Hi-Plan: Hierarchical Text-Planning for Argumentative Essay Generation

Anonymous ACL submission

Abstract

Argumentative Essay Generation (AEG) aims to generate a complete argumentative text on a specific controversial topic or debate. Due to the need to organize arguments into a complete chapter, AEG is challenging in terms of article structure. Existing approaches use text-planning methods to augment generative language models, neglecting the necessary attention to the hierarchical logical structure of argumentative texts. This causes difficulties in generating a complete logic flow, significantly reducing the persuasiveness of argumentative texts. To solve this problem, we designed a hierarchical text-planning method: Hi-Plan, introducing the structural and logical information of argumentative texts into text-planning explicitly. To be more specific, it allows this method to control the logical relevance in a high-level, fine-grained way simultaneously. Since Hi-Plan requires a fine-grained article outline, we propose an automated data augmentation method to extend existing AEG datasets. The hierarchical text planning and fine-grained data augmentation methods form a complete cycle from data construction to model training. Experiments show that our proposed method outperforms strong baseline models in terms of logical validity and persuasiveness and achieves comparable results with super-large scale pre-trained language models, such as ChatGPT.

1 Introduction

From decision-making to public discussions, argumentative texts serve to proclaim ideas or defend a point of view in a wide range of scenarios Sato et al. (2015). Argumentative essay generation (AEG) is a task designed to generate a persuasive argumentative text containing several arguments on a given controversial debate topic (e.g., whether euthanasia should be legalized). Generally, an ideal argumentative essay has 4 paragraphs and 400-500 words, first introducing background information on the topic (1 paragraph), followed by some balanced presentation of arguments organized coherently (2 paragraphs) and a convincing conclusion (1 paragraph). It requires a higher long-range correlation and logicality to ensure intra-paragraph consistency and inter-paragraph coherence. Therefore, to generate a persuasive argumentative essay, a complete logical flow needs to be formed between its major claims, and each major claim needs to be logically and effectively proved. This imposes more rigorous demands on the structure and logical coherence of AEG, making it more challenging.

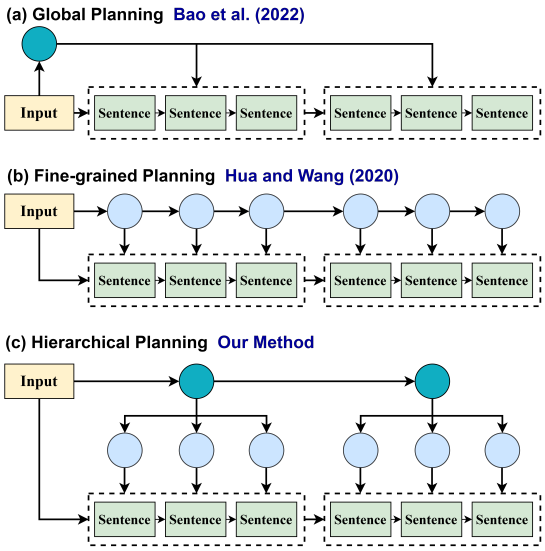

Figure 1: Methods in figure (a) use global planning to guide the generation of the whole essay, it is difficult to achieve sentence-level control. Methods in figure (b) generate a separate planner for each sentence, it is hard to guide chapter-level structure. Figure (c) shows our approach, it uses hierarchical planning structure, which guides both essay-level structure and sentence-level connections. Figure (a) and (b) comes from Hu et al. (2022)

To solve this problem, Bao et al. (2022). achieved excellent results by introducing the text-planning method to the AEG task. The core of this

method lies in designing a series of intermediate steps, rather than using an end-to-end approach to generate target text directly. Apparently, the well-designed planning signals enable the model to generate longer texts with a certain logic chain generation capability. There are two types of text-planning methods, (i) **Global planning**: guide the text by developing planning over the whole text; (ii) **Fine-grained planning**: design planning over each sentence. Both types have their drawbacks when performing the AEG task. Type (i) works well in guiding the main content and logic of the whole text, but loses control of keeping intra-paragraph self-consistency; while type (ii) better guides fine-grained sentence relevance, but fails in developing inter-paragraph logic chains on the chapter level. This dilemma creates trouble for current approaches to balance global and local consistency simultaneously, and even leads to self-contradiction.

To deal with the limitations of the current text-planning method, we propose Hi-Plan, a hierarchical text-planning method that leverages the strengths of both types. Inspired by the outlining method commonly employed in human writing, we introduced an explicit hierarchical structure into text-planning. This structure aims to enhance the guidance of the logic chain at both the chapter level between paragraphs and the sentence level within paragraphs in the AEG task. We have observed that human writers often make a detailed outline before writing an argumentative essay. This outline typically consists of several major claims that are relevant to the topic and includes a chain of supporting arguments or logical proofs for these major claims (Flower and Hayes, 1981). We abstract this kind of outline into a two-tier text-planning framework to finely control both logic flow and writing details. At the first level, a set of major claims is generated to prompt the overall logical structure. At the second level, arguments, logical reasoning and other supplementary writing materials are generated to guide writing in the sentence level. Finally, the comprehensive two-tier writing outline is expanded to produce a well-structured and persuasive argumentative text. Figure 1 shows the key differences between our approach and the existing two types of planning methods.

