Exploring Quality and Diversity in Synthetic Data Generation for Argument Mining

Jianzhu Bao[1,2*], **Yuqi Huang**[1*], **Yang Sun**[1], **Wenya Wang**[2], **Yice Zhang**[1], **Bojun Jin**[1], **Ruifeng Xu**[1,3†]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Nanyang Technological University, Singapore
[3]Peng Cheng Laboratory, Shenzhen, China
jianzhubao@gmail.com, 210110113@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

The advancement of Argument Mining (AM) is hindered by a critical bottleneck: the scarcity of structure-annotated datasets, which are expensive to create manually. Inspired by recent successes in synthetic data generation across various NLP tasks, this paper explores methodologies for LLMs to generate synthetic data for AM. We investigate two complementary synthesis perspectives: a quality-oriented synthesis approach, which employs structure-aware paraphrasing to preserve annotation quality, and a diversity-oriented synthesis approach, which generates novel argumentative texts with diverse topics and argument structures. Experiments on three datasets show that augmenting original training data with our synthetic data, particularly when combining both quality- and diversity-oriented instances, significantly enhances the performance of existing AM models, both in full-data and low-resource settings. Moreover, the positive correlation between synthetic data volume and model performance highlights the scalability of our methods.

1 Introduction

Understanding the underlying logical reasoning embedded within natural language text is a fundamental challenge for artificial intelligence. Argument Mining (AM) aims to address this challenge by identifying and outlining the structure of arguments within a document (Stab and Gurevych, 2014, 2017; Lawrence and Reed, 2019). This process involves pinpointing key textual segments that function as argument components, such as claims stating a stance or premises providing support. It also requires determining the relations between these components, such as whether one supports or attacks another. Having this structured view of arguments provides valuable insights into logical reasoning and persuasive techniques, proving beneficial across various domains (Nguyen and Litman, 2018; Slonim et al., 2021; Fabbri et al., 2021; Elaraby and Litman, 2022).

Unfortunately, the advancement of AM models is often hindered by a significant bottleneck: the scarcity of annotated data (Dutta et al., 2022; Morio et al., 2022). The complexity inherent in AM tasks—requiring not just span identification but also fine-grained component classification and the intricate mapping of relational structures—makes manual annotation a particularly challenging and labor-intensive endeavor. Consequently, existing benchmark datasets for AM (Park and Cardie, 2018; Mayer et al., 2020), while invaluable for research, tend to be limited in size (e.g., only 400+ essays in the AAEC dataset (Stab and Gurevych, 2017)). This data limitation poses a substantial challenge for training robust, high-performance AM systems. Addressing this critical data scarcity is therefore a key motivation of this work.

Recent advancements leveraging LLMs for synthetic data generation have shown significant promise for augmenting data in NLP tasks (Havrilla et al., 2024). Inspired by this, we argue that LLM-based synthetic data generation holds potential for alleviating the data bottleneck in AM. However, generating synthetic data for AM is non-trivial due to the inherent complexity of its annotation scheme, involving component spans, types, and directed relations. This difficulty is consistent with existing work showing that off-the-shelf LLMs (e.g., GPT-4o) struggle to perform well on AM (Bao et al., 2025). Therefore, expecting LLMs to consistently generate both the argumentative text and its precise structural annotation through simple prompting is an even greater challenge.

To address these challenges, this paper explores methodologies for effectively leveraging LLMs to generate synthetic data for AM, ultimately aiming to enhance the performance of existing AM models. Inspired by Havrilla et al. (2024), we investigate

[*] Equal Contribution
[†] Corresponding Author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26581–26604
November 4-9, 2025 ©2025 Association for Computational Linguistics

two complementary synthesis perspectives, focusing on synthetic data quality and diversity, respectively. We refer to them as the **quality-oriented** synthesis (QOS) and **diversity-oriented** synthesis (DOS) approaches.

The **QOS** approach prioritizes the quality of synthetic data. It employs a structure-aware prompting technique that instructs an LLM to paraphrase existing training instances while strictly preserving their original annotation labels (i.e., component spans, types, and relations). The resulting synthetic data exhibits high label quality, mirroring the gold standard, but consequently offers limited diversity in terms of topics and argument structures. Conversely, the **DOS** approach focuses on generating synthetic data with greater diversity. It first employs an LLM to brainstorm a wide range of new topics. Then, for a new topic, it prompts the LLM to generate a new argumentative text by imitating an existing training instance. During this process, we provide the LLM with the argumentation pattern of this reference—a concise representation of the argument structure—and instruct it to modify this pattern before generating new argumentative text, thereby fostering structural diversity. Finally, these new texts are automatically annotated by a baseline AM model. This method yields synthetic data with greater topical and structural diversity, albeit with potentially less reliable labels than QOS.

We conduct comprehensive experiments on three AM datasets. Our empirical results demonstrate that augmenting the original training data with synthetic data–generated by either QOS or DOS–leads to significant performance improvements over two strong baseline AM systems, both in full-data and low-resource settings. Furthermore, we observe that combining synthetic data derived from both data synthesis approaches yields additional performance gains. Our analyses also reveal a generally positive correlation between the volume of synthetic data incorporated and the resulting model performance, highlighting the potential scalability of our methods.

2 Related Work

2.1 Argument Mining

Argument Mining is a multifaceted research area within NLP focused on automatically extracting argumentative structures from text (Lawrence and Reed, 2019). Given its inherent complexity, many efforts have sought to make the problem tractable by selectively focusing on specific sub-tasks (Chen et al., 2024; Kuribayashi et al., 2019; Li et al., 2022; Bao et al., 2021; Liang et al., 2023). These include, but are not limited to, argument component segmentation and classification, which involves locating textual spans corresponding to argumentative units like claims and premises (Moens et al., 2007; Wang et al., 2020; Cheng et al., 2022); and argumentative relation identification, which aims to determine the relations between these components (Cocarascu and Toni, 2017; Jo et al., 2021).

While such focused research has yielded valuable insights, the interdependencies between these sub-tasks have motivated a growing body of work on joint modeling and end-to-end approaches (Eger et al., 2017; Morio et al., 2022). These methods attempt to parse the entire argument structure in a single, unified framework, capturing richer contextual information and mitigating error propagation common in pipeline systems. Prominent end-to-end strategies largely fall into two main categories: those adapting traditional natural language parsing techniques (Persing and Ng, 2016; Ye and Teufel, 2021; Morio et al., 2022), and the more recent rise of generative models (Kawarada et al., 2024; Sun et al., 2024; Bao et al., 2022, 2025). In this paper, our focus aligns with these approaches, aiming to improve end-to-end analysis of argument structure.

The development of end-to-end AM systems is hampered by the scarcity of annotated corpora (Dutta et al., 2022; Morio et al., 2022). Annotating the comprehensive argument structures—encompassing argument components, their types, and their interrelations—that end-to-end approaches strive to model is a meticulous endeavor. Thus, existing datasets, while foundational, are limited in size (Park and Cardie, 2018; Accuosto and Saggion, 2020). For instance, the widely used Argument-annotated Essays Corpus (AAEC) (Stab and Gurevych, 2017) contains 402 persuasive essays; the AbstRCT corpus (Mayer et al., 2020) comprises 500 abstracts of clinical trials. This data scarcity poses a critical hurdle for developing effective end-to-end AM systems.

2.2 Synthetic Data Generation

Synthetic data generation, particularly empowered by LLMs, has become a crucial strategy for augmenting data across numerous NLP tasks (Guo and Chen, 2024; Bao et al., 2023; Wang et al., 2023; Havrilla et al., 2024). While LLMs excel at generating fluent text for tasks with simpler output structures, their application to complex, structured-output tasks like dependency parsing (Zhang et al., 2024), semantic parsing (Nicosia et al., 2021), and information extraction (Josifoski et al., 2023; Dong et al., 2023) requires more tailored approaches to handle intricate linguistic annotations.

Despite these advancements, research on synthetic data generation specifically for end-to-end AM remains relatively underexplored. To the best of our knowledge, this paper is among the pioneering efforts to systematically investigate LLM-based synthetic data generation as a means to enhance end-to-end AM systems.

3 Task Formulation

Formally, the task of end-to-end AM takes an input argumentative text X. The objective is twofold: (1) **Argument Component Identification (ACI)** aims to identify a set of argument components $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$. Each component $c_i \in \mathcal{C}$ is represented as a tuple (s_i, e_i, t_i^c), where s_i and e_i are the start and end token indices of the component's span within X, and t_i^c is its type (e.g., "Claim", "Premise") drawn from a predefined set of component types. (2) **Argumentative Relation Identification (ARI)** aims to identify a set of argumentative relations $\mathcal{R} = \{r_1, r_2, \ldots, r_{|\mathcal{R}|}\}$. Each relation $r_i \in \mathcal{R}$ connects a source component $c_i^{src} \in \mathcal{C}$ to a target component $c_i^{tgt} \in \mathcal{C}$. The relation is also assigned a type t_i^r (e.g., "Support", "Attack") from a predefined set of relation types. Thus, r_i can be represented as a tuple $(c_i^{src}, c_i^{tgt}, t_i^r)$.

Let $\mathcal{M}$ denote an AM model designed to perform this end-to-end task, typically trained on a manually annotated dataset $\mathcal{D}_{\text{orig}} = \{(X_k, \mathcal{C}_k, \mathcal{R}_k)\}_{k=1}^{N_{\text{orig}}}$. The core objective of this paper is to investigate methods for generating synthetic training data, $\mathcal{D}_{\text{syn}}$, by leveraging LLMs. Our goal is to demonstrate that by augmenting the original training data with these synthetic instances, we can significantly enhance the performance of existing AM models.

4 Methodology

We explore two complementary synthetic data generation approaches for AM: **Quality-Oriented Synthesis (QOS)**, which prioritizes the quality of the synthetic data by ensuring high fidelity to gold-standard annotations, and **Diversity-Oriented Synthesis (DOS)**, which focuses on the diversity of the synthetic data in terms of topics and argument structures. Figure 1 provides an overview of our proposed framework.

4.1 Quality-Oriented Synthesis (QOS)

The QOS approach employs a paraphrase-based strategy. It aims to generate synthetic data that is lexically and syntactically varied from original training instances, while meticulously preserving both their original semantics and the integrity of their human-annotated argument structures. This ensures that the synthetic data exhibits high-quality labels, thereby minimizing the risk of injecting significant label noise into the training process.

Structure-aware Paraphrasing. QOS leverages a carefully designed prompt that instructs an LLM to perform structure-aware paraphrasing. As shown in Figure 1 (a)[1], for each original training instance $(X_k, \mathcal{C}_k, \mathcal{R}_k) \in \mathcal{D}_{\text{orig}}$, we format the input for the LLM as a JSON object. In this object, we explicitly separate the context text from the argument components text. The context text contains placeholders (e.g., "[AC1]", "[AC2]") indicating the positions where argument components are to be inserted. The "argument_component_info" part of the JSON object provides the actual content and pre-defined type of each argument component. The LLM is then tasked to paraphrase both the context and the content of each argument component. In essence, the LLM aims to enhance linguistic diversity while strictly preserving the original meaning, component types, and overall textual coherence, returning the output in the same JSON format.

