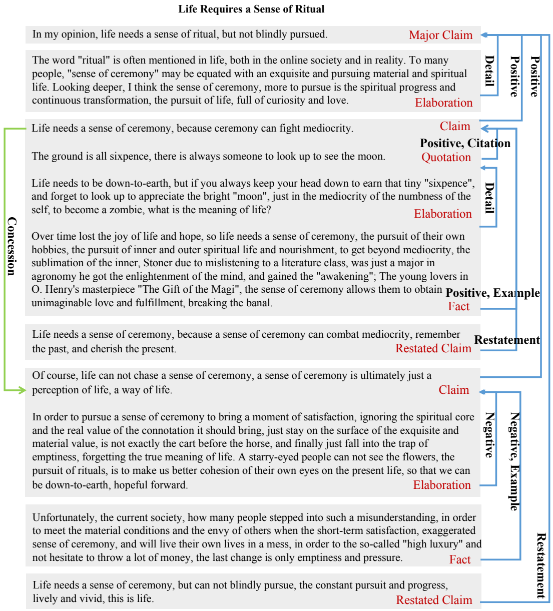
Towards Comprehensive Argument Analysis in Education: Dataset, Tasks, and Method

Yupei Ren[1,2,3], Xinyi Zhou[4], Ning Zhang[5], Shangqing Zhao[3],
Man Lan[1,2,3*], Xiaopeng Bai[1,2,4]
[1]Lab of Artificial Intelligence for Education, East China Normal University
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University
[3]School of Computer Science and Technology, East China Normal University
[4]Department of Chinese Language and Literature, East China Normal University
[5]College of Education, Zhejiang University
ypren@stu.ecnu.edu.cn, mlan@cs.ecnu.edu.cn

Abstract

Argument mining has garnered increasing attention over the years, with the recent advancement of Large Language Models (LLMs) further propelling this trend. However, current argument relations remain relatively simplistic and foundational, struggling to capture the full scope of argument information. To address this limitation, we propose a systematic framework comprising 14 fine-grained relation types from the perspectives of vertical argument relations and horizontal discourse relations, thereby capturing the intricate interplay between argument components for a thorough understanding of argument structure. On this basis, we conducted extensive experiments on three tasks: argument component prediction, relation prediction, and automated essay grading. Additionally, we explored the impact of writing quality on argument component prediction and relation prediction, as well as the connections between discourse relations and argumentative features. The findings highlight the importance of fine-grained argumentative annotations for argumentative writing assessment and encourage multi-dimensional argument analysis.[1]



Life Requires a Sense of Ritual

In my opinion, life needs a sense of ritual, but not blindly pursued.

The word "ritual" is often mentioned in life, both in the online society and in reality. To many people, "sense of ceremony" may be equated with an exquisite and pursuing material and spiritual life. Looking deeper, I think the sense of ceremony, more to pursue is the spiritual progress and continuous transformation, the pursuit of life, full of curiosity and love.

Life needs a sense of ceremony, because ceremony can fight mediocrity.

The ground is all sixpence, there is always someone to look up to see the moon.

Life needs to be down-to-earth, but if you always keep your head down to earn that tiny "sixpence", and forget to look up to appreciate the bright "moon", just in the mediocrity of the numbness of the self, to become a zombie, what is the meaning of life?

Over time lost the joy of life and hope, so life needs a sense of ceremony, the pursuit of their own hobbies, the pursuit of inner and outer spiritual life and nourishment, to get beyond mediocrity, the sublimation of the inner, Stoner due to mislistening to a literature class, was just a major in agronomy he got the enlightenment of the mind, and gained the "awakening"; The young lovers in O. Henry's masterpiece "The Gift of the Magi", the sense of ceremony allows them to obtain unimaginable love and fulfillment, breaking the banal.

Life needs a sense of ceremony, because a sense of ceremony can combat mediocrity, remember the past, and cherish the present.