For the implementation of Hi-Plan, fine-grained data similar to human writing outlines are required, which can finely annotate the major claims and logical structure of the essay. The available AEG dataset lacks this fine-grained annotation, so we design a complete data augmentation method to construct high-quality training data. As shown in Figure 2, data augmentation method resembles the inverse process of the generation step, constructing text-planning ground truth step by step from the argumentative text paradigm. We use GPT-4 instead of manually for the automatic construction of trainning data, which not only reduces the cost of the data construction, but also makes our method with the possibility of iterative optimization.

We conduct experiments to evaluate our proposed model, using a publicly available AEG dataset with results indicating that our model achieved State-Of-Art across various automatic evaluation metrics. Meanwhile, manual evaluation further confirms that our model significantly outperforms the strong baseline with its superiority in terms of metrics such as fluency, logic and persuasiveness, and results are comparable to those of ChatGPT and other LLMs.

To conclude, our model made the following contributions

- We propose a hierarchical text-planning method that mimics the human writing outline, enabling the model to generate better-structured and more persuasive argumentative texts.

- We design a data augmentation method that automatically completes the fine-grained annotation of AEG, and constructs high-quality training data for hierarchical text-planning.

- Our model demonstrates superior performance over existing models across a wide range of human and automatic evaluation metrics and achieved comparable results to ChatGPT with much smaller parameters.

2 Task Definition

The AEG task can be formally defined as follows: given a writing prompt $X = \{w_x^1, w_x^2, ..., w_x^n\}$, generate coherent and persuasive argumentative text $Y = \{w_y^1, w_y^2, ..., w_y^n\}$ associated with the writing prompt. Methodology of Hierarchical context-planning Inspired by previous work(Hua and Wang, 2020; Hu et al., 2022), we adopt a two-step approach, involving planning generation followed by chapter generation, to accomplish the AEG task. In previous work, planning can be defined as a hidden variable z_t, a

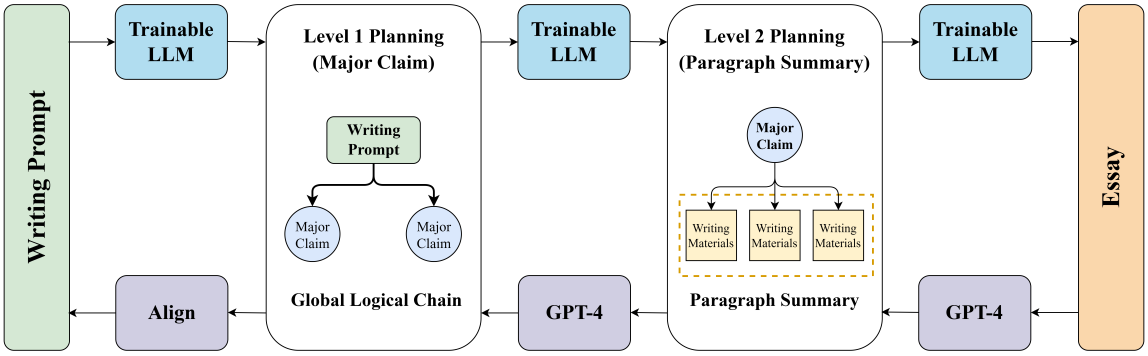

Figure 2: The full flow chart of hierarchical text-planning. The arrows from left to right show the process of generating text-planning from writing prompt and finally generating argumentative text, while the arrows from right to left show the process of gradually building hierarchical text-planning from cloud truth using ChatGPT. Detailed hierarchical text-planning samples are given in and Appendix A and Appendix B .

keyword set $\{w_i\}_{i=1}^{n}$,or a keyword-location pair $\{(w_i, pos_i)\}_{i=1}^{n}$.

During the generation process, the model first determines the planning based on the writing prompt and then relies on the planning to guide the current time generation.

Unlike previous approaches that employed global or sentence-level planning, our method defines a hierarchical text-planning approach to guide the generation of hierarchical logic chains within argumentative texts. Moreover, we use natural text for planning instead of using unreadable keywords or hidden variables.

In particular, our planning is structured as a hierarchical set of texts, denoted as $U^h = \{w_{u_1}^h, w_{u_2}^h, ..., w_{u_n}^h\}$, where "$h$" indicates the level of the planning and $w_{u_1}^h, w_{u_2}^h, ..., w_{u_n}^h$ represent individual tokens within the text.

3 Method

3.1 Overview

According to Figure 2, two stages of Hi-Plan form a whole cycle. Inspired by the two tiers outline method used by professional TOEFL or IELTS writers, we design the hierarchical text-planning into two tiers: the first-tier controls inter-paragraph logical flow, and the second controls intra-paragraph relevance and logical structure. This approach significantly improves the metrics of the generated text, especially logical validity.

Hierarchical text-planning needs hierarchical fine-grained annotation for training. Ideally, an outline of the human writing that corresponds to the ground truth text is necessary, including the major claim and the detailed logical structure. However, it can be a problem to obtain these outlines from professional writers. Therefore, we design an automatic data augmentation method, using GPT-4 to generate hierarchical text-planning from the provided argumentative text paradigm for training.

Overall, the hierarchical context-planning is a step-by-step process of unfolding a writing prompt into a complete argumentative text, whereas fine-grained data augmentation, on the contrary, compresses a longer argumentative text into writing prompt. This shows that our proposed approach covers all the steps of the data construction and training phases. Meanwhile, after the above two steps of generation-compression, the original writing prompt X will become a new writing prompt X'. This is similar to the training process of an autoencoder or VAE. It suggests that we are able to self-optimize the model in few-shot scenarios. In future work, we will further explore this potentiality of our approach.