This explicit separation of context and argument components offers significant advantages. By treating argument components as distinct units inserted into placeholders, their textual boundaries are inherently maintained during paraphrasing. This circumvents the issue in free-form paraphrasing where original span annotations often become misaligned in the paraphrased text. Notably, argumentative relations are not explicitly provided to the LLM, due to the challenge of representing complex relational structures in a textual format that LLMs can robustly interpret. Instead, we assume these relations are implicitly encoded within the context and components' semantics, and are preserved through strict meaning preservation during paraphrasing.

Annotation Inheritance. Once the LLM generates the paraphrased JSON output, the new synthetic text can be reconstructed by inserting the

[1]The specific prompt is shown in Figure 9 of Appendix N.

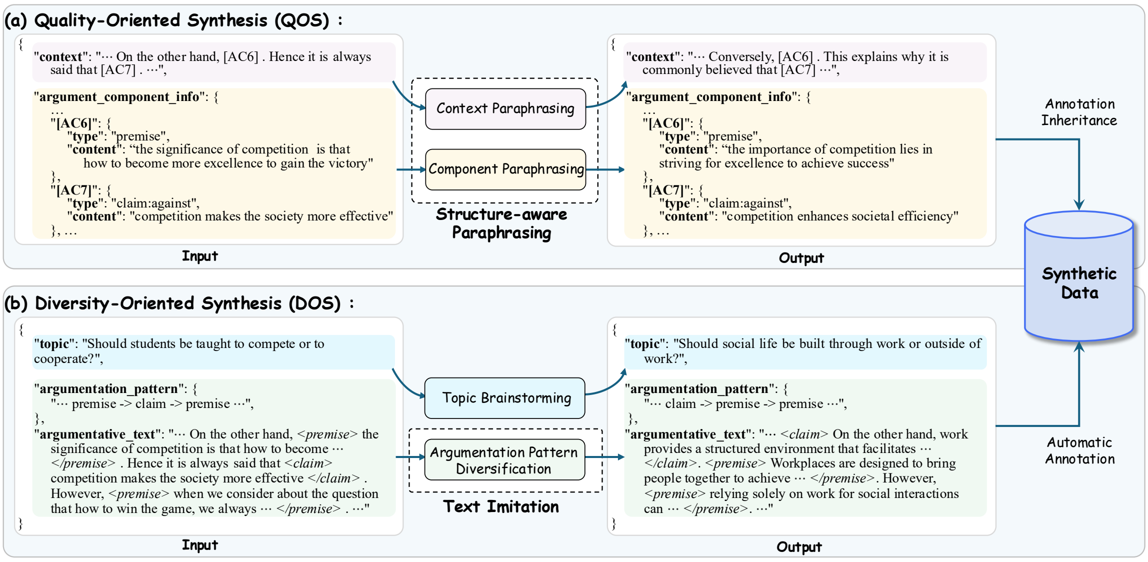


Figure 1: Overview of our synthetic data generation framework. (a) Quality-Oriented Synthesis (QOS) inputs an original training sample (Input) and uses structure-aware paraphrasing with inheriting annotations to produce a synthetic sample (Output). (b) Diversity-Oriented Synthesis (DOS) first employs topic brainstorming to generate diverse new topics. Then, for a new topic and an original sample (Input), it generates a diversified argumentation pattern and imitates a novel argumentative text (Output), which is automatically annotated by a baseline AM model.

paraphrased component contents into their corresponding placeholders within the paraphrased context. The span information for each component is directly derivable from this process. Critically, the labels of both the component types and the argumentative relations are inherited from the original training instance's annotations. The resulting synthetic dataset generated via QOS, denoted as $\mathcal{D}_{\text{syn-qos}}$, therefore exhibits high label quality[2]. However, as a consequence of its reliance on paraphrasing existing data, $\mathcal{D}_{\text{syn-qos}}$ offers limited novelty in terms of topics and argument structures when compared to the original training set.

4.2 Diversity-Oriented Synthesis (DOS)

DOS prioritizes the generation of synthetic data that exhibits greater novelty, particularly in terms of the topics and the underlying argument structures. Its goal is to expose the AM model to a wider spectrum of argumentation and reasoning patterns than those present in $\mathcal{D}_{\text{orig}}$. This is achieved through a three-stage process: first brainstorming diverse topics, then generating new argumentative texts by text imitation while encouraging structural variation, and finally automatically annotating the generated texts using a baseline AM model trained on existing training data.

Diverse Topic Brainstorming. The initial stage of DOS aims to generate a pool of novel topics to serve as the foundation for new argumentative texts. This is achieved by prompting an LLM to brainstorm diverse topics, drawing inspiration from the thematic domains in the original training data $\mathcal{D}_{\text{orig}}$. Specifically, we first randomly sample a small set of existing topics from $\mathcal{D}_{\text{orig}}$. For the AAEC dataset (Stab and Gurevych, 2017), we view essay titles as topics. For other datasets such as CDCP (Park and Cardie, 2018) or AbstRCT (Mayer et al., 2020), where topics are not readily available, we prompt an LLM to summarize a concise topic for each argumentative text[3]. Once these topics are acquired, a randomly selected subset is provided as input to an LLM prompt designed for topic brainstorming[4]. This prompt instructs the LLM to generate a specified number of new, diverse topics that are thematically related to the provided examples. This ensures the newly generated topics maintain relevance to the dataset's domain yet offer sufficient novelty.

Text Imitation with Argumentation Pattern Diversification. Following the generation of diverse topics, this stage focuses on creating new argumentative texts for each new topic. This process employs an imitation prompt, as exemplified

[2]This high quality is empirically evidenced in Appendix D.

[3]The specific prompt is shown in Figure 10 of Appendix N.

[4]The specific prompt is shown in Figure 11 of Appendix N.

in Figure 1 (b)[5], which guides an LLM to generate a new argumentative text by taking as input both a new topic and an example training instance randomly sampled from $\mathcal{D}_{\text{orig}}$. This example instance is provided in a JSON format, containing its original topic, its full argumentative text, and a central element we introduce: its "argumentation pattern". The argumentation pattern serves as a simplified representation of the reference instance's argument structure, formalized as a sequence of its component types (e.g., "claim $\rightarrow$ premise $\rightarrow$ premise"). To facilitate the LLM's understanding of this pattern and its direct correspondence to the actual text, the argumentative text of the example instance within the JSON input includes inline tags (e.g., <premise>...</premise>) that explicitly mark the spans and types of its argument components. These tags will be removed after generation.

Furthermore, a key mechanism for fostering diversity lies in the explicit instruction within the prompt: the LLM is required to first modify the provided argumentation pattern to create a new one (e.g., from "claim $\rightarrow$ premise $\rightarrow$ premise" to "premise $\rightarrow$ claim $\rightarrow$ premise"). Specifically, the LLM is encouraged to introduce variations by changing, adding, or removing components in the provided argumentation pattern. Then, it generates a new argumentative text following this modified pattern. This strategy can prevent the LLM from merely copying the reasoning flow of the reference text, thereby promoting the generation of more diverse argument structures.

Automatic Annotation. Once the new argumentative text is generated through the imitation process, it must be annotated to function as a training instance. This annotation is performed automatically by employing a baseline AM model, denoted $\mathcal{M}_{\text{base}}$. This model is trained exclusively on the original human-annotated data $\mathcal{D}_{\text{orig}}$. The combination of the new argumentative text and its automatically annotated argument components and argumentative relations then constitutes a synthetic training instance. We denote the synthetic dataset as $\mathcal{D}_{\text{syn-dos}}$. This method yields data with the desired topical and structural diversity[6]. However, the quality of these automatically generated labels is inherently contingent upon the performance of the baseline model, potentially introducing a degree of label noise. This represents a deliberate trade-off to achieve greater diversity compared to QOS.

4.3 Training Strategy with Synthetic Data

We use the synthetic datasets $\mathcal{D}_{\text{syn-qos}}$ and $\mathcal{D}_{\text{syn-dos}}$ to augment the original training data $\mathcal{D}_{\text{orig}}$ using a two-stage training strategy. First, the AM model $\mathcal{M}$ is trained on a synthetic dataset–either $\mathcal{D}_{\text{syn-qos}}$, $\mathcal{D}_{\text{syn-dos}}$, or a mixture of them–leveraging the larger volume for initial learning. Subsequently, the model is further trained on the original human-annotated dataset $\mathcal{D}_{\text{orig}}$ to refine predictions with high-quality labels and align with the target data distribution. The effectiveness of this approach is evaluated on the standard test sets of the respective datasets.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our methods on three widely-used AM datasets: **AAEC** (Stab and Gurevych, 2017), **CDCP** (Park and Cardie, 2018), and **AbstRCT** (Mayer et al., 2020). Details of these datasets are shown in Appendix A.1. Our experiments are performed under two training data settings: using 100% of the original training data and a low-resource setting with only 5% of the training data, which is randomly sampled. For all experiments, we follow the train, validation, and test splits used in prior work (Morio et al., 2022).

Implementation Details. For all LLM-based synthetic data generation approaches, we utilize the gpt-4o-2024-05-13 model. In Appendix H, we also experiment with replacing GPT-4o with open-source LLMs. In our main experiments[7], the volume of synthetic data employed for augmentation is consistently set to twice the size of the original training data used in a given setting. For the combined QOS+DOS approach in these experiments, the synthetic data is composed of 25% QOS and 75% DOS instances[8]. Specific hyper-parameters are provided in Table 5 of Appendix A.2. For all experiments, we report the average results over five runs with different random seeds. We use three micro F1 metrics adopted from previous work (Morio et al., 2022): F1$_{span}$ for identifying argument component spans, F1$_{aci}$ for further classifying their types, and F1$_{ari}$ for detecting relations between them.

[5]The specific prompt is shown in Figure 12 of Appendix N.

[6]This diversity is empirically evidenced in Appendix 5.8.

[7]Code: https://github.com/YuqiHuang2003/QOS_DOS

[8]This ratio is determined by experiments in Section 5.7.