Of course, life can not chase a sense of ceremony, a sense of ceremony is ultimately just a perception of life, a way of life.

In order to pursue a sense of ceremony to bring a moment of satisfaction, ignoring the spiritual core and the real value of the connotation it should bring, just stay on the surface of the exquisite and material value, is not exactly the cart before the horse, and finally just fall into the trap of emptiness, forgetting the true meaning of life. A starry-eyed people can not see the flowers, the pursuit of rituals, is to make us better cohesion of their own eyes on the present life, so that we can be down-to-earth, hopeful forward.

Unfortunately, the current society, how many people stepped into such a misunderstanding, in order to meet the material conditions and the envy of others when the short-term satisfaction, exaggerated sense of ceremony, and will live their own lives in a mess, in order to the so-called "high luxury" and not hesitate to throw a lot of money, the last change is only emptiness and pressure.

Life needs a sense of ceremony, but can not blindly pursue, the constant pursuit and progress, lively and vivid, this is life.

Figure 1: An Annotation Example (Excerpt). The red font indicates argument component types, the blue arrows on the right signify vertical argument relations, and the green arrow on the left represent horizontal logical relations. The content above the arrows corresponds to the respective relation types.

1 Introduction

Argument Mining (AM) aims to automatically extract structured argumentation information from unstructured texts, encompassing the analysis of argument units, comprehending their roles and interactions within a document, and ultimately forming a cohesive argumentation (Lippi and Torroni, 2016). The automatic identification of argument structure holds significant promise, providing valuable support for various downstream Natural Language Processing (NLP) tasks such as quality evaluation (Stahl et al., 2024) and text generation (Lin et al., 2023; Chen et al., 2023).

Existing research on argument mining has proposed various argument annotation schemes and tasks, mainly focusing on two aspects: (a) *Argument Component Prediction*, and (b) *Argument Relation Prediction*. Each element embodies a unique aspect of argument mining. Regarding argument component prediction, Guo et al. (2023) emphasize that a comprehensive understanding of argumentative texts requires knowledge of the viewpoints (i.e., claims) presented in the text, the validity of those viewpoints (i.e., supporting evidence), and the source of the evidence (i.e., evidence types). In

[*] Corresponding author.

[1]Our code and dataset are released at https://github.com/cubenlp/CEAMCv2.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14215–14231
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

terms of argument relation prediction, Mochales and Moens (2011) argue that the core of argument analysis lies in comprehending the content of the argument chains, analyzing linguistic structures, and determining the relations between argument units to reveal the argument structure of the text.

Despite significant progress in the field of argument mining, the current argument relations are still relatively simple and basic, making it difficult to capture complete argument information, especially to meet the representation needs of complex argument structures in real scenarios. For example, most argument mining studies (Cheng et al., 2022; Schaller et al., 2024) only categorizes argument relations into *support* and *attack* based on stances, lacking the characterization of critical information such as argument strategies and patterns, which are essential for a thorough understanding and evaluation of the overall structure and quality of arguments. Furthermore, existing quality assessment of argumentative essays primarily focuses on scoring annotations for argumentative attributes such as strength, relevance (Wambsganss and Niklaus, 2022), content, and style (Schaller et al., 2024). While these innovative annotation methods enhance the granularity of assessment, they overlook the intrinsic value of argument components and relations as key argumentative features, hindering the effective integration of argument component prediction and relation prediction with quality assessment.

To address the shortcomings of existing research, we propose an innovative relation annotation scheme to characterize the argument strategies and patterns within the **C**hinese **E**ssay **A**rgument **M**ining **C**orpus (**CEAMC**) (Ren et al., 2024). As shown in Figure 1, each argumentative essay undergoes meticulous annotation. We argue that these annotations address the key limitations in prior work: **First**, it overcomes the issue of simplified argument relations prevalent in previous studies. By deeply integrating argument relations with discourse relations, it introduces 14 fine-grained relation types from both vertical and horizontal dimensions, comprehensively depicting the complex interactions between argument components and providing a deeper understanding of argument structures. **Second**, it breaks away from the isolationist approach previous studies. With the integration of argument component, relations, and essay grading, it provides a more comprehensive understanding of argument analysis. **Lastly**, the detailed annotations adeptly capture the subtle nuances of real-world argumentative texts, providing a more reliable basis for argument evaluation and instruction.