3.2 Hierarchical Text-planning Generation

We referred to the outline method that human uses extensively when working on argumentative essays such as TOEFL or IELTS writing tasks. This outline usually contains two components: (i) **Intra-Paragraph Logical Chain**: a list of major claims based on the essay's topic, and each of them should build a complete logical chain to argue for a particular topic or express a specific point of view. (ii) **Inter-Paragraph Reasoning**: a list of supporting examples or the process of logical reasoning to justify each claim respectively. Finally, the human writer can expand the outline into a complete argumentative essay.

As we can see, the human writer's outline re-

flects the two tiers of hierarchical structure: major claims establish inter-paragraph logical structure, while arguments and logical reasoning set up intra-paragraph logical structure. Based on this observation, we also design the hierarchical text-planning into two tiers, formally notated as U^1 and U^2. Furthermore, unlike most of the previous work, we employ text-planning at each level constructed in natural language similar to a human outline instead of using unreadable hidden variables or sets of keywords as text-planning. This is because utilizing methods like self-instruction learning that allows models to understand natural, readable planning, can be seen as a more user-friendly approach for generative LLM.

More specifically, the first step is to train the model $P_\theta(U^1|X)$, this model aims to generate the major claims for each paragraph, considering their relevance to the given topic and the logical structure, such as juxtaposition and progression among these claims, which shape the overall logical flow of the full text.

After that, the next level model $P_\theta(U^2|X, U^1)$ is trained. This model needs to generate arguments for each major claim in form of brief paragraph, considering proper logical reasoning to support the claim accordingly, which consist the intra-paragraph logical structure.

Finally, the generate model $P_\theta(Y|X, U^1, U^2)$ expands its second level of text-planning to complete the AEG task. Figure 2 illustrates detailed process diagram of this method, while figure in Appendix B provides an in-depth overview of the hierarchical text-planning design. Through this methodology, a hierarchical logical structure and the content control flow are explicitly introduced into the generation process, enabling the model to generate hierarchical logic chains similar to human counterparts and better fulfill the requirements of the AEG task.

3.3 Fine-grained Training Data Augmentation

The process of generating hierarchical text-planning data from a given range of argumentative text acts as the inverse process of the AEG. Drawing inspiration from the hierarchical summary technique proposed by Wu et al. (2021), we generate each level of text-planning in reverse order.

Firstly, we use GPT-4 to generate summaries for each paragraph of the given text. These summaries are instructed to contain complete logical structure and writing materials such as examples and logical reasoning sentences, and serve as the second level of text-planning U^2. We use direct prompt to guide GPT-4 to generate the text-planning we need. The specific prompt is: *"The following text is a paragraph in an argumentative text about [Recently , courts have...](**Writing prompt**). Please summarize the text. The generated summary needs to contain complete logical flow, key examples and other writing materials should be missed."*

Then these summaries are used as a foundation to extract the paragraph major claims, which could constitute the first level of text-planning U^1. In this step we use a combination method of rule-base and GPT-4. We have observed that in argumentative texts, the first sentence of a paragraph is often the major claim, so we first use GPT-4 to determine whether the sentence is a major claim. If not, we use GPT-4 reading the full text and summary of the paragraph for extraction.

We illustrate this pair of processes in Figure 2, it is clear that the data augmentation process of building text-planning is similar to long text compression. Several studies have shown that LLMs such as ChatGPT and GPT-4, are good data annotators and generators because they could produce data whose quality is comparable to those manually created by humans. As a result, the quality of the generated text-planning using this approach can be guaranteed.

3.4 Training and Inference

As for training process, we train three generative models $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_0$ is the writing prompt to generate first-level text-planning (major claim), $\mathcal{L}_1$ represents the second-level text-planning (paragraph summary) generated with the help of the first-level text-planning, and $\mathcal{L}_2$ shows the generated argumentative text based on the second level of text-planning.

The specific definitions are as follows:

$$\mathcal{L}_0 = -\sum_{t=1}^{|X|} logP(U_t^1|X_{<t}) \quad (1)$$

$$\mathcal{L}_1 = -\sum_{t=1}^{|X|} logP(U_t^2|U_{<t}^1, X) \quad (2)$$

$$\mathcal{L}_2 = -\sum_{t=1}^{|X|} logP(Y_t|U_{<t}^2, U_{<t}^1, X) \quad (3)$$

where X, U_t^1, U_t^2, and Y represent writing

prompt, first-level text-planning, second-level text-planning, and argumentative text, respectively. As for the reasoning process, we generate U_t^1, U_t^2 and Y at one time using a form of pipeline, and all generation steps are done using an autoregressive large language model.

4 Experiments

4.1 Datasets

We evaluated our model on the ArgEssay(Bao et al., 2022), the largest standardized assessment dataset currently available for the AEG task. This dataset comprises 11,000 high-quality argumentative essays and corresponding writing topics sourced from various international standardized English writing tests (e.g., IELTS and TOEFL). It covers a range of common controversial topics such as technological advances, educational approaches, and environmental issues. The final dataset consists of 11,282 writing topic-argumentative text pairs, of which 9,277 pairs are allocated to the training set, 1,002 pairs to the validation set, and 1,003 pairs to the test set.