AM Model	Setting	Method	AAEC				AbstRCT				CDCP			
			$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg	$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg	$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg
ST	100%	Origin	84.31	75.56	53.49	71.12	70.15	64.57	36.01	56.91	82.01	67.88	30.85	60.25
		EDA	84.81	76.30	53.51	71.54	70.88	65.14	37.65	57.89	82.26	67.92	30.75	60.31
		FTGA	85.32	76.68	53.93	71.98	71.61	65.21	37.74	58.19	81.78	67.82	32.73	60.78
		JTLS	85.21	76.03	53.89	71.71	71.29	65.55	38.22	58.45	82.11	67.97	30.73	60.27
		QOS	85.23*	77.08*	56.68*	73.00*	71.80	65.60	38.47*	58.62*	82.77	67.45	33.62*	61.28*
		DOS	84.80	76.11	55.44*	72.12*	72.76*	66.28*	38.72*	59.25*	82.40	67.84	32.47*	60.90
		QOS+DOS	**86.38***	**78.04***	**56.94***	**73.79***	**74.10***	**67.49***	**40.03***	**60.54***	**83.76***	**69.01***	**36.10***	**62.96***
	5%	Origin	67.91	50.33	16.91	45.05	57.52	50.76	21.57	43.28	76.54	49.51	4.03	43.36
		EDA	66.72	48.40	16.43	43.85	58.35	51.79	21.59	43.91	76.24	48.28	4.27	42.93
		FTGA	69.17	52.84	20.83	47.61	56.96	50.93	23.02	43.64	76.19	49.76	5.07	43.67
		JTLS	70.83	52.42	20.18	47.81	59.47	50.95	22.27	44.23	77.31	49.44	3.83	43.52
		QOS	70.90*	55.55*	**24.49***	50.31*	63.18*	56.66*	27.84*	49.23*	76.60	50.27	6.49*	44.45*
		DOS	71.47*	54.39*	22.49*	49.45*	61.81*	54.83*	26.93*	47.86*	76.98	52.12*	8.06*	45.72*
		QOS+DOS	**72.10***	**55.60***	23.68*	**50.46***	**64.36***	**57.93***	**28.08***	**50.12***	**77.34***	**54.19***	**8.23***	**46.59***
UniASA†	100%	Origin	87.12	76.37	54.72	72.74	78.88	71.73	37.10	62.57	81.93	66.72	27.12	58.59
		EDA	87.42	76.93	55.46	73.27	77.85	72.02	37.39	62.42	81.68	66.72	26.13	58.18
		FTGA	87.58	76.92	55.21	73.24	79.11	71.78	38.48	63.12	82.13	67.57	26.28	58.66
		JTLS	87.48	76.86	55.07	73.14	79.19	72.20	38.40	63.26	82.54	67.51	26.63	58.89
		QOS	87.40	77.74*	56.57*	73.90*	80.39*	72.64*	38.87*	63.97	81.48	66.34	26.12	57.98
		DOS	87.79*	77.52	56.59*	73.97*	80.33*	73.06*	40.24*	64.54*	81.57	67.30*	26.77	58.55
		QOS+DOS	**87.96***	**78.22***	**56.91***	**74.36***	**80.94***	**73.11***	**42.25***	**65.43***	**82.55***	**67.82***	**27.96***	**59.44***
	5%	Origin	51.44	31.47	1.06	27.99	60.36	46.98	7.96	38.43	69.21	37.06	2.05	36.11
		EDA	58.62	38.02	2.06	32.90	61.14	44.07	8.23	37.81	70.55	40.71	3.82	38.36
		FTGA	66.79	44.41	3.69	38.30	60.98	48.83	10.58	40.13	71.34	39.36	4.46	38.39
		JTLS	55.78	35.69	2.08	31.18	60.71	45.98	11.61	39.43	71.11	42.86	2.71	38.89
		QOS	74.73*	50.95*	8.45*	44.71*	61.65	54.45*	21.29*	45.80*	71.88*	47.14*	**8.66***	42.56*
		DOS	72.58*	48.01*	7.30*	42.63*	61.66*	50.79	16.18*	42.88*	71.80*	46.16*	6.55*	41.50*
		QOS+DOS	**75.34***	**53.62***	**12.05***	**47.00***	**63.63***	**56.78***	**23.97***	**48.13***	**73.06***	**48.10***	8.35*	**43.17***

Table 1: Main experimental results on the AAEC, AbstRCT, and CDCP datasets under full (100%) and low-resource (5%) training data settings. "Origin" indicates models trained solely on the original training data. UniASA† denotes the single-view version of the UniASA model; we use this variant as its performance is generally comparable to the multi-view version while being significantly less time-consuming. The best results for each metric within each setting are highlighted in **bold**. "Avg" is the arithmetic mean of the three F1 scores. "*" indicates the results obtained by our methods are statistically significant (p-value < 0.05) based on a paired t-test.

5.2 Baseline AM Models

End-to-end AM models predominantly fall into two categories: those adapting traditional natural language parsing techniques and, more recently, generative models. To evaluate the utility of synthetic data across these two main categories, we measure performance improvements on two strong AM models, each representative of one category: (1) **ST** (Morio et al., 2022): A strong representative of models adapting natural language parsing techniques[9]. (2) **UniASA** (Bao et al., 2025): A generative model that formulates AM as a sequence-to-sequence task[10]. Note that, for the automatic annotation step in DOS, $\mathcal{M}_{\text{base}}$ refers to the specific model being evaluated (either ST or UniASA).

5.3 Compared Methods

Given the scarcity of existing work on synthetic data generation specifically tailored for end-to-end AM, we evaluate our QOS and DOS methods against several widely applicable data augmentation and synthetic data generation methods: (1) **Easy Data Augmentation (EDA)** enhances argumentative text by applying lexical operations. (2) **Few-shot Text Generation and Auto-annotation (FTGA)** uses GPT-4o with few-shot examples to generate new argumentative texts, which are then annotated by a baseline AM model. (3) **Joint Text and Label Synthesis (JTLS)** prompts GPT-4o with text-annotation pairs to simultaneously generate new argumentative texts and their full structural annotations. Details of these methods are provided in Appendix A.3.

[9] https://github.com/hitachi-nlp/graph_parser

[10] https://github.com/HITSZ-HLT/UniASA

5.4 Main Results

Table 1 shows results for our proposed synthetic data generation methods (QOS, DOS, QOS+DOS) and other compared methods on ST and UniASA models, revealing several key points:

Enhanced Performance with QOS and DOS. Augmenting original training data with QOS or DOS synthetic instances significantly improves performance for both ST and UniASA models. This enhancement is consistent across all three datasets and in both the full data and low-resource settings when compared to models trained solely on the

Set.	Method	AAEC				AbstRCT				CDCP			
		$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg
100%	QOS	85.23	**77.08**	**56.68**	**73.00**	71.80	**65.60**	**38.47**	**58.62**	82.77	67.45	**33.62**	61.28
	w/o Struct.-aware Paraph.	**85.34**	75.63	55.99	72.32	**72.64**	64.19	38.26	58.36	**82.98**	**68.19**	33.60	**61.59**
	w/o Original Annotations	84.92	76.23	54.47	71.87	68.02	61.87	36.82	55.57	81.97	67.77	31.28	60.34
	DOS	84.80	76.11	55.39	72.10	**72.76**	**66.28**	**38.72**	**59.25**	**82.40**	67.84	**32.47**	**60.90**
	w/o Topic Brainstorming	84.92	76.30	**55.60**	**72.27**	69.58	63.42	37.59	56.86	81.16	67.31	31.93	60.13
	w/o Argumentation Pattern	**85.52**	**76.35**	54.90	72.26	70.32	64.53	37.36	57.40	82.30	**68.30**	30.87	60.49
5%	QOS	**70.90**	**55.55**	**24.49**	**50.31**	**63.18**	**56.66**	**27.84**	**49.23**	76.60	**50.27**	**6.49**	**44.45**
	w/o Struct.-aware Paraph.	69.76	53.51	23.06	48.78	56.74	50.51	23.30	43.52	76.53	49.46	4.58	43.52
	w/o Original Annotations	68.82	49.86	18.06	45.58	55.01	49.08	21.94	42.01	**76.87**	48.59	4.97	43.48
	DOS	**71.47**	**54.39**	**22.49**	**49.45**	**61.81**	**54.83**	**26.93**	**47.86**	**76.98**	**52.12**	**8.06**	**45.72**
	w/o Topic Brainstorming	70.24	53.17	21.92	48.44	59.22	52.71	23.78	45.24	76.64	51.26	7.15	45.02
	w/o Argumentation Pattern	69.50	53.16	21.86	48.17	57.57	51.32	23.97	44.29	76.67	50.39	4.88	43.98

Table 2: Ablation study of QOS and DOS on the ST model. "Set." is short for "Setting".

original data ("Origin").

Synergistic Gains from Mixing QOS and DOS. The mixture of QOS and DOS data ("QOS+DOS") generally yields the most substantial performance gains, highlighting a powerful synergy between the two approaches. By leveraging both high-quality paraphrases of existing training data and novel argumentative texts with diverse topics and argument structures, the AM models are exposed to a richer and more comprehensive training signal, leading to superior performance.

Pronounced Benefits in Low-Resource Scenarios. The advantages of our methods are particularly striking in the 5% low-resource setting. Here, the performance uplift from QOS, DOS, and especially QOS+DOS over the "Origin" is often more pronounced than with full data. This crucial finding demonstrates the potential of our methods to effectively mitigate the challenge of data scarcity, enabling performant AM systems even with limited human-annotated data.

Superiority over Compared Data Augmentation Methods. Our methods, particularly "QOS+DOS", generally exhibit superior performance compared to the other data augmentation techniques (EDA, FTGA, and JTLS). While these compared methods offer some improvements over the "Origin", gains from our methods are typically more consistent and significant across all datasets, AM models, and data settings. This suggests that our tailored designs for QOS and DOS are more effective for the complex task of end-to-end AM.

5.5 Ablation Study

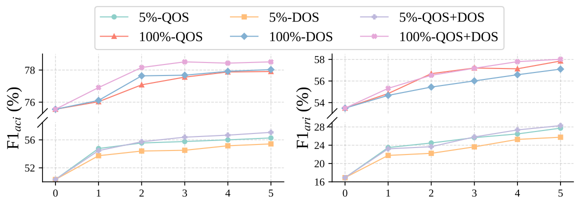
To validate the effectiveness of the key components of our proposed QOS and DOS approaches, we conduct an ablation study on ST (Table 2).

QOS Ablation Results. First, "*w/o* **Struct.-aware Paraph.**" means that the LLM paraphrases the entire text without separating context and components, with annotations generated by the baseline AM model. This has mixed effects in the 100% setting but significantly hurts performance in the 5% low-resource setting. Second, "*w/o* **Original Annotations**" denotes that the LLM performs structure-aware paraphrasing, but annotations come from the baseline AM model instead of gold-standard labels. This consistently reduces performance across all settings, underscoring the importance of high-quality labels in QOS.