Our contributions can be summarized as follows:

- We have revised and enhanced a comprehensive multi-task dataset for argument analysis, enhancing understanding of Chinese high school student argumentative essays.

- We provide comprehensive benchmarks for each task, systematically evaluate the performance of existing methods, and offer reference points for future research.

- Through insightful experiments, we illustrate the impact of writing quality on argument component prediction and relation prediction, and explore the connections between discourse relations and argumentative features, encouraging multi-dimensional argument analysis.

2 Related Work

2.1 Argument Mining

Most argument mining studies (Fergadis et al., 2021; Wambsganss and Niklaus, 2022; Jundi et al., 2023) have focused on identifying the basic argument components and relations, namely the three components of *major claim*, *claim* and *premise*, as well as the two relations of *support* and *attack*.

Existing studies have delved into argument components, including refining the categories from the perspective of sentence functions (Song et al., 2021; Kennard et al., 2022) and further categorizing them based on evidence attributes (Niculae et al., 2017; Guo et al., 2023). Recently, research on argumentative essays in German (Schaller et al., 2024; Stahl et al., 2024) and Chinese (Ren et al., 2024) schools has advanced the study of argumentation education through multi-level granularity annotation. While these efforts have facilitated an understanding of arguments, they lack a thorough exploration of argument relations.

Regarding argument relations, several studies have refined additional relations based on discourse relations, such as *detail*, *sequence* (Kirschner et al., 2015), *by-means*, *info-required* and *info-optional* (Accuosto et al., 2021), which hold significant value in scientific literature. Similarly, Jo et al. (2021) adopted *causal* and *normative* relations as supplements in debate analysis. Recently, Liu et al. (2024) defined *affiliation*, *co-occurrence* and *co-relevance* relations to characterize the argument structure within documents from online financial forums. Although the general argumentation schemes proposed by Walton et al. (2008) (e.g., *Argument from Example* and *Hypothesis*) provide a theoretical framework for computational argumentation, practical applications remain predominantly focused on basic argumentation methods, lacking in-depth understanding and systematic analysis of argument structures. Moreover, most of these findings are concentrated in out-of-education domains and mainly in English and German, limiting the further development and application of argument mining research.

2.2 Discourse Relation Recognition

Discourse Relation Recognition (DRR) aims at detecting semantic relations between text units, thereby modeling the logical structure of discourse. Existing research on discourse relations is mainly based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Discourse Treebank (PDTB) (Prasad et al., 2008). On this basis, considering the nuances of Chinese discourse, Wu et al. (2023) formulated a framework consisting of four tiers and thirteen labels. This framework encompasses a wide range logical semantic types in Chinese discourse and promotes the development of discourse relation research.

It is noteworthy that discourse structure is closely associated with argument structure, and the research of discourse relations plays a critical role in guiding argument mining. Cabrio et al. (2013) and Stab and Gurevych (2017) have both emphasized the significance of discourse relations in argument mining research and advocate for their integration to foster insightful investigation. Existing frameworks, such as the Inference Anchoring Theory (IAT), have explicitly modeled both argument structure and dialogue structure (Ruiz-Dolz et al., 2024; Górska et al., 2024; Janier et al., 2014). This theory primarily involves two core types of relations: propositional relations and illocutionary relations. Propositional relations include inference, conflict, and rephrase; while illocutionary relations existing between locutions uttered in the dialogue and the argumentative propositions associated with them, reflecting speaker intentions (e.g., asserting, agreeing, disagreeing, questioning, etc.). The establishment of IAT provides a significant foundation for the organic integration of dialogical discourse analysis and argumentation research. However, it should be noted that there are significant differences between dialogical intentions and writing logic, particularly considering that the main purpose of argumentative writing is to enhance persuasiveness, where confrontational or conflict relations rarely occur (Wambsganss and Niklaus, 2022; Stahl et al., 2024). Furthermore, the discourse relations between arguments contribute to a comprehensive understanding of argument structure and quality from a holistic perspective, yet this aspect remains largely underexplored in current research.