4.2 Baselines

We compared our proposed approach with several strong baseline models to showcase its effectiveness.

DD-KW: The text-planning method proposed by (Bao et al., 2022) . uses a series of keyword sets $\{m_i\}_{i=1}^n$ as planning, and designs a dual decoder structure based on BART, one is a planning decoder (PD) to generate text-planning, and the other is a writing decoder (WD) finally generates argumentative texts.

DD-Rel: Another text-planning method proposed by (Bao et al., 2022) . still uses the BART+dual decoder structure in the design of the generation model, and changes the text-planning from keywords to the set of relational triples $\{(s_i, o_i, r_i)\}_{i=1}^n$ drawn from off-the-shelf OpenI, where s_i,o_i,r_i denote the subject, object and relation in the i-th triad, respectively

LLaMA: (Touvron et al., 2023) presents a collection of baseline language models from 7B to 65B parameters LLaMA, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, while LLaMA-65B is competitive with the best models Chinchilla-70B and PaLM-540B, which is a very strong baseline model, especially for text generation tasks. As with the BART and OPT baselines, we also directly tuned LLaMA2-13B-chat using writing prompt and argumentative text

4.3 Implementation Details

We finetune LLaMA2-13B-chat with all parameters with the help of huggingface and DeepSpeed. AdamW optimizer is adopted for optimization, and initial learning rates are set to 1e-5 with linear descent schedule. We train the model 5 epochs. The batch size per-device is set to 8. All experiments are conducted with 8*NVIDIA Tesla A100 GPU.

4.4 Evaluation Metrics

Automatically evaluating open-domain text generation text is a challenging task(Celikyilmaz et al., 2021). For the AEG task, evaluating the quality of generated text solely based on automatic metrics such as BLEU can be insufficient, as there can be multiple valid argumentative essays in the ground truth. To solve this problem, some works in recent years have utilized GPT-4 for ground-truth-free evaluation, which significantly enhance the consistency of automatic evaluation metrics with manual evaluation(Hu et al., 2023; Liu et al., 2023; Fu et al., 2023). Inspired by these work, we propose a GPT-4-based AEG evaluation approach.

4.4.1 Automatic Metrics

Based on the systems theory of argumentation(Van Eemeren et al., 2004), we evaluate the persuasive of argumentative essay from the following five aspects: **Relevance**, **Validity of the Reasoning**, **Credibility of Evidence**, **Language and Rhetoric** and **Overall Persuasiveness**. We design prompts for each evaluation aspect with specific task instructions and a comprehensive list of detailed criteria. To reduce the variance of the evaluation results, we prompt model to give a detailed explanation for evaluation results before the final prediction of score. Detailed prompts for each evaluation aspect will be given in Appendix C.

4.4.2 Human Evaluation Metrics

We employed three skilled English speakers with NLP background to perform manual evaluation of the generated results. We set up the following two evaluation tasks:

Scoring the generated text: Same as GPT-4 based evaluation, score the generated results in terms of five aspects with the same criteria as the prompt used in GPT-4.

Method	Relevance	Validity of Reasoning	Credibility of Evidence	Language and Rhetoric	Overall Persuasiveness
DD-KW	58.80	52.25	25.00	54.13	52.26
DD-Rel	60.73	54.45	25.93	55.48	53.66
LLaMA2-base	90.40	80.26	45.92	72.28	70.54
Ours	**90.75**	**84.64**	**49.20**	**75.52**	**75.04**
-w/o Level 1 planning	89.23	82.35	48.02	73.12	72.42
-w/o Level 2 planning	90.00	83.33	46.92	72.34	73.19

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

Method	Relevance	Validity of Reasoning	Credibility of Evidence	Language and Rhetoric	Overall Persuasiveness
DD-KW	3.82	2.86	3.30	3.04	3.00
DD-Rel	3.72	2.9	3.34	3.08	3.00
LLaMA2-base	4.92	4.33	4.45	4.45	4.51
ChatGPT	**4.97**	4.62	**4.88**	**4.70**	**4.82**
Ours	4.94	**4.68**	4.76	4.60	4.76

Table 2: The results of comparison of baselines on human evaluation metrics. **Bold** numbers denote the best performance among all methods on each dataset.

Model	Relevance	Quality
LLaMA2-13B	4.90	3.75
ChatGPT	4.92	4.15
GPT-4	**4.95**	**4.33**

Table 3: Comparison of the effects of different models on the construction of fine-grained training Data. The human annotator evaluated 50 samples in terms of both relevance and quality out of 5 points. **Bold** numbers denote the best performance among all methods on each dataset.

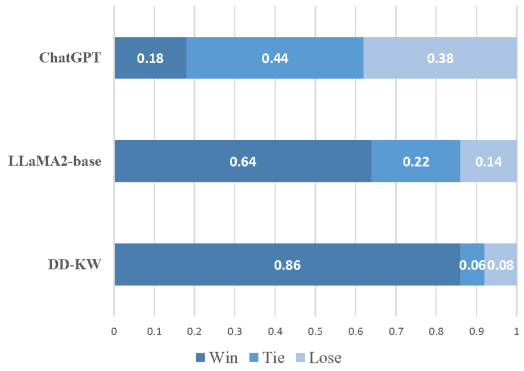

Figure 3: Hi-Plan compared to other baselines. Human raters compared different model generations and and scored them accordingly.