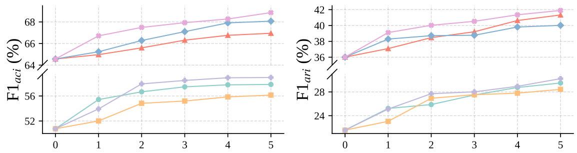
DOS Ablation Results. First, "*w/o* **Topic Brainstorming**" denotes that the LLM generates new texts using only topics present in the original training data. This lowers performance, especially in the 5% setting, highlighting the value of topical diversity. Second, "*w/o* **Argumentation Pattern**" removes argumentation pattern guidance and diversification instruction from the text imitation prompt (Figure 12). This leads to a decrease in overall performance, confirming the benefit of the explicit pattern diversification instruction in DOS.

5.6 Impact of Synthetic Data Scale

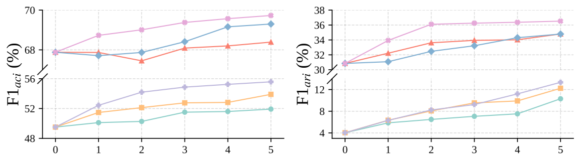
We examine the impact of varying the volume of synthetic data (e.g., 1x, 2x the original training data size) on the ST model's performance. Figure 2 presents these results for QOS, DOS, and their combination (25% QOS + 75% DOS as in the main experiments). Generally, increasing the volume of synthetic data consistently improves performance. This positive trend is observed for all synthetic data types and across all datasets and training settings. The benefits are particularly pronounced in the low-resource setting. Crucially, the combi-



(a) AAEC



(b) AbstRCT



(c) CDCP

Figure 2: Impact of synthetic data scale on the ST model performance across three datasets for $F1_{span}$, $F1_{aci}$, and $F1_{ari}$ metrics. The x-axis indicates the synthetic data volume, expressed as N times the size of the original training data used. Results are shown for both low-resource (5%) and full-data (100%) settings.

nation of QOS and DOS synthetic data generally outperforms using QOS or DOS data individually.

5.7 Impact of QOS and DOS Mixture Ratios

In the main experiments, "QOS+DOS" utilizes a mixture of 25% QOS and 75% DOS data, which generally yielded strong performance gains. Here, we further analyze different mixture ratios of QOS and DOS data, maintaining a total synthetic data volume of 2x the original training data. Figure 3 shows results for the ST model on CDCP. Results for AAEC and AbstRCT are shown in Figure 6 (Appendix B). It can be seen that a blend often outperforms using solely QOS or DOS data. Specifically, a mixture with 25% QOS data and 75% DOS data frequently yields the best performance.

5.8 Diversity Analysis of DOS Data

To better understand how DOS enhances diversity in synthetic data, we conduct a visual analysis of both topical and structural diversity in the DOS data.

Topical Diversity. We use Instructor embeddings (Su et al., 2023) for text representation and t-SNE for visualization. Figure 4 illustrates the topic distributions of original training data compared to DOS data across the three datasets. We can see that the topics of DOS data consistently show more expansive and evenly distributed patterns in the embedding space. For AAEC (Figure 4a), DOS topics cover a similar semantic space but with increased density. More notably, for AbstRCT (Figure 4b) and CDCP (Figure 4c), DOS significantly extends the topical range, as evidenced by the wider dispersion of points in the visualization. These results demonstrate that the topic brainstorming process in DOS substantially enhances topical diversity across all datasets.

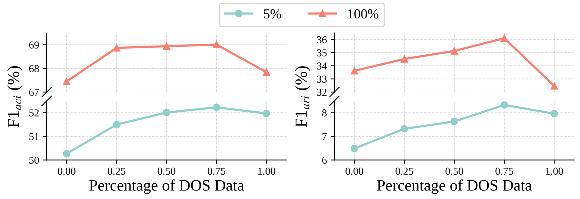
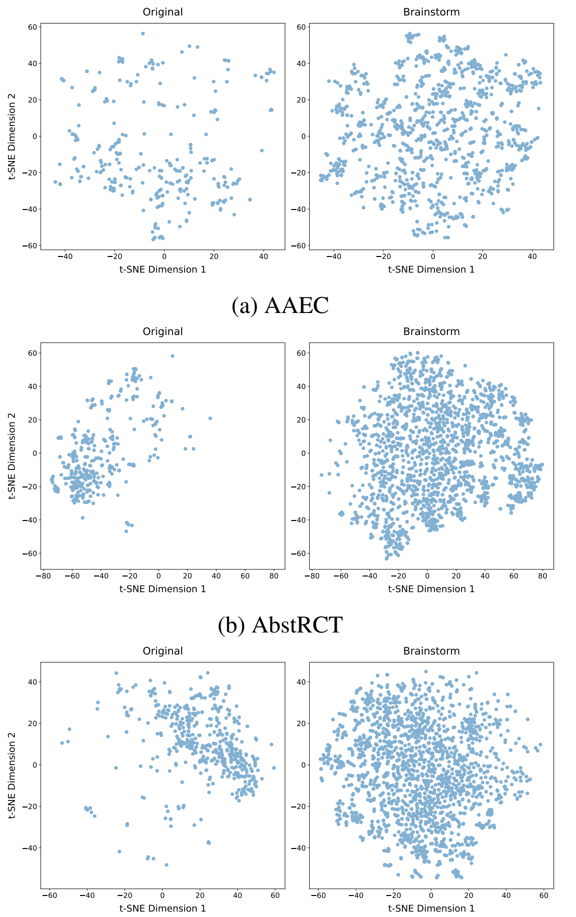


Figure 3: Impact of QOS and DOS mixture ratios for the ST model on CDCP. The x-axis represents the percentage of DOS data, while the remaining portion is QOS data.

Argument Structural Diversity. To analyze the diversity of argument structures, we first view the argument structure annotations as argument graphs, and employ graph2vec (Narayanan et al., 2017) to embed the graphs and visualize them using t-SNE. Importantly, this graph embedding method considers only the structural information composed of argument component types and their argumentative relations (including relation types), without incorporating any textual semantic information. Figure 5 shows the results of both the original training data and the DOS data. It can be seen that the argument structures in the original training data typically form concentrated clusters, indicating limited structural patterns. In contrast, DOS data exhibits significantly wider distribution, expanding beyond the boundaries of original structures across all datasets. These visualizations confirm that our DOS approach effectively generates texts with diverse argument structure annotations.

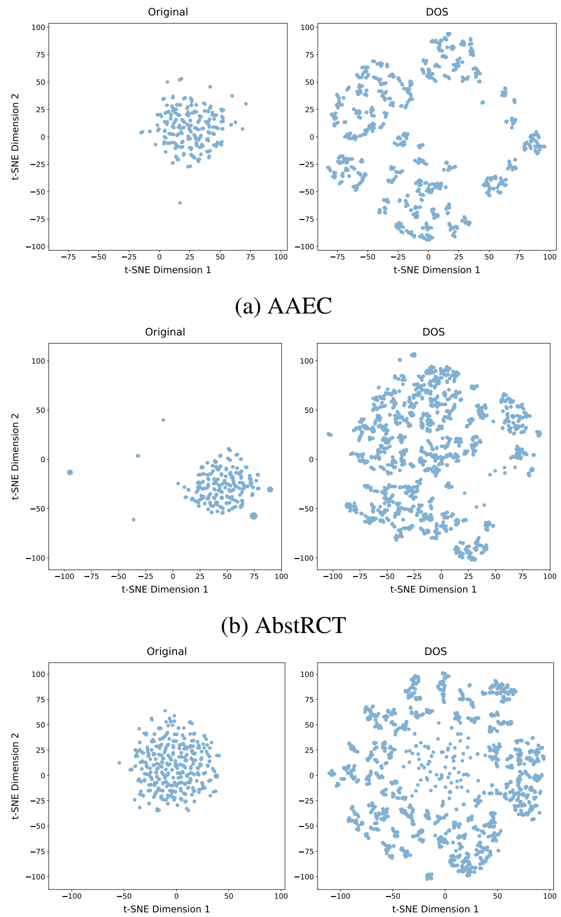
5.9 Training Time Analysis

The introduction of synthetic data inevitably increases model training time. All AM models are trained on a single Tesla A100 GPU. Under the full-data setting, training the ST model on original data takes approximately 40 minutes, while train-



(a) AAEC (b) AbstRCT (c) CDCP

Figure 4: t-SNE visualization of topic distributions in original training data (left) versus DOS-generated data (right) under full-data (100%) setting. We use Instructor (Su et al., 2023) for topic embedding.



(a) AAEC (b) AbstRCT (c) CDCP

Figure 5: t-SNE visualization of argument structure distributions in original training data (left) versus DOS-generated data (right) under full-data (100%) setting. We use graph2vec (Narayanan et al., 2017) for graph embedding.

ing with augmented synthetic data extends this to about 1 hour. For the UniASA model, the original training completes in roughly 2 hours, increasing to about 3 hours with synthetic data. This increase in training time represents a worthwhile investment given the substantial performance improvements demonstrated throughout our experiments. Importantly, our methods do not introduce any additional overhead during model inference.

5.10 Further Analyses

We conduct additional analyses, detailed in the appendices. These include deeper investigations into synthetic data quality (Appendix D), case study (Appendix L), error analysis (Appendix M), and integration with self-training (Appendix C). We also verify the generalizability of our methods on an additional AM model (Appendix E), two more datasets (G), and in a cross-dataset transfer setting (Appendix I). For reproducibility, the detailed prompts are available in Appendix N.

6 Conclusion

This paper investigates leveraging LLMs for synthetic data generation to alleviate data scarcity in AM. We propose two complementary approaches: quality-oriented synthesis, which focuses on label fidelity through structure-aware paraphrasing, and diversity-oriented synthesis, which emphasizes topical and structural novelty via topic brainstorming and argumentation pattern diversification. Extensive experiments on three datasets demonstrate that augmenting training data with instances from either QOS or DOS significantly enhances the performance of existing AM models, particularly in low-resource scenarios. Also, combining both approaches yields further synergistic improvements.

Limitations

While our proposed synthetic data generation methods demonstrate significant promise for alleviating data scarcity in AM, we acknowledge certain limitations.

First, incorporating synthetic data, despite its benefits, inevitably increases the overall model training time due to the larger volume of training instances. We discuss this in Section 5.9. Second, although effective, both our QOS and DOS approaches currently rely on some existing human-annotated data as a reference. Generating high-quality, structured argumentative data entirely from scratch, without any reference to gold-standard annotations, remains a significant challenge for future work. Finally, our study is primarily focused on the AM domains represented by the existing benchmark datasets. Expanding these synthetic data generation techniques to more diverse, open-domain AM scenarios presents an important avenue for future research.

Ethics Statement

This work uses publicly available datasets widely adopted in previous AM studies. Our use of these datasets, and all other software and resources, strictly complies with their respective licenses and intended purposes. The chosen datasets are understood to be free of personally identifiable information and offensive content. We acknowledge that using LLMs for synthetic data generation may introduce potential risks such as bias amplification, unintended factual inaccuracies. However, thoroughly addressing these risks falls beyond the scope of our current AM task focus. AI assistants are employed solely for grammar checking and text polishing in manuscript preparation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China 62176076 and 62576120, Natural Science Foundation of Guang Dong 2023A1515012922, the Major Key Project of PCL2023A09, CIPSC-SMP-ZHIPU Large Model Cross-Disciplinary Fund ZPCG20241119405 and Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No.2024ZD020.