3 Corpus Construction

In this section, we briefly describe the corpus we use, i.e. the CEAMC corpus (Ren et al., 2024). In addition, we present our annotation scheme, the procedure, and results.

Writing Score Level	Counts	Writing Quality Group	% of Total
I (63 - 70)	21	High-quality	34.51%
II (52 - 62)	57		
III (39 - 51)	146	Low-quality	65.49%
IV (21 - 38)	2		
V (0 - 20)	0		
Total	226	-	100.00%

Table 1: Distribution of argumentative essay scores in the research data.

3.1 Source Data

The CEAMC corpus (Ren et al., 2024) comprises 226 argumentative essays from high school examination context, annotated with 4 coarse-grained and 10 fine-grained sentence-level argument components (i.e., *Assertion*: *major claim*, *claim* and *restated claim*; *Evidence*: *fact*, *anecdote*, *quotation*, *proverb*, and *axiom*; *Elaboration*; and *Others*). Essays from authentic educational settings encapsulate rich argumentative information, offering a unique perspective for insightful exploration of argument strategies and structures. Based on this, we conduct extensive relation annotations.

Table 1 details the score distribution of the essays, classified according to the the Chinese National College Entrance Examination (Gaokao) scoring criteria. Given the limited sample size and imbalanced category distribution, the Category I and II essays are combined into a high-quality group, and the Category III, IV, and V into a low-quality group, enabling subsequent in-depth comparative analysis of argumentation features across different quality tiers.

3.2 Annotation Scheme

In persuasive writing, different argument components and targets lead to various argument strategies. To provide a comprehensive representation and profound analysis of argumentative essays, we innovatively annotate the relations within argumentative essays from both vertical and horizontal dimensions, based on education practice and by integrating both argument and discourse relations.

3.2.1 Vertical Dimension

The vertical dimension focus on the relations between different types of argument components, aiming to reveal the internal logic and reasoning chains of arguments. We have defined ten types of argument relations from three aspects to comprehensively characterize the collaborative interactions between argument components.

Stance-Based Argument Relations Most argument mining research categorizes argument relations into support and attack based on stance, but occurrences of attack relations are quite rare (Stab and Gurevych, 2014, 2017; Stahl et al., 2024). Additionally, no attack relations were observed in the analysis of online comments by Park and Cardie (2018), nor in the research on business proposals by German university students conducted by Wambsganss and Niklaus (2022). Furthermore, Song et al. (2021) did not annotate the relations in the Chinese online argumentative essays and subtly implied that there was a support relation between the evidence and the claim. Taking into account the fact that when writing persuasive essays, students aim to argue for their major claims in the most persuasive manner, typically without overemphasizing attack relations between argument components, but more commonly from the opposite side to strengthen their reasoning. Based on these observations, we propose three stance-based argument relations, namely **Positive**, **Negative**, and **Comparative** argumentation, to understand students' argumentative essays.

Evidence-Based Argument Relations Different types of evidence lead to different modes of argumentation. In conjunction with educational practice, evidence-based argument relations include **Example** and **Citation** argumentation.