Compare the results with baselines: the annotator will compare the outputs of our proposed method with all the baselines and give an overall evaluation of **Win**, **Loss** and **Tie** for each test sample pair.

5 Results and Analysis

5.1 Main Results

Performance on automatic metrics: In Table 1, we present the results of automatic evaluation metrics, demonstrating the superior text generation quality, and novelty of out model compared to the current State-Of-Art models DD-Rel and LLaMA. Specifically, we outperform DD-Rel by 49.43%, 62.00%, 89.74%, 36.12% and 74.04% on Relevance, Validity of the Reasoning, Credibility of Evidence, Language and Rhetoric and Overall Persuasiveness, respectively. Furthermore, our designed method also significantly stronger than directly fintune LLaMA model, by outperforming 0.4%, 5.46%, 7.14%, 4.48% and 6.38% on Relevance, Validity of the Reasoning, Credibility of Evidence, Language and Rhetoric and Overall Persuasiveness. The performance on the automatic evaluation metrics proves the effectiveness of our proposed hierarchical text-planning method. In particular, our method has the most obvious improvement in Validity of the Reasoning and Credibility of Evidence. It is because our hierarchical text-planning method can significantly reduce the difficulty to generate complex logical structures.

Performance on human evaluation: The human evaluation metrics are presented in Table 2 and Figure 4. Results in Table 2 show that our model outperforms the strong baseline models BART, LLaMA and DD-KW in terms of all five aspects, which indicates that our proposed approach effec-

tively guides the model to produce diverse and high-quality texts. In particular, our model is significantly better than the strong baseline model in both Validity of the Reasoning and Credibility of Evidence scores, and achieves comparable results with ChatGPT. The results in Figure 3 demonstrate a direct comparison of our method with strong baselines. Results shows that our model outperform the current SOTA DD-KW and LLaMA by 64% and 86%, respectively, and is not inferior to ChatGPT by 62%. According to human evaluation, we found that our model achieved comparable results to ChatGPT on multiple scores, and did not fall behind in direct comparison. It means our Hi-Plan method can effectively gain smaller parameters model to achieve comparable results with larger parameter models on AEG tasks.

5.2 Ablation Study

Effectiveness of Hierarchical Text-planning Method: We conducted ablation experiments to determine the contribution of each layer of text-planning and model base to the final results. The results are presented in Table 1, demonstrating the effectiveness of our hierarchical text-planning approach with the fact that Level 1 text-planning contributes 3.49% of Overall Persuasiveness, while Level 2 text-planning contributes 2.47% of Overall Persuasiveness, demonstrating the effectiveness of the hierarchical text-planning approach we designed.

Effectiveness of Fine-grained Training Data Augmentation: Further, in Table 3, we analyze the effectiveness of different base models on the data augmentation method by manual evaluation. The results show that GPT-4 performs well in both relevance and quality of the generated plannings. Meanwhile, the selection of LLaMA2-13B caused only about 15% performance loss in data construction, so our method also has good compatibility with smaller models.

5.3 Case Study

We give an example in the Appendix D that includes all baselines and our model outputs. In this example, the Writing Prompt is "In countries with a high rate of unemployment, numerous pupils should do offered only predominant education. There is no point in offering secondary education to those who have no hope of finding a job. To what extent do you agree or disagree with this statement". It can be seen that our method and ChatGPT provide a clear and complete global logic chain, and also have good performance in logical coherence within paragraphs.

In contrast, the two claims given by the LLaMA2-base are not deeply related to unemployment in Writing Prompt, and the internal logic is not smooth with no examples or logical reasoning to prove its own opinions. The DD-KW method generated an article that overall conforms to the topic, but the logical structure is very confusing. In the first paragraph, DD-KW believes that "the government should not offer tertiary education to those who are unemployed" is wrong, but the following two major claims turn to support this view. The DD-Rel method is similar to the DD-KW method and produces serious logical errors. For example, DD-Rel believes that "offering secondary education to pupils who have no hope of finding employment is not beneficial", but in the following two paragraphs, one introduces university education at length, and the other is more inclined to support this view.

6 Related Work

6.1 Text-planning Method in Neural Generation

Given that natural language generation models often produce texts that lack relevance to the topic or exhibits a deficiency in high-level structure(Zhao et al., 2017; Xu et al., 2020), the idea of using external information as text-planning to guide generative models has a long history.

During the early stages of deep learning, a group of researchers began integrating text-planning methods with the Seq2Seq framework. For example, Yao et al. (2019) used storylines as global text-planning in story generation, resulting in more coherent generated stories. Wiseman et al. (2018) applied an implicit Half Markov Model (HSMM) decoder, enabling the model to learn discrete underlying text-planning alongside the generation process.

Currently, researchers have shown great interest in applying text-planning to large-scale pre-trained language models such as GPT. Tan et al. (2021) used several keywords as planning and progressively refined them via multiple stages into generate coherent long texts. Mirowski et al. (2023) built structural context via prompt chaining and large language model to enhance usefulness for long-form creative writing. Guan et al. (2023) ex-

tended Transformer model to dynamically learn and update text planning, which led a more coherent and diverse stories and news generation.