References

Pablo Accuosto and Horacio Saggion. 2020. Mining arguments in scientific abstracts with discourse-level embeddings. *Data Knowl. Eng.*, 129:101840.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6354–6364, Online. Association for Computational Linguistics.

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

Jianzhu Bao, 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, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *Preprint*, arXiv:2501.08828.

Kuicai Dong, Aixin Sun, Jung-jae Kim, and Xiaoli Li. 2023. Open information extraction via chunks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15390–15404, Singapore. Association for Computational Linguistics.

Jiawei Du, Xin Zhang, Juncheng Hu, Wenxing Huang, and Joey Tianyi Zhou. 2024. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *Advances in Neural Information Processing Systems*, volume 37, pages 119443–119465. Curran Associates, Inc.

Subhabrata Dutta, Jeevesh Juneja, Dipankar Das, and Tanmoy Chakraborty. 2022. Can unsupervised knowledge transfer from social discussions help argument mining? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7774–7786, Dublin, Ireland. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Xu Guo and Yiqiang Chen. 2024. Generative AI for synthetic data generation: Methods, challenges and the future. *CoRR*, abs/2403.04190.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, Duy Phung, Maia Iyer, Dakota Mahan, Chase Blagden, Srishti Gureja, Mohammed Hamdy, Wen-Ding Li, Giovanni Paolini, Pawan Sasanka Ammanamanchi, and Elliot Meyerson. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *CoRR*, abs/2412.02980.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2002–2014, St. Julian's, Malta. Association for Computational Linguistics.

Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An empirical study of span representations in argumentation structure parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4691–4698, Florence, Italy. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. 2025. CoIR: A comprehensive benchmark for code information retrieval models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22074–22091, Vienna, Austria. Association for Computational Linguistics.

Yinzi Li, Wei Chen, Zhongyu Wei, Yujun Huang, Chujun Wang, Siyuan Wang, Qi Zhang, Xuanjing Huang, and Libo Wu. 2022. A structure-aware argument encoder for literature discourse analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7093–7098, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jingcong Liang, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Hi-ArG: Exploring the integration of hierarchical argumentation graphs in language pretraining. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14606–14620, Singapore. Association for Computational Linguistics.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pages 225–230. ACM.

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.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning Distributed Representations of Graphs. *arXiv e-prints*, arXiv:1707.05005.

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5892–5899. AAAI Press.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and 34 others. 2021. An autonomous debating system. *Nat.*, 591(7850):379–384.

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

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

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Yang Sun, Guanrong Chen, Caihua Yang, Jianzhu Bao, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. 2024. Discourse structure-aware prefix for generation-based end-to-end argumentation mining. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11597–11613, Bangkok, Thailand. Association for Computational Linguistics.

Hao Wang, Zhen Huang, Yong Dou, and Yu Hong. 2020. Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5480–5493, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 257–268, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rui Wang, Jianzhu Bao, Fei Mi, Yi Chen, Hongru Wang, Yasheng Wang, Yitong Li, Lifeng Shang, Kam-Fai Wong, and Ruifeng Xu. 2023. Retrieval-free knowledge injection through multi-document traversal for dialogue models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6619, Toronto, Canada. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678, Online. Association for Computational Linguistics.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. LLM-assisted data augmentation for Chinese dialogue-level dependency parsing. *Computational Linguistics*, 50(3):867–891.

Appendix

A Additional Experimental Settings

A.1 Dataset Information

We conduct experiments on three AM datasets: the **Argument-annotated Essays Corpus (AAEC)** (Stab and Gurevych, 2017), comprising persuasive essays; the **Consumer Debt Collection Practices (CDCP)** corpus (Park and Cardie, 2018), containing user comments on e-rulemaking; and **AbstRCT** (Mayer et al., 2020), consisting of abstracts from clinical trials. Details of these datasets are presented in Table 4.

A comprehensive list of all component and relation types for each dataset is shown in Table 3. It is important to note that for the data synthesis process on the AAEC dataset, we simplify the annotation by treating both “Claim:For” and “Claim:Against” as a single, unified “Claim” type. This simplification is adopted because the LLM struggles to reliably distinguish between these two nuanced subtypes during generation. This does not compromise the final AM model’s predictive capabilities, as the model will finally be trained on the original training data.

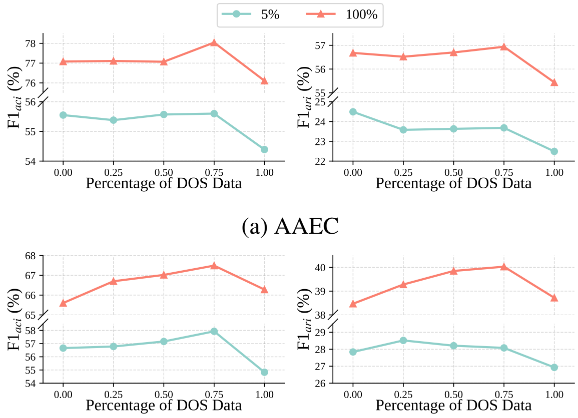
A.2 Hyper-parameters of the Main Experiments

The hyper-parameters of the main experiments are shown in Table 5. For training on the original training data, we adhere to the hyperparameter configurations reported in the original papers of the AM models. For AAEC, we conduct experiments at the essay level, as this represents a more complete and challenging setting.

A.3 Details about the Compared Methods

We compare our data synthesis methods with the following data augmentation and synthetic data generation methods:

- **Easy Data Augmentation (EDA)**: This method applies lexical operations to the argumentative text. While the original EDA (Wei and Zou, 2019) framework includes synonym replacement, random insertion, deletion, and swapping, the latter three operations can inevitably disrupt the original argument structure annotations. Therefore, we exclusively apply synonym replacement to the argumentative text.



(a) AAEC

(b) AbstRCT

Figure 6: Impact of QOS and DOS mixture ratios for the ST model on AAEC and AbstRCT. The x-axis represents the percentage of DOS data, while the remaining portion is QOS data. The total synthetic data volume is 2x the original training data.

- **Few-shot Text Generation and Auto-annotation (FTGA)**: This method prompts GPT-4o with a few example argumentative texts from $\mathcal{D}_{\text{orig}}$ to generate new texts, which are then automatically annotated by a baseline AM model.

- **Joint Text and Label Synthesis (JTLS)**: This approach prompts GPT-4o with a few examples of texts paired with their full structural annotations from $\mathcal{D}_{\text{orig}}$ to directly generate new texts along with their corresponding structural annotations.

B Additional Results for QOS and DOS Mixture Ratios

This section provides supplementary results to the analysis in Section 5.7, illustrating the impact of varying QOS and DOS mixture ratios on the ST model’s performance for the AAEC and AbstRCT datasets. Figure 6 shows these results. Similar to the findings for CDCP, specific blends of QOS and DOS data often yield better performance than using either synthetic data type exclusively.

C Analysis of Iterative Self-Training

Self-training is an effective approach that leverages additional data. We therefore further explore integrating our data synthesis approaches with self-training. Notably, our DOS approach involves automatically annotating new synthetic texts using a baseline AM model (pseudo-labeling), a process

Dataset	Component Types	Relation Types
AAEC	*MajorClaim*, *Claim:For*, *Claim:Against*, *Premise*	*Support*, *Attack*
CDCP	*Value*, *Fact*, *Policy*, *Testimony*, *Reference*	*Reason*, *Evidence*
AbstRCT	*MajorClaim*, *Evidence*, *Claim*	*Support*, *Attack*, *Partial-Attack*

Table 3: Component and relation types as defined in each dataset.

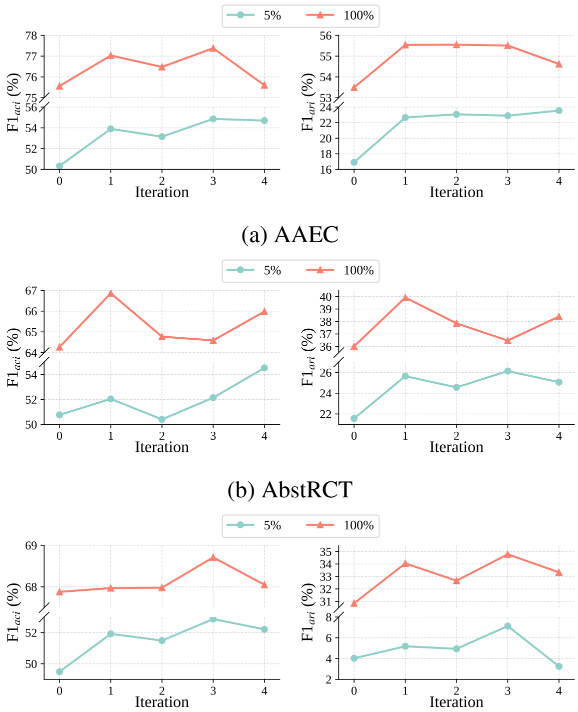
Dataset	# Instance	# Components	# Relations
AAEC	402	6,089	3,832
CDCP	731	4,779	1,353
AbstRCT	500	3,279	2,060

Table 4: Statistics of the datasets used in our experiments.

that can be iteratively applied. In this iterative scheme, the model trained on data from preceding iterations is used to generate pseudo-labels for new synthetic texts, thus leveraging model improvements across iterations.

Our iterative self-training primarily follows the framework presented in Wang et al. (2021). The specific setup is as follows: We start with the model trained solely on the original data D_{orig} as the initial model (Iteration 0). For each subsequent iteration i (where $i \geq 1$), we employ the model trained in iteration $i-1$ to pseudo-label a set of newly generated diverse texts (created using the text generation process from the DOS approach). We then filter these pseudo-labeled instances based on their confidence scores, retaining only those with confidence values between 0.7 and 0.9 to avoid both noisy low-confidence samples and overly simplistic high-confidence ones. The confidence for each instance is determined by averaging all its predicted classification probabilities by the AM model. For training the model in iteration i, we combine the original training data D_{orig} with a fixed amount of QOS data (generated once at the start, equivalent to 0.5x the size of D_{orig}) and the cumulative set of high-confidence pseudo-labeled DOS data collected from all iterations up to i. The model is first trained on this combined synthetic and original dataset, and then fine-tuned on D_{orig}. This process is repeated for 4 iterations beyond the initial training (Iterations 1 through 4).

Figure 7 shows the results for the ST model on AAEC, AbstRCT and CDCP. Generally, performance increases with iterations, particularly in the initial steps, though later iterations exhibit diminishing or slightly negative gains.