Discourse-Based Argument Relations To comprehensively depict the argumentation process of students, we integrate discourse analysis theory to further expand and refine the argument relations. Drawing on the framework of RST (Mann and Thompson, 1988), we introduce three new categories of relations, namely **Background**, **Detail**, and **Restatement**, to enhance the understanding of argument structure. Based on Walton et al. (2008), we introduce **Hypothetical Argumentation**, which plays a significant role in Chinese argumentation. Moreover, considering the importance of metaphoric rhetoric in argumentative activities (Pilgram and van Poppel, 2021), we further define **Metaphorical Argumentation**.

3.2.2 Horizontal Dimension

The horizontal dimension focuses on the relations between argument components of the same type, aiming to analyze the interconnections between elements at the same level (such as the relations between claims and how they collectively support the major claim), thus facilitating a comprehensive understanding of argument structure. Grounded in Chinese argumentation teaching, we draw on the research by Wu et al. (2023) and utilize four discourse relations, namely **Coherence**, **Progression**, **Contrast**, and **Concession**, to annotate the logical transitions between arguments of the same type.

For a detailed overview of relation types and samples, please refer to Appendix A.1.

3.3 Annotation Process

Annotation team consists of three undergraduates, three postgraduates specializing in linguistics and education, and two experts with extensive experience in Chinese teaching. Before the formal annotation work, the team underwent a series of training sessions and pre-annotation exercises to better familiarize and master the task requirements. Building on this, we discussed their understanding of the guidelines and variations in annotations, making appropriate adjustments to the guide. Each essay was independently annotated by two annotators, with domain experts responsible for coordinating and resolving any disagreements between them.

It is noteworthy that, before undertaking the relation annotation task, we asked two experts to clearly demarcate the boundaries of argument units based on the results of sentence-level argument component annotations. This step was particularly crucial because we observed that in the Chinese context, there are often multiple consecutive sentences of the same type discussing the same content, which poses challenges to annotating relations between arguments. On this basis, we annotated the relations from both vertical and horizontal dimensions, encompassing 226 argumentative essays with a total of 4,837 relations.

3.4 Inter-Annotator Agreements

We followed Cheng et al. (2022) and Liu et al. (2024), employing Cohen's kappa to measure Inter-Annotator Agreement (IAA). A total of 3,458 argument units were derived from 4,726 sentences, with an IAA score of 0.95 for the annotation of argument unit boundaries, which indicates a high degree of consistency, providing reliable outcomes for subsequent relation annotation. Based on this, a total of 4,837 relations were annotated with an IAA score of 0.68, which is a reasonable and relatively high agreement considering the diversity and complexity of relation annotations. The confusion matrix for relation type annotations is provided in Appendix A.2.

3.5 Data Statistics

The final corpus consists of 226 Chinese argumentative essays, comprising 3,458 argument units, 3,923 argument pairs, and 4,837 relations (multiple relations may exist between each argument pair). As shown in Table 2, there are significant differences in the distribution of various argument relations and discourse relations, indicating that students exhibit diversity and complexity in mastering argument structures and relations in argumentative writing.

Dimension	**Aspect**	**Label**	**Counts**	**% of Total**
Vertical (4,102)	Stance-Based	Positive	1,599	33.04%
		Negative	396	8.19%
		Comparative	27	0.56%
	Evidence-Based	Example	661	13.67%
		Citation	216	4.47%
	Discourse-Based	Metaphorical	31	0.64%
		Hypothetical	6	0.12%
		Restatement	203	4.20%
		Detail	698	14.43%
		Background	265	5.48%
Horizontal (735)	-	Coherence	277	5.73%
		Progression	305	6.31%
		Contrast	46	0.95%
		Concession	107	2.21%
Total	-	-	4,837	100.00%

Table 2: Distribution of relations.

4 Experiments

4.1 Tasks

Our annotated dataset serves as the foundation for three core tasks, each delving into distinct facets of argument analysis:

	Train	**Dev.**	**Test**	**Total**
Argument Component Prediction				
# Sentences	3,805	451	470	4,726
# Arguments	2,767	346	345	3,458
Relation Prediction				
# Positives	3,133	398	392	3,923
# Relations	3,866	484	487	4,837
# Negatives	16,410	1,992	2,039	20,441
Automated Essay Grading				
# Essays	180	23	23	226

Table 3: Data split statistics for benchmark testing.