The study of the AEG began relatively late and research on relevant methodologies is still in its infancy due to its remarkable challenging nature. Bao et al. (2022) first proposed ArgEssay, a large-scale standard dataset that collected 11,000 high-quality argumentative 1essays and corresponding writing topics covering common controversial topics including technological progress, educational methods, and environmental issues.

In AEG-related tasks, such as argument generation, we could also witness great attempts to utilize text-planning. Hua and Wang (2020) picked up keywords and masks to create a fill-in-the-blank template as planning, enhancing the relevance and consistency of argument generation. Hua et al. (2021) used concepts and knowledge relevant to viewpoints as planning to increase the persuasiveness of argument generation. Hu et al. (2022) amplified Transformer decoder with latent representations and maintained sentence-level text-planning based on word packets, succeeding in generating more content-rich coherent text.

However, all these existing methods failed in generating the hierarchical chapter structure of argumentative texts. Moreover, current planning methods often rely on unreadable hidden variables or a series of keywords, which is lack of interpretation for language models.

6.2 Opinion Text Generation

Opinion generation produces arguments for or against a given topic. Unlike the AEG task, generating a number of major claims is enough for opinion generation task and it does not require a full essay with detailed argumentative process.

Some early work used retrieval methods to obtain relevant argumentative content from existing argument repositories, such as Levy et al. (2018) and Stab et al. (2018), or implemented argument retrieval tools to rewrite existing arguments, as demonstrated by Wachsmuth et al. (2018) and Hua and Wang (2018).

With the development of large-scale pre-trained generative language models, recent research has emphasized the direct generation of arguments using generative methods. Gretz et al. (2020) devised an opinion generation pipeline using GPT-2. Schiller et al. (2021) proposed a controlled opinion generation model capable of producing sentence-level arguments based on given topics, positions, and aspects. Al Khatib et al. (2021) developed three argumentative knowledge graphs and extracted knowledge from that to formulate prompts for training end-to-end opinion generation models. Hu et al. (2023) proposed a agent interaction framework for opinion and counter-arguments generation using ChatGPT.

The above approaches mainly focused on generating single and brief arguments, without considering the generation of coherent long argumentative texts encompassing multiple arguments.

7 Conclusion

In this paper, we propose Hi-Plan, a complete cycle for AEG containing a hierarchical text-planning method and fine-grained data augmentation method. Our model presents the ability to generate complex logical argumentative structures by designing a hierarchical text-planning guidance model similar to the writing outline written by humans. Automatic and manual evaluation metrics show that our model could produce fluent, coherent, and persuasive argumentative text with a significant improvement in the logical relevance within and between paragraphs.

8 Limitation

Despite the excellent results of our model on the AEG task, a potential limitation exists in our model due to the absence of explicit modeling for causal relationships among major claims, topics, and writing materials within paragraphs. In future studies, we plan to use causal modeling approaches to target logical structures in argumentative texts for more in-depth control.

9 Ethical Consideration

The ethical risks of our proposed methods and models are low. This is because we used a publicly published dataset that has passed ethical review and does not contain sensitive or private content. Moreover, the open source models we use such as LLaMA have undergone strict security training, and the output content complies with ethical standards.

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.

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

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey.

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

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

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 the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Jian Guan, Zhenyu Yang, Rongsheng Zhang, Zhipeng Hu, and Minlie Huang. 2023. Generating coherent narratives by learning dynamic and discrete entity states with a contrastive framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12836–12844.

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.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2023. Americano: Argument generation with discourse-driven decomposition and agent interaction.

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

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

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

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

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

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.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,

Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

A Example of Hierarchical Logical Structure

An example of a hierarchical logical structure that exists in real argumentative essays is given in Figure 4.

B Detailed Design Of Hierarchical Text-Planning

The detailed design of the hierarchical text-planning is given in Figure 5.

C Automatic Evaluation Prompts

For GPT4-based automatic evaluation as described in section Experiments, we present the detailed prompts for each metric in Table 4. The scoring range of the model is given after each prompt, and the model is required to give a brief explanation before scoring. This is to constrain the model output and reduce the randomness and variance.

D Example of Outputs

We present additional examples with different model outputs in Table 5 and 6.

Metric	Prompt
Relevance	Rate the direct relevance of the above argumentative essay to the topic. The criterion is: all the claims in the argumentative essay should be related to the topic or the major claim of the essay. Rating score is out of 100. Please give a short reason for the rating first, followed by a direct score in the form of xx/100.
Validity of the Reasoning	Rate the clarity and coherence of the logic of the above argumentative essay. The criteria are: the main ideas presented in the essay should have logical relationships, including progression, concessions, etc., and be able to argue the major claim of the essay. Each main idea should be logically and consistently argued. Rating score is out of 100. Please give a short reason for the rating first, followed by a direct score in the form of xx/100.
Credibility of Evidence	Rate the credibility of the evidence used in the above argumentative essay. The criterion is: ideas should be supported by credible evidence or sources, and unsubstantiated ideas are not as persuasive as rebuttals based on corroborating evidence. Rating score is out of 100. Please give a short reason for the rating first, followed by a direct score in the form of xx/100.
Language and Rhetoric	Rate the effectiveness of the language and rhetoric used in the above argumentative essay. The criterion is: argumentative essay should effectively use persuasive language and rhetorical techniques, such as appealing to logic or evidence, to convince the reader. Rating score is out of 100. Please give a short reason for the rating first, followed by a direct score in the form of xx/100.
Overall Persuasiveness	Rate the overall persuasiveness of the above argumentative essay. The criterion is: the argumentative essay should effectively argue its point of view in a way that the reader is convinced and agrees with it. Please give a short reason for the rating first, followed by a direct score in the form of xx/100.

