinement.

Claim: More renewable energy sources are harnessed to generate electricity these days.

Essay (Initial):

More renewable energy is being used to generate electricity, offering substantial environmental benefits—reduced pollution and CO_2 emissions—and economic gains like job creation and energy security. ...

Renewables provide long-term stability compared with volatile fossil fuels. ...

However, challenges remain: intermittency of wind and solar complicates grid integration, high upfront infrastructure costs (offshore wind still expensive) ...

Social and ethical issues require transparent, inclusive planning to distribute benefits equitably and provide training. ...

In conclusion, the shift to renewables is essential for sustainability but needs careful planning and investment to maximize benefits and mitigate drawbacks ...

Essay (Iter.1):

The global shift toward renewable energy is driven by climate concerns and the need to reduce fossil fuel dependence, offering major environmental benefits but also notable challenges. ...

Renewable sources such as solar and wind significantly cut greenhouse gas emissions, reduce air pollution, and enhance energy security. ...

However, intermittency, high initial investment costs, and outdated grid infrastructure hinder large-scale adoption, requiring advanced storage technologies and grid modernization. ...

Despite these obstacles, declining technology costs, government incentives, and regional grid integration can mitigate limitations. ...

In conclusion, renewable energy is essential for a sustainable future, but its success depends on strategic planning, technological support, and effective policy implementation. ...

Essay (Iter.2):

The global shift toward renewable energy is driven by climate concerns and the need to reduce reliance on fossil fuels, offering clear environmental benefits but also notable challenges. ...

Renewable sources such as solar and wind significantly cut greenhouse gas emissions, reduce air pollution, and conserve natural resources like water. ...

However, intermittency due to weather dependence complicates reliable power supply, requiring advanced and costly storage technologies. ...

High upfront investment and outdated grid infrastructure further limit large-scale adoption, despite declining technology costs and long-term economic viability. ...

Strategic solutions—including energy storage, grid modernization, government incentives, and regional integration—can address these issues. ...

In conclusion, renewable energy is essential for sustainability, and with balanced policies and technological support, its environmental, economic, and social benefits can be fully realized ...

Essay (Iter.3):

The global shift toward renewable energy is driven by climate change concerns and the need to reduce dependence on fossil fuels, offering major environmental benefits alongside significant challenges. ...

Renewable sources such as solar and wind reduce greenhouse gas emissions, lower air pollution, improve public health, and conserve natural resources like water. ...

However, intermittency caused by weather dependence makes it difficult ..., requiring advanced energy storage and transmission technologies that still demand high investment. ...

High upfront costs and infrastructure constraints further hinder large-scale adoption, even though technology costs have declined and long-term savings are substantial. ...

Existing power grids require modernization to accommodate decentralized renewable systems, and regional integration can improve stability. ...

A balanced strategy combining renewable energy, storage technologies, grid upgrades, government incentives, and supportive policies is essential. ...

In conclusion, while renewable energy is vital for sustainability, careful planning and integrated solutions are necessary to overcome its limitations and realize long-term environmental, economic, and social benefits ...

Figure 6: Examples of the argumentative essay generation.

Prompt for Automatic Evaluation

Given a claim on a controversial topic: {**input_claim**}

Given an argumentative Essay: {**input_essay**}

Task: 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.

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.

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.

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

Figure 7: Prompt for automatic evaluation.

Prompt for plan generation

Given a claim on a controversial topic: {**input_claim**}

Background:You are a writer of argumentative essay. ...

Task: Generate a structured outline that briefly summarizes the core content of each section. This outline should be a skeleton of the essay, not the fleshed-out argument itself. Avoid providing specific evidence, data, or extended reasoning at this stage. At the same time, be careful not to deviate from the topic and maintain a high level of relevance to it.

#The final output must follow this format and contain only the summarized points:
Introduction:
(Briefly present the controversy and its significance)
...
Conclusion:
(Concisely summarize the main points of discussion from both sides)

Figure 8: Plan generation prompt in Equation 1.

Prompt for Essay Writing

Given a claim on a controversial topic: {**input_claim**}

Given a writing plan: {**input_plan**}

Background:You are a writer of argumentative essay. ...

Task: Write a coherent, persuasive,well-structured argumentative article to discuss the claim based on the plan.Please further consider and expand the content based on the plan.Provide specific evidence, data, and detailed reasoning for each part of the writing plan.

#Requirements for output content:
Your final output only includes one article.
Now，generate the essay.

Figure 9: Essay writing prompt in Equation 2.

Prompt for Abstract Generation

Given a claim on a controversial topic: {**input_claim**}

Given an argumentative Essay: {**input_essay**}

Background:You are provided with an argumentative essay aimed at analyzing and discuss a given claim on a controversial topic.

Task: Summarize the argumentative essay in one concise sentence (under 30 words). This sentence must clearly reflect the main argument,and the main reasons for making judgement.Avoid any extra details or elaboration.

Now, generate the final result given the claim and essay:

Figure 10: Abstract generation prompt in Equation 3.

Prompt for Plan Refining

Given a claim on a controversial topic: {**input_claim**}

Given a writing plan: {**input_plan**}

Given an abstract: {**input_abstract**}

Background:You are provided with a writing plan,this plan is used to generate an argumentative essay aimed at analyzing and discussing the give claim.And then,you are provided with an abstract that summarize the above argumentative essay in one sentence.

Task: According to the abstract and the writing plan, analyze what other arguments can be supplemented to better analyze and discuss the given claim or support the viewpoint you hava already put forward,and whether the existing arguments need to be supplemented, modified, or deleted.
And then,base on the analysis results,modify the writing plan.
The Plan is a structured outline that briefly summarizes the core content of each section. ...

#The final output must follow this format and contain only the summarized points:
Introduction:
(Briefly present the controversy and its significance)
…
Conclusion:
(Concisely summarize the main points of discussion from both sides)

Figure 11: Plan refining prompt in Equation 4.