Argument Component Prediction. This task aims to detect and classify all potential argument components. We formulate it as a sentence-level classification task, utilizing IOB tagging to represent structural span information.

Relation Prediction. This task aims to detect and classify all relations between argument components. We frame it as argument-pair classification task: given a pair of argument components, predict the types of relations between them, noting the multi-label nature due to multiple relation types.

Automated Essay Grading. This task aims to evaluate the overall quality of students' argumentative essays. We frame it as a four-classification task, with detailed writing quality levels provided in Table 1.

To address above tasks, we split our data as summarized in Table 3. Across all tasks, a total of 226 labeled argumentative essays are split in an approximate 8:1:1 ratio. It should be noted that in the second task of relation prediction, our data statistics indicates that setting an argument distance of 15 covers almost 99% of the positive argument pairs. Therefore, we construct negative samples based on a forward-backward distance of 15 and argument component types, that is, argument pairs with no existing relations.

4.2 Baselines and Metrics

We experiment on two well-established pretrained language models (PLMs): *BERT* (Devlin et al., 2019) and *RoBERTa* (Liu et al., 2019). Given the recent unparalleled achievements of LLMs in various NLP tasks, we also employ the LoRA technique (Hu et al., 2021) to conduct supervised fine-tuning (SFT) on three open-source Chinese LLMs, *Qwen* (Yang et al., 2025), *DeepSeek* (Guo et al., 2025), and *ChatGLM* (GLM et al., 2024), to evaluate their performance on each of our argument analysis tasks. Additionally, we assess the performance of OpenAI's ChatGPT[2], specifically *GPT-4-turbo*, under zero-shot and few-shot prompting conditions, to serve as a reference.

Argument Component Prediction: We fine-tune various PLMs and LLMs on the training dataset, leveraging their powerful language modeling capabilities. We evaluate the performance of models using Precision (P), Recall (R), Micro-F_1, and Macro-F_1. More precisely, the true positive for calculation is defined as the number of predicted argument components that exactly match a gold standard argument component, i.e., their boundaries and category labels are identical.

Relation Prediction: We evaluate the performance of models using Micro-F_1, Macro-F_1, and Pos.-F_1. Precisely, Pos.-F_1 aims to measure the models' ability to identify positive samples with relations, focusing solely on whether a relation exists between arguments, without distinguishing the specific types of relations. It is noteworthy that the argument pairs labeled with no-relation far exceeds other types of relations (as shown in Table 3). To overcome this challenge, we adopt negative sampling techniques (Mikolov et al., 2013). During the training process, we randomly select a certain amount of unrelated arguments for each argument as negative samples. These negative samples, along with all other arguments, form a new training dataset.

Automated Essay Grading: In addition to using the original essay as input, we incorporate argument components and relations to explore the impact of fine-grained argumentative information on essay grading (see Appendix B.1 for complete prompts). We evaluate model performance using P, R, F_1, Accuracy (Acc), and Quadratic weighted Kappa (QWK) (Vanbelle, 2016).

4.3 Implementation Details

For PLMs, we implement BERT-Base-Chinese and Chinese-RoBERTa-wwm-ext, using an AdamW optimizer with a learning rate of $2e^{-5}$ to update the model parameters, and set the batch size to 8. For the open-source LLMs, we use Qwen3-8B-Base, DeepSeek-R1-Distill-Qwen-7B, and ChatGLM-4-9B-Base, employing LoRA throughout all training sessions with a LoRA rank of 8 and a dropout rate of 0.1. Training configurations include the learning rate of $5e^{-5}$ and the batch size of 2. All our experiments are conducted on a single NVIDIA RTX 3090 GPU. All other parameters are initialized with the default values in PyTorch Lightning[3], and our models are entirely implemented by Transformers[4]. Each experiment was run three times with averaged results reported.