Table 4: Prompts for all GPT4-based automatic metrics.

Writing Prompt: In countries with a high rate of unemployment, numerous pupils should do offered only predominant education. There is no point in offering secondary education to those who have no hope of finding a job. To what extent do you agree or disagree with this statement.
DD-KW:These days, a large number of the students are unable to finish their studies in the secondary schools due to the reasons. Some argue that the government should not offer tertiary education to those who are unemployed. I completely disagree with this view and in this essay, I will discuss my own opinion in detail. First and foremost, offering secondary education to the students Suggestionin the primary schools is not beneficial to all of the pupils. If the primary school does not have enough resources to provide higher education, it will not be beneficial for the students. For example, if the secondary school does offer only primary school, the students who are unable find a job will not tend to study there. In addition, the primary institution will not have adequate resources to train the students with the right skills. Moreover, the secondary education will not provide the job seekers with sufficient opportunities to work. The children who are not able to find a well-paid job will be unable to afford the tuition fees. As a result, they will not get the necessary skills and knowledge to continue their studies. For instance, In the USA, the majority of the school leavers are unable as they are not capable of learning the subjects required to become a skilled worker. Therefore, it is apparent that the secondary level will not benefit the students as they do not have sufficient resources to study. In conclusion, offering primary education has few advantages. The government should provide a job opportunity to the pupils who are unready to study the subjects. **DD-Rel**: In some countries, many pupils are keen to get only predominant education. However, I believe that offering secondary education to pupils who have no hope of finding employment is not beneficial. I will explain why many students are keen on focusing on their secondary education and the reasons why I agree with this view. Firstly, focusing on secondary education is a waste of money. For example, the price of a university course is much higher than another school course. For this reason, some pupils prefer to study a school subject that is related to their interest. As a consequence, they are willing to pay more money for a good university. Moreover, they feel that studying the subject that they really like will help them to get a job. It means that if students study only secondary education, they will not get a good job in the future. Secondly, focusing only on primary education can be harmful to the students. There are many subjects which are useless in the current market. For instance, English is considered to be less important than science. Many students who have a talent in English tend to choose a programming program that is not related to science, which is not useful for their job. This means that they will be more likely to be unemployed. In conclusion, in my opinion, it is not an effective way to offer a secondary education for those who do not have any hope of getting a job right away. Therefore, focusing more on secondary studies can be a better way to encourage students to study. **LLaMA2-base**:In the developing countries, the unemployment is a major problem. Some people think that the students should be provided with the vocational training only, as the secondary education is useless for them. I completely disagree with this statement. Firstly, the education is the basic right of every human being. The government should provide the education to all its citizens. The unemployment is not the fault of the students. They should be provided with the education, so that they can also serve the nation in the future. Moreover, the education is the only way to reduce the poverty and the crime rate in the society. Secondly, the education is not only the right, but it is also the responsibility of the government. The government should provide the education to all its citizens. The government should not discriminate between the rich and the poor, the literate and the illiterate. The government should provide the education to all its citizens. The education is the only way to reduce the poverty and the crime rate in the society. Thirdly, the education is the basic requirement of the human being. The education is the only way to reduce the poverty and the crime rate in the society. The education is the only way to serve the nation. The education is the only way to get the job. The education is the only way to get the respect in the society. To conclude, the education is the basic right of every human being. The education is the only way to reduce the poverty and the crime rate in the society. The education is the only way to serve the nation. The education is the only way to get the job. The education is the only way to get the respect in the society.

Table 5: Output example of different baselines and ours.

Writing Prompt: In countries with a high rate of unemployment, numerous pupils should do offered only predominant education. There is no point in offering secondary education to those who have no hope of finding a job. To what extent do you agree or disagree with this statement.
ChatGPT:In some countries, many pupils are keen to get only predominant education. However, I believe that offering secondary education to pupils who have no hope of finding employment is not beneficial. I will explain why many students are keen on focusing on their secondary education and the reasons why I agree with this view. Firstly, focusing on secondary education is a waste of money. For example, the price of a university course is much higher than another school course. For this reason, some pupils prefer to study a school subject that is related to their interest. As a consequence, they are willing to pay more money for a good university. Moreover, they feel that studying the subject that they really like will help them to get a job. It means that if students study only secondary education, they will not get a good job in the future. Secondly, focusing only on primary education can be harmful to the students. There are many subjects which are useless in the current market. For instance, English is considered to be less important than science. Many students who have a talent in English tend to choose a programming program that is not related to science, which is not useful for their job. This means that they will be more likely to be unemployed. In conclusion, in my opinion, it is not an effective way to offer a secondary education for those who do not have any hope of getting a job right away. Therefore, focusing more on secondary studies can be a better way to encourage students to study. **Ours**: In many countries, the unemployment rate is increasing at an alarming rate. Some people believe that students from poor families should only be offered predominant education and there is no point in offering secondary education to them as they will not be able to find a job in the future. I completely disagree with this statement and I will discuss why I disagree with this statement in this essay. Firstly, the main reason why I disagree with this statement is that the secondary education is the fundamental step for higher education. This education provides the necessary knowledge and skills to the students so that they can pursue higher studies. For example, in most of the countries, higher education is not accessible for the students who do not have secondary education. Therefore, secondary education is a pre-requisite for higher education. Secondly, secondary education is the basic right of every citizen. Every child has the right to education irrespective of their financial background and future job prospects. It is the responsibility of the government to ensure that every citizen has access to this basic right. For instance, in India, the right to education act of 2009 was passed to ensure that every child between the age group of 6 to 14 has a fundamental right to education. In conclusion, I strongly disagree with the statement that only predominant education should be offered to the students in countries with high unemployment rate. I believe that every child has a right to education and it is the responsibility of the government to ensure that every child has access to this right.