(a) AAEC

(b) AbstRCT

(c) CDCP

Figure 7: Iterative self-training results of the ST model on AAEC, AbstRCT and CDCP. The x-axis represents the iteration number.

D Synthetic Data Quality Analysis: Training with QOS and DOS Alone

To specifically evaluate and compare the quality of synthetic data produced by QOS and DOS, we conduct an experiment where AM models are trained exclusively on synthetic data without any original human-annotated training data. This setup aims to isolate the impact of synthetic data quality by removing the influence of original training instances. As in our main experiments, the volume of synthetic data used is twice the size of the original training data for each setting. We show the results of AM models (ST) trained with only data generated by QOS, DOS, or their combination (25%QOS+75%DOS) in Table 6.

QOS consistently outperforms DOS across all datasets and settings, with particularly significant

AM Model	Training Data	Dataset	BS	LR	Epoch
ST	Synthetic Data	AAEC AbstRCT CDCP	8	2e-5	10
	Original Training Data	AAEC AbstRCT CDCP	4 4 4	9.1e-5 8.1e-5 5.6e-5	20 20 20
UniASA	Synthetic Data	AAEC AbstRCT CDCP	2	2e-5	10
	Original Training Data	AAEC AbstRCT CDCP	1 1 1	1e-4 1e-4 2e-4	35 10 40

Table 5: Hyper-parameters of the main experiments. "BS" and "LR" denote batch size and learning rate. For full-data and low-resource settings, and different data synthesis/data augmentation methods, we use the same hyper-parameters as above.

gaps in the low-resource setting. These results confirm that QOS produces synthetic data with higher-quality labels due to its structure-aware paraphrasing approach. The performance difference between QOS and DOS is more pronounced in low-resource scenarios, indicating that label quality becomes increasingly critical when working with limited data. The QOS+DOS combination generally performs better than DOS alone but slightly underperforms QOS in most metrics. These findings demonstrate a clear quality-diversity trade-off between our approaches. QOS effectively preserves label quality, while DOS offers valuable diversity at the cost of some label reliability.

E Results on Another AM Model

We further conduct additional experiments on another AM model: DENIM (Sun et al., 2024). Our choice of ST and UniASA is to ensure our methods are evaluated on models representative of the two dominant paradigms in end-to-end AM (parsing-based and generation-based). Adding DENIM, a generation-based model with a discourse structure-aware prefix, helps demonstrate the broad applicability of our approach.

We followed the experimental setup from the original DENIM paper, evaluating on the AbstRCT dataset. Table 7 presents the results. The findings clearly show that augmenting the training data with our synthetic instances—QOS, DOS, and their combination—yields consistent performance improvements for DENIM in both full-data and low-resource settings.

F Task-Specific Performance Analysis for ACI and ARI

To provide a more granular understanding of where performance gains originate, we conduct a separate analysis of the two primary sub-tasks: Argument Component Identification (ACI) and Argumentative Relation Identification (ARI). We evaluate the ST model on each task independently to determine which benefits most from our synthetic data augmentation. Since span identification is a prerequisite for both tasks, we report $F1_{span}$ for both evaluations, alongside $F1_{aci}$ for the ACI task and $F1_{ari}$ for the ARI task.

The results, presented in Table 8, demonstrate that our synthetic data augmentation benefits both the ACI and ARI tasks. Although the specific gains for ACI and ARI vary across datasets, our methods generally provide similar levels of enhancement for both tasks.

G Results on Additional Datasets

To further assess the generalizability of our findings, we extend our evaluation to two additional AM datasets: MTC and AASD. Given their smaller size, we report results on the full (100%) training data setting, using a 7:1:2 random split for train/validation/test. The results for the ST model, shown in Table 9, demonstrate that our methods—particularly the QOS+DOS combination—continue to provide significant performance gains. These findings across a total of five diverse datasets strengthen the evidence for the broad applicability of our proposed data synthesis strategies.

Setting	Method	AAEC				AbstRCT				CDCP			
		$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg	$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg	$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg
100%	QOS	81.74	73.17	52.50	69.13	70.25	64.07	**36.39**	56.90	**79.96**	**64.31**	**28.91**	**57.73**
	DOS	78.75	68.89	47.24	64.96	66.45	58.78	24.93	50.05	78.56	57.63	19.96	52.05
	QOS+DOS	**81.75**	**73.38**	**52.70**	**69.28**	**71.24**	**64.21**	35.67	**57.04**	79.92	63.49	26.79	56.73
5%	QOS	**68.61**	**51.95**	**22.54**	**47.70**	**62.01**	**55.11**	**22.94**	**46.69**	73.03	**48.46**	**4.87**	42.12
	DOS	63.70	45.41	10.99	40.03	28.30	23.13	4.54	18.66	60.18	35.55	0.90	32.21
	QOS+DOS	64.06	46.83	15.62	42.17	51.33	44.71	19.94	38.66	**75.13**	48.17	4.83	**42.71**

Table 6: Performance comparison of ST models trained exclusively on synthetic data without any original training data.

Setting	Method	$\mathrm{F1}_{span}$	$\mathrm{F1}_{aci}$	$\mathrm{F1}_{ari}$	Avg
5%	Origin	45.07	39.22	17.72	34.00
	QOS	54.90	**49.53**	**21.25**	**41.89**
	DOS	55.37	49.56	19.42	41.45
	QOS+DOS	**55.59**	48.39	20.67	41.55
100%	Origin	76.85	69.17	39.57	61.86
	QOS	77.09	69.90	**40.80**	62.60
	DOS	77.09	70.08	39.63	62.27
	QOS+DOS	**78.30**	**71.15**	40.23	**63.23**

Table 7: Performance of DENIM model on AbstRCT dataset with our synthetic data generation methods.

H Data Synthesis with Open-source LLMs

To demonstrate the general applicability of our data synthesis method, we conduct experiments using two open-source LLMs: Qwen2.5-14B-Instruction and Llama3.1-70B-Instruction. The results for the ST model are presented in Table 10. As can be seen, our method consistently improves performance regardless of the underlying LLM. Overall, Llama3.1-70B-Instruction yields slightly better gains than Qwen2.5-14B-Instruction, though both are outperformed by GPT-4o. This suggests that the benefits of our approach scale with the capability of the LLM used for data synthesis.

I Cross-Dataset Transferability Analysis

We conduct cross-dataset transfer experiments using our DOS method. Specifically, we pre-train the ST model on DOS data synthesized from a source dataset, and then fine-tune and test it on a target dataset.

The results are presented in Table 11. In the vast majority of cases, pre-training on synthetic data—even from a different source dataset—improves final performance. This indicates that the argumentative patterns learned from the synthetic data are transferable to some extent. However, we also observe a few instances where cross-dataset pre-training leads to a minor performance decrease compared to the baseline, likely due to significant differences in dataset characteristics. For example, pre-training on CDCP DOS data and then fine-tuning on AbstRCT in the 5% setting results in a slight drop in performance. Overall, these experiments show promising results for cross-dataset transfer.

J Analysis of the Auto-Annotation Method in DOS

To validate the design choice for the auto-annotation step in our DOS approach, we conduct a comparative experiment. We compare our method, which uses a fine-tuned baseline AM model for annotation (DOS-BL), against an alternative that uses few-shot GPT-4o for the same task (DOS-LLM).

The results, presented in Table 12, consistently show that using a specialized, in-domain model (DOS-BL) for auto-annotation is more effective than using a general-purpose LLM in a few-shot setting (DOS-LLM). This is likely because a model specifically trained on the AM task's complex annotation scheme provides higher-quality pseudo-labels.

K Quantitative Analysis of Semantic Diversity

Diversity is an important aspect emphasized by many datasets and benchmarks, as it is crucial for building robust and generalizable models (Du et al., 2024; Xie et al., 2025; Dong et al., 2025; Li et al., 2025). To quantitatively measure the diversity of the synthetic data, we conduct an analysis of semantic distance. Specifically, for each synthetic instance, we compute its semantic distance to every instance in the original training set. This distance is derived from the cosine similarity of their embeddings, generated by OpenAI's text-embedding-

Setting	Method	AAEC				AbstRCT				CDCP			
		ACI		ARI		ACI		ARI		ACI		ARI	
		$F1_{span}$	$F1_{aci}$	$F1_{span}$	$F1_{ari}$	$F1_{span}$	$F1_{aci}$	$F1_{span}$	$F1_{ari}$	$F1_{span}$	$F1_{aci}$	$F1_{span}$	$F1_{ari}$
5%	Origin	69.79	52.08	69.32	16.71	57.86	51.12	55.47	21.35	76.06	44.60	76.91	3.01
	QOS	72.31	55.11	71.13	**22.90**	62.04	53.94	58.92	24.48	**77.39**	51.37	**77.67**	8.08
	DOS	**72.98**	56.43	70.80	22.54	61.49	54.03	56.90	22.92	77.11	53.74	76.35	7.25
	QOS+DOS	72.83	**56.67**	**71.99**	22.54	**63.22**	**55.29**	**59.26**	**25.44**	76.87	**54.82**	76.49	**9.96**
100%	Origin	84.77	74.81	84.79	54.42	69.15	62.68	69.38	36.08	82.70	**69.17**	82.32	32.48
	QOS	85.01	75.81	85.44	55.94	72.77	66.03	69.03	37.57	82.57	68.42	82.32	32.34
	DOS	85.37	75.33	85.02	55.14	72.51	65.16	68.32	36.10	82.21	68.73	82.06	34.37
	QOS+DOS	**86.00**	**76.89**	**85.72**	**56.37**	**74.26**	**67.02**	**69.47**	**38.12**	**82.84**	69.08	**82.71**	**35.43**

Table 8: Task-specific performance analysis for ACI and ARI on the ST model in the low-resource (5%) and full-data (100%) settings.

Dataset	Method	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg
MTC	Origin	85.85	78.05	43.64	69.18
	QOS	86.21	77.93	45.48	69.87
	DOS	85.90	79.75	48.32	71.32
	QOS+DOS	**88.83**	**80.51**	**50.67**	**73.34**
AASD	Origin	91.76	75.96	63.94	77.22
	QOS	93.36	77.08	64.22	78.22
	DOS	93.27	78.56	67.06	79.63
	QOS+DOS	**94.23**	**78.81**	**69.93**	**80.99**

Table 9: Performance on two additional datasets, MTC and AASD, using the ST model with 100% training data.

ada-002 model. We then average these distances to get a diversity score for that instance. The final diversity score for a method (QOS or DOS) is the average score across all its generated instances. A higher score indicates greater semantic novelty relative to the original training set. We also compute the internal diversity of the original training set for comparison.

As shown in Table 13, the data generated by our DOS method consistently exhibits higher semantic diversity compared to both the data from QOS and the original training set itself. This quantitative analysis further validates the effectiveness of DOS in enhancing data diversity.