Model	$P(\%)$	$R(\%)$	Micro-F_1	Macro-F_1
BERT$_{ft}$	39.67	45.80	42.51	25.85
RoBERTa$_{ft}$	43.93	49.18	46.41	27.58
DeepSeek$_{ft}$	53.49	51.30	52.37	42.17
Qwen$_{ft}$	56.46	57.39	56.92	45.67
ChatGLM$_{ft}$	**58.43**	**59.61**	**59.02**	**50.78**
GPT-4$_{0-shot}$	29.50	34.20	31.68	20.55
GPT-4$_{1-shot}$	27.01	27.44	27.19	19.80
GPT-4$_{3-shot}$	32.66	33.04	32.85	22.56

Table 4: Results for Argument Component Prediction.

4.4 Results and Analysis

4.4.1 Argument Component Prediction

Table 4 showcases the performance of various models on the *Argument Component Prediction* task. In the SFT setting, LLMs outperform PLMs on all metrics, demonstrating the exceptional ability of LLMs in identifying and predicting argument components. This superiority is attributed to the extensive knowledge and learning capabilities of LLMs, confirming the scaling laws that larger models tend to yield better performance (Kaplan et al., 2020). Furthermore, the larger parameter-sized ChatGLM (9B) surpasses Qwen (8B) and DeepSeek (7B) in achieving the best performance, which further corroborates this principle.

In contrast, while GPT-4 performs commendably under 0-shot and few-shot conditions, it significantly lags behind the models under SFT, highlighting the advantages of SFT and the importance of data annotation. Moreover, adding prompt examples does not significantly enhance GPT-4's deep understanding of the task, with its performance visibly diminishing in the 1-shot setting while showing only a slight improvement in the 3-shot setting. This seems to confirm the sensitivity and instability of LLMs to prompt samples, suggesting that procuring high-quality samples to improve the performance of LLMs warrants further investigation.

[2] https://openai.com/blog/chatgpt

[3] https://github.com/Lightning-AI/lightning

[4] https://github.com/huggingface/transformers

Model	Micro-F_1	Macro-F_1	Pos.-F_1
BERT$_{ft}$	**73.72**	14.41	19.89
RoBERTa$_{ft}$	69.39	15.97	28.41
DeepSeek$_{ft}$	65.14	22.00	37.17
Qwen$_{ft}$	65.32	**32.11**	40.04
ChatGLM$_{ft}$	67.11	31.52	**42.07**
GPT-4$_{0-shot}$	2.64	4.65	27.82
GPT-4$_{1-shot}$	10.61	5.72	28.14
GPT-4$_{3-shot}$	4.97	4.89	27.94

Table 5: Results for Relation Prediction task.

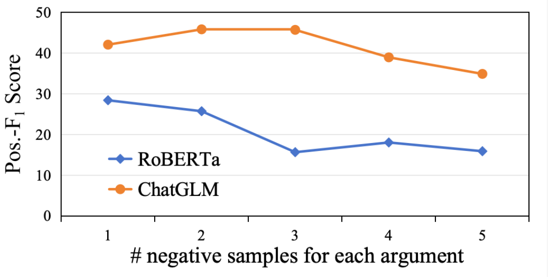


Figure 2: Effect of negative sampling for *Relation Prediction* task with RoBERTa and ChatGLM models.