Table 6: Output example of different baselines and ours.

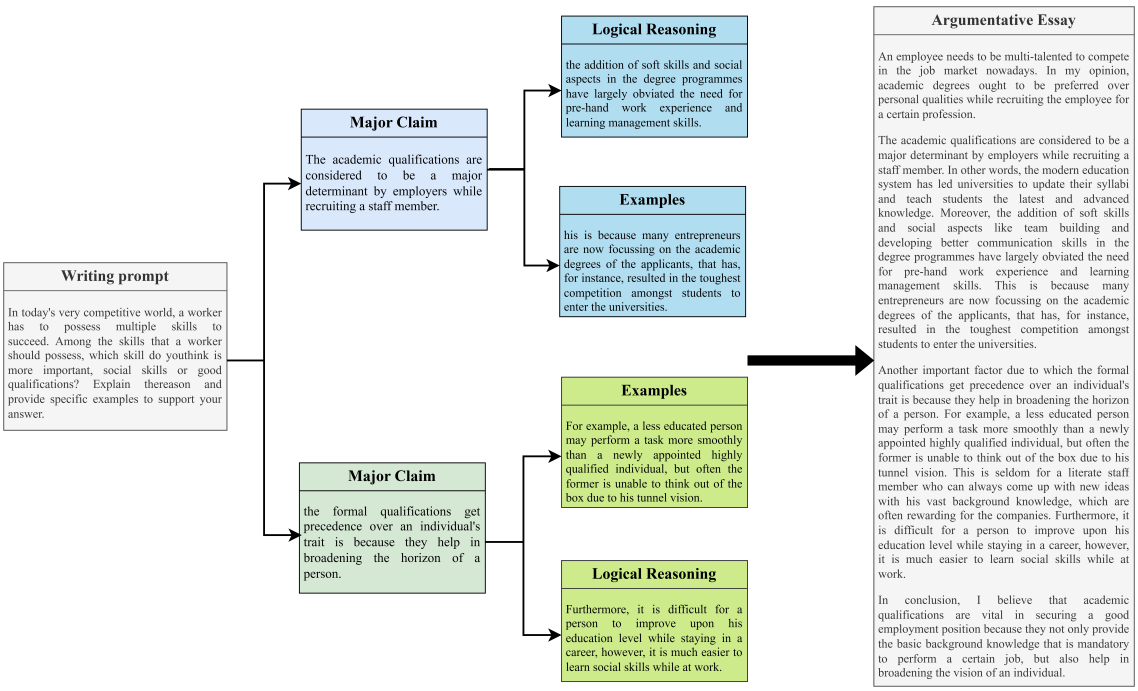

Figure 4: Example of hierarchical logical structure in human-authored argumentative text. The leftmost writing prompt extends two Major claims, after which each Major claim expands into several examples or logical reasoning. Finally the argumentative text is written according to the outline formed by the hierarchical logical structure.

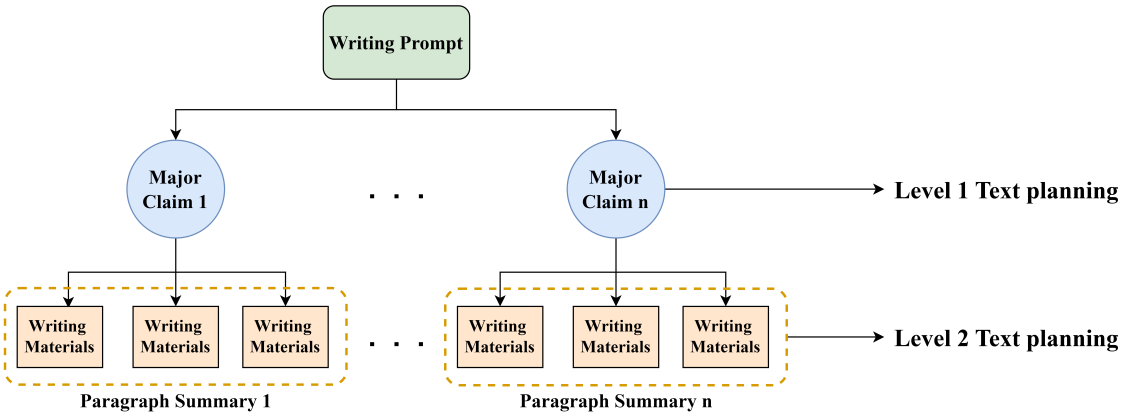

Figure 5: The specific design of hierarchical Text-planning. Two levels of text-planning are shown from top to bottom: the first level is the major claim, and the second level is a paragraph summary consisting of writing material.