L Case Study

This section provides a concrete example comparing synthetic data generated by both QOS and DOS approaches alongside the original reference training instance. As shown in Figure 8, the QOS approach preserves the original argumentative structure while paraphrasing the content, maintaining the same component types and relation types. In contrast, the DOS approach generates text on a completely different topic, with a modified argumentation pattern that introduces new argument structures.

M Error Analysis of Synthetic Data

To provide deeper insight into the characteristics of our generated data, we conduct a manual error analysis. We randomly sample 5 instances generated by GPT-4o for both the QOS and DOS methods and examine the quality of the annotations.

QOS Data The structure-aware paraphrasing approach of QOS proves highly effective. In our analysis of the sampled instances, we observe no significant errors, such as span misalignments or incorrect component type inheritance. This result confirms the high-fidelity nature of the QOS method.

DOS Data For the DOS data, which relies on auto-annotation, our analysis of the 5 samples (containing 87 component spans and 42 relations) reveals the following mispredictions or omissions:

- **Span-related errors:** 10 instances (e.g., incorrect boundaries, identifying non-argumentative text).
- **Component type errors:** 7 instances (e.g., misclassifying a Claim as a Premise).
- **Relation errors:** 6 instances (e.g., incorrect support/attack links).

We consider this error rate acceptable, as our main experimental results (Table 1) consistently show that the diversity introduced by DOS data provides a net positive impact on model performance, despite these imperfections. Below are representative examples of observed errors.

Setting	LLM	Method	AAEC				AbstRCT				CDCP			
			$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg
5%	-	Origin	67.91	50.33	16.91	45.05	57.52	50.76	21.57	43.28	76.54	49.51	4.03	43.36
	Qwen2.5-14B-I	QOS	70.67	53.22	22.56	48.82	59.48	53.10	24.60	45.73	76.49	47.90	5.40	43.26
		DOS	70.51	53.17	22.81	48.83	56.11	49.62	23.74	43.16	76.29	50.34	5.21	43.95
		QOS+DOS	**70.96**	**54.19**	**23.34**	**49.50**	**59.56**	**53.37**	**24.91**	**45.95**	**76.89**	**50.35**	**6.02**	**44.42**
	Llama3.1-70B-I	QOS	69.37	52.56	22.10	48.01	59.72	53.19	24.17	45.69	76.57	49.75	4.91	43.74
		DOS	70.59	54.23	23.49	49.44	60.47	53.78	24.29	46.18	76.35	50.53	5.37	44.08
		QOS+DOS	**70.64**	**54.31**	**23.69**	**49.55**	**61.21**	**54.24**	**25.81**	**47.09**	**76.81**	**51.32**	**6.78**	**44.97**
100%	-	Origin	84.31	75.56	53.49	71.12	70.15	64.57	36.01	56.91	82.01	67.88	30.85	60.25
	Qwen2.5-14B-I	QOS	84.65	76.31	54.75	71.90	72.10	64.16	39.43	58.56	82.79	68.74	32.87	61.47
		DOS	85.30	77.01	55.94	72.75	70.61	65.06	38.91	58.19	83.01	67.81	30.75	60.52
		QOS+DOS	**85.33**	**77.22**	**56.36**	**72.97**	**72.53**	**66.23**	**39.46**	**59.41**	**83.19**	**68.77**	**33.52**	**61.83**
	Llama3.1-70B-I	QOS	85.19	76.44	55.50	72.38	71.34	65.58	37.58	58.17	82.26	**70.11**	33.79	62.05
		DOS	84.57	76.40	54.69	71.89	71.61	65.84	37.69	58.38	82.45	67.33	34.29	61.36
		QOS+DOS	**85.24**	**76.64**	**55.82**	**72.57**	**72.62**	**66.79**	**39.47**	**59.63**	**83.15**	68.82	**35.06**	**62.34**

Table 10: Performance of the ST model with synthetic data generated by two open-source LLMs. Qwen2.5-14B-I and Llama3.1-70B-I are short for Qwen2.5-14B-Instruction and Llama3.1-70B-Instruction.

Setting	Source	Target: AAEC				Target: AbstRCT				Target: CDCP			
		$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg
5%	None	67.91	50.33	16.91	45.05	57.52	50.76	21.57	43.28	76.54	49.51	4.03	43.36
	AAEC	**71.47**	**54.39**	**22.49**	**49.45**	58.90	51.82	25.41	45.38	76.33	46.57	2.96	41.95
	AbstRCT	70.54	52.80	21.58	48.31	**61.81**	**54.83**	**26.93**	**47.86**	76.52	49.89	5.09	43.83
	CDCP	69.74	52.39	22.23	48.12	56.76	49.95	21.55	42.75	**76.98**	**52.12**	**8.06**	**45.72**
100%	None	84.31	75.56	53.49	71.12	70.15	64.57	36.01	56.91	82.01	67.88	30.85	60.25
	AAEC	84.80	**76.11**	**55.44**	**72.12**	72.71	**66.65**	38.26	59.21	82.13	67.38	31.43	60.31
	AbstRCT	84.35	75.65	55.12	71.71	**72.76**	66.28	**38.72**	**59.25**	**82.91**	**69.11**	31.48	**61.17**
	CDCP	**84.94**	75.77	54.68	71.80	71.98	65.01	37.10	58.03	82.40	67.84	**32.47**	60.90

Table 11: Cross-dataset transfer experiment results in both low-resource (5%) and full-data (100%) settings. "None" indicates the baseline without pre-training on synthetic data.

Setting	Method	AAEC				AbstRCT				CDCP			
		$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg	$F1_{span}$	$F1_{aci}$	$F1_{ari}$	Avg
5%	Origin	67.91	50.33	16.91	45.05	57.52	50.76	21.57	43.28	76.54	49.51	4.03	43.36
	DOS-BL	**71.47**	**54.39**	**22.49**	**49.45**	**61.81**	**54.83**	**26.93**	**47.86**	**76.98**	52.12	**8.06**	**45.72**
	DOS-LLM	70.67	52.58	21.43	48.23	57.16	50.44	24.23	43.94	75.63	**52.12**	6.13	44.62
100%	Origin	84.31	75.56	53.49	71.12	70.15	64.57	36.01	56.91	82.01	67.88	30.85	60.25
	DOS-BL	**84.80**	**76.11**	**55.44**	**72.12**	**72.76**	**66.28**	**38.72**	**59.25**	**82.40**	67.84	**32.47**	**60.90**
	DOS-LLM	84.24	75.53	53.91	71.23	70.26	64.49	38.21	57.65	82.33	**68.60**	31.17	60.70

Table 12: Comparison of auto-annotation methods for DOS. DOS-BL uses a fine-tuned baseline AM model, while DOS-LLM uses few-shot GPT-4o. Results show the performance of the ST model when trained on the resulting synthetic data.

Example 1: Span-related Error. This example illustrates a case where a non-argumentative discourse marker is incorrectly identified as a component. Here, <Claim 8> is a discourse marker introducing a viewpoint, not a distinct argumentative component itself. It was erroneously identified as a span.

Argumentative Text with Argument Component Annotations:
. . . On the other hand, <Claim 8> there are concerns that </Claim 8> <Claim 9> school choice might exacerbate educational inequality </Claim 9>. Critics argue that <Premise 10> it could lead to a disparity between well-resourced schools and those with fewer resources </Premise 10>. . . .

Argumentative Relations:
<Premise 10> Support <Claim 9>

Method	AAEC	AbstRCT	CDCP
DOS	**0.2423**	**0.2573**	**0.2438**
QOS	0.2020	0.2035	0.2192
Origin	0.2036	0.1987	0.2173

Table 13: Quantitative analysis of semantic diversity. Scores represent the average semantic distance from the original training set. A higher score indicates greater diversity.

Example 2: Component Type and Relation Errors. This example shows how a component misclassification leads to an invalid relation.

- Component type error: The component <Premise 4> functions as a concluding "Claim" summarizing a benefit, but it is misclassified as a "Premise".
- Relation error: As a consequence of the type error, the annotated "Support" link from <Premise 4> to <Claim 1> is invalid.

Argumentative Text with Argument Component Annotations:
...To begin with, <Claim 1> parents are most aware of their children's needs and aspirations </Claim 1>. <Premise 2> They possess intimate knowledge of their child's learning style, strengths, and weaknesses </Premise 2>. <Premise 3> This ...</Premise 3>. Moreover, <Premise 4> the ability to choose can lead to a more personalized and effective educational experience, enhancing the child's academic and social development </Premise 4>. ...

Argumentative Relations:
<Premise 2> Support <Claim 1> | <Premise 3> Support <Claim 1> | <Premise 4> Support <Claim 1>

N Prompt Details

This section provides the detailed prompts used in our data synthesis approaches:

- Figure 9: Structure-aware paraphrasing of QOS.
- Figure 10: Topic summarization of DOS.
- Figure 11: Topic brainstorming of DOS.
- Figure 12: Text imitation with argumentation pattern diversification of DOS.

Reference Instance from Original Training Data:

Argumentative Text with Argument Component Annotations:
Television is one of the greatest innovations that we use every day. Besides, watching television has some cons and at the same time has very good pros. However, while some people argue that TV has devastated communication among friends and families, I believe it has done the opposite. I think <MajorClaim 0> TV programs are among the popular topics in every day talk as well as a great time for gathering </MajorClaim 0>. In addition <Claim 1> Modern TV has smart system that let you be connected with people </Claim 1>.
First, it's true <Claim 2> watching TV take a lot of your day </Claim 2>, nevertheless <Premise 3> it makes people meet and start a conversation about what they watch at other times like weather forecast and sport programs </Premise 3>. Second, <Claim 4> people like having meals while they are watching TV </Claim 4>. <Claim 5> Families usually tend to watch television either with each other or friends </Claim 5>. For example <Premise 6> my friends and I used to sit together and enjoy watching movies using VHS player at nights </Premise 6>. <Premise 7> It was fun and a little bit similar to the cinema </Premise 7>. Third, <Premise 8> TV can be one of communication ways </Premise 8>. Moreover, <Premise 9> TV can be connected to multimedia devices that have smart operating system like Google TV, WDTV using Skype, tango and other apps </Premise 9>. In addition <Premise 10> Teenagers can play games online with their friends </Premise 10> ; therefore <Claim 11> we can communicate with each other through TV </Claim 11>.
To sum up, <Claim 12> TV may take some of your time away from your family and friends </Claim 12> but <Claim 13> it can also make family get together and help to get in conversation </Claim 13>. As well as <Claim 14> it's one way of recent communications </Claim 14>, based on that I think <MajorClaim 15> TV helps you to get connected with people </MajorClaim 15>.

Argumentative Relations:
<Claim 2> Attack <Premise 3> <Claim 5> Support <Premise 6> <Claim 5> Support <Premise 7>
<Claim 11> Support <Premise 8> <Claim 11> Support <Premise 9> <Claim 11> Support <Premise 10>

Synthetic Instance from QOS:

Argumentative Text with Argument Component Annotations:
Television is undoubtedly one of the most significant innovations we engage with daily. On the one hand, watching television has its drawbacks, yet it also offers substantial benefits. Despite claims that TV has weakened communication among friends and families, I am convinced it has actually strengthened it. I believe <MajorClaim 0> TV shows are among the trending subjects in everyday conversations and serve as an excellent occasion for social gatherings </MajorClaim 0>. Additionally, <Claim 1> Modern television features smart systems enabling connectivity with others </Claim 1>. Initially, while it's correct that <Claim 2> watching TV consumes a significant portion of your day </Claim 2>, it nevertheless <Premise 3> it encourages people to come together and initiate discussions about what they view, such as weather forecasts and sports broadcasts </Premise 3>.
Furthermore, <Claim 4> many enjoy eating while watching TV </Claim 4>. <Claim 5> Families often prefer to watch television with either each other or their friends </Claim 5>. To illustrate, <Premise 6> my friends and I used to gather and relish watching movies on a VHS player during the evenings </Premise 6>. <Premise 7> It was entertaining and somewhat resembled a cinema experience </Premise 7>.
Moreover, <Premise 8> TV can serve as a means of communication </Premise 8>. Besides, <Premise 9> TV can connect to multimedia devices featuring smart operating systems like Google TV and WDTV, utilizing apps such as Skype and Tango </Premise 9>. Also, <Premise 10> Teenagers have the option to play online games with friends </Premise 10> ; thus <Claim 11> we can interact with each other through television </Claim 11>.
In conclusion, <Claim 12> TV might consume time that could be spent with family and friends </Claim 12> but <Claim 13> it also facilitates family gatherings and stimulates conversation </Claim 13>. Furthermore, <Claim 14> it is a modern form of communication </Claim 14>, and based on this, I assert <MajorClaim 15> TV fosters connections among people </MajorClaim 15>.

Argumentative Relations:
<Claim 2> Attack <Premise 3> <Claim 5> Support <Premise 6> <Claim 5> Support <Premise 7>
<Claim 11> Support <Premise 8> <Claim 11> Support <Premise 9> <Claim 11> Support <Premise 10>

Synthetic Instance from DOS:

Argumentative Text with Argument Component Annotations:
The relationship between democracy and capitalism has long been a subject of heated debate. While some argue that these systems are inherently at odds, I believed <MajorClaim 0> emocracy and capitalism can coexist and even complement each other </MajorClaim 0>. In fact, <Claim 1> capitalism creates the economic prosperity that allows democratic ideals to flourish </Claim 1>.
First, <Claim 2> democracy thrives on individual freedoms and choice </Claim 2>. <Premise 3> Capitalism, as an economic system, also emphasizes freedom by allowing individuals to make choices about what to buy, sell, or invest in </Premise 3>.This alignment of values suggests that <Claim 4> both systems can support one another </Claim 4>.
Second, <Claim 5> capitalism can provide the resources necessary for a democratic society to function effectively </Claim 5>. <Premise 6> Economic growth driven by capitalism enables governments to fund public services and infrastructure </Premise 6>. At the same time, <Premise 7> democracy provides the political framework to regulate capitalism and ensure fairness </Premise 7>. For example, <Premise 8> laws against monopolies and exploitation are often enacted in democratic societies </Premise 8>.
Third,critics argue that <Claim 9> capitalism promotes inequality, which can undermine democracy </Claim 9>. However, <Premise 10> democratic institutions can mitigate these inequalities through policies like progressive taxation and social welfare programs </Premise 10>. Furthermore, <Premise 11> capitalism encourages innovation and competition, which can empower individuals to improve their circumstances </Premise 11>. <Premise 12> This empowerment aligns with democratic ideals of equal opportunity </Premise 12>.
To sum up,while <Claim 13> capitalism may pose challenges to democracy </Claim 13>, <Claim 14> it also offers mechanisms for economic freedom and growth that reinforce democratic principles </Claim 14>. Additionally, <Premise 15> democracy has the tools to regulate capitalism and ensure it serves the public good </Premise 15>. Therefore, <MajorClaim 16> democracy and capitalism are compatible systems that, when balanced, can mutually reinforce one another </MajorClaim 16>.

Argumentative Relations:
<Claim 4> Support <Premise 3> <Claim 5> Support <Premise 6> <Claim 5> Support <Premise 7>
<Premise 7> Support <Premise 8> <Claim 9> Attack <Premise 10> <Claim 9> Attack <Premise 11>
<Claim 9> Attack <Premise 12>

Figure 8: Examples of synthetic data from QOS and DOS, taken from the AAEC dataset.

Your task is to paraphrase the provided argumentative text.
The text is given in a JSON format, which consists of a main text (context) with placeholders ([AC1], [AC2], etc.), and the argument components (argument_component_info) that will be inserted at these placeholders.
The types of the argument components are defined as follows:

- "MajorClaim": The central standpoint of the author on the topic.
- "Claim": A statement that supports or attacks the author's central standpoint (MajorClaim).
- "Premise": A statement serving as a reason, justification, or evidence to support or attack either a Claim or another Premise.

Please adhere to the following rules:

- Preserve the original meaning of both the context and argument components
- Enhance expression diversity and language variety
- After paraphrasing, ensure smooth and natural flow when components are reintegrated into the context
- Maintain each argument component's designated type

Below is the provided argumentative text, please return the answer in a similar JSON format.

```
{
    "context": "··· From this point of view, I firmly believe that [AC1] .\nFirst of all, [AC2] . [AC3] . ··· ",
    "argument_component_info": {
    "[AC1]": {
        "type": "MajorClaim",
        "content": "we should attach more importance to cooperation during primary education"
        },
    },
    "[AC2]": {
        "type": "Claim",
        "content": "through cooperation, children can learn about interpersonal skills which are significant in the future life of all students"
    },
    ···
}
```

Figure 9: Example of the prompt used for Quality-Oriented Synthesis (QOS). It instructs the LLM to perform structure-aware paraphrasing. The input text is provided in a JSON format with context (containing placeholders like '[AC1]') and argument component information (content and type). The LLM is tasked to paraphrase both while preserving original meaning, component types, and ensuring natural reintegration.

Your task is to summarize the topic of the following argumentative text:

Any collector who uses a robocall, without first having a live person call to verify that the phone number is correct, is lazy and irresponsible. Aside from being a major nuisance, · · ·

Rules:
- Please refer to the following examples and provide a topic of similar length.
- Return the result in a similar JSON format.

Example 1:

Input:

Too many collectors call and never report the "mini-Miranda warning". They call all times of the day and night, and multiple times of the day. If they don't get you because your have called ID, they will · · ·

Output:

```
{
    "topic": "Debt collectors often harass consumers and lack proper documentation."
}
```

Example 2:

Input:

I think the "unless" part of the rule about contacting a person more than once should be scrapped. They should not be allowed to contact anyone (other than the debtor him/herself) more than once. If the person · · ·

Output:

```
{
    "topic": "Collectors should not repeatedly contact third parties about debts."
}
```

Figure 10: Example of the prompt used for summarizing the topic of an argumentative text. This is employed in the Diversity-Oriented Synthesis (DOS) approach when explicit topics are not readily available in the original dataset (e.g., CDCP, AbstRCT).

Referring to the argumentative text writing topics below, please brainstorm and write 16 diverse topics.

Please adhere to the following rules:

- Return the results in a similar JSON format.
- Ensure the generated topics cover diverse aspects within the same domain.

```
{
    "topics": [
        "Should students be taught to compete or to cooperate?",
        "International tourism is now more common than ever before",
        "Will newspapers become a thing of the past?",
        "Government budget focus, young children or university?",
        "Roommates quality and their importance",
        "Should governments spend more money on improving roads and highways",
        "Physical exercise",
        "Advance in transportation and communication like the airplane and the phone",
    ]
}
```

Figure 11: Example of the prompt used for topic brainstorming in the Diversity-Oriented Synthesis (DOS) approach. The LLM is given a list of existing topics as examples and instructed to generate a specified number of new, diverse topics within the same thematic domain.

Your task is to imitate the provided reference text to write a new argumentative text.
The topic of the new text should be:

"Should social life be built through work or outside of work?"

Please adhere to the following rules:

- Ensure the number of paragraphs is the same as the reference text, and the text length is similar.
- Make adjustments in aspects such as the organization of the argumentative structure, logical reasoning patterns, or the selection of evidence types.
- In the ‘argumentation_pattern‘, the sequence of argument components describes the logic flow of the entire text. The types and definitions of these components are as follows:
    - "MajorClaim": The central standpoint of the author on the topic.
    - "Claim": A statement that supports or attacks the author’s central standpoint (MajorClaim).
    - "Premise": A statement serving as a reason, justification, or evidence to support or attack either a Claim or another Premise.
- First, adjust the provided ‘argumentation_pattern‘ to create a *new* ‘argumentation_pattern‘. Then, generate a new argumentative text following this new ‘argumentation_pattern‘. To adjust the ‘argumentation_pattern‘, you can choose to perform one or more of the following operations:
    - Add new argument components.
    - Remove existing argument components.
    - Adjust the order of the argument components.
    - Adjust the type of the argument components.

Below is the reference text, please return the answer in a similar JSON format.

{
    "topic": "Should students be taught to compete or to cooperate?",
    "argumentation_pattern": {
        "paragraph_1": "MajorClaim",
        "paragraph_2": "Claim $\rightarrow$ Premise $\rightarrow$ Premise $\rightarrow$ Premise",
        . . .
    },
    "argumentative_text": "It is always said that competition can effectively promote the development of economy. In order to survive in the competition, companies continue to improve their products and service, and as a result, the whole society prospers. However, when we discuss the issue of competition or cooperation, what we are concerned about is not the whole society, but the development of an individual’s whole life. From this point of view, I firmly believe that <MajorClaim> we should attach more importance to cooperation during primary education </MajorClaim> .\nFirst of all, <Claim> through cooperation, children can learn about interpersonal skills which are significant in the future life of all students </Claim> . <Premise> What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others </Premise> . <Premise> During the process of cooperation, children can learn about how to listen to opinions of others, how to communicate with others, how to think comprehensively, and even how to compromise with other team members when conflicts occurred </Premise> . <Premise> All of these skills help them to get on well with other people and will benefit them for the whole life </Premise> . · · · "
}

Figure 12: Example of the prompt used for text imitation with argumentation pattern diversification in the Diversity-Oriented Synthesis (DOS) approach. The LLM is provided with a new topic and a reference argumentative text (including its original topic, argumentation pattern, and full text with inline component tags). It is instructed to first modify the reference argumentation pattern and then generate a new argumentative text on the new topic, following the diversified pattern.