4.4.2 Relation Prediction

Table 5 displays the performance of various models on the *Relation Prediction* task, where 1 negative sample is randomly selected for each argument component. Evidently, LLMs underperform PLMs in Micro-F_1 but significantly surpass them in Macro-F_1 and Pos.-F_1. This suggests that LLMs possess strong relational reasoning capabilities, particularly in identifying positive samples with existing relations and handling imbalanced data. Notably, GPT-4 achieves remarkably low scores in both Micro-F_1 and Macro-F_1 across all settings, indicating significant challenges in relation prediction. Analysis of the outputs reveals that GPT-4 tends to misclassify negative samples as positive, despite negative (no-relation) samples vastly outnumbering positive ones. This may stem from GPT-4's extensive understanding of relations, which exceeds the scope of our defined argument pair relations, leading it to classify most samples as positive. These findings underscore the importance of domain-specific fine-tuning for LLMs using annotated data. (For fine-grained results of the relation prediction task, see Appendix B.2.)

As depicted in Figure 2, we also compare the performance of the RoBERTa and ChatGLM models with varying numbers of negative samples per argument. Interestingly, ChatGLM peaks in performance with 3 negative samples, while RoBERTa shows the worst performance at this sampling size, reflecting significant differences between LLM and PLM. This may be due to the LLMs' extensive pre-training data and larger parameter size, which allow for better generalization and learning from more diverse data. Conversely, RoBERTa, as a smaller PLM, potentially suffers from an insufficient capacity for abstraction when processing larger numbers of negative samples, particularly in the context of complex relation judgments, thus exhibiting a marked decrease in performance as the number of negative samples rises. Notably, both models exhibit poor performance when the proportion of negative samples is excessively high. These insights highlight the importance of considering model size and learning capabilities when designing sampling strategies.

4.4.3 Automated Essay Grading

Table 6 showcases the performance of various models on the *Automated Essay Grading* task. It is noteworthy that the models achieved promising results in this four-classification task, demonstrating the strengths of current methods in assessing students' overall writing proficiency. Overall, PLMs with smaller parameter sizes demonstrate superior performance compared to larger-scale LLMs. Analysis of the prediction results indicates that LLMs tend to assign lower writing ratings. This may stem from their exposure to a higher proportion of high-quality argumentative texts during pretraining, while high school students' argumentative writing skills are still developing and generally fall below adult level, leading to stricter evaluations by LLMs. Additionally, data scarcity somewhat limits LLMs' performance in domain-specific tasks.

Incorporating fine-grained argument components and relations information into the input significantly enhances the performance of most models. Specifically, Longformer showed notable improvements across all evaluation metrics and achieved the optimal Precision score, while QWen and ChatGLM demonstrated substantial gains in key metrics such as QWK and Recall. However, the performance of Deepseek significantly declined, which may be attributed to inherent architectural differences in reasoning models. These findings suggest that fine-grained annotation can enhance model performance in writing evaluation to a certain extent, though further exploration is needed to effectively

Model	$P(\%)$	$R(\%)$	F_1 (%)	$Acc(\%)$	QWK
BERT$_{ft}$	71.30	69.31	69.63	72.46	0.6888
RoBERTa$_{ft}$	78.85	**81.75**	**79.95**	**78.26**	**0.7537**
Longformer$_{ft}$	70.31	70.63	70.32	69.57	0.7025
DeepSeek$_{ft}$	67.43	61.11	61.06	72.46	0.6721
Qwen$_{ft}$	61.94	48.15	50.96	62.32	0.5220
ChatGLM$_{ft}$	73.89	69.58	71.06	71.01	0.7049
Longformer$_{ft}^{\dagger}$	**79.65**	73.81	72.71	75.36	0.7271
DeepSeek$_{ft}^{\dagger}$	50.00	43.12	42.81	60.87	0.4272
Qwen$_{ft}^{\dagger}$	62.39	51.32	54.29	60.87	0.5649
ChatGLM$_{ft}^{\dagger}$	71.49	70.24	69.94	71.74	0.7244
GPT-4$_{0-shot}$	64.44	74.60	65.97	65.22	0.5952
GPT-4$_{1-shot}$	46.19	51.59	42.91	39.13	0.2014
GPT-4$_{3-shot}$	35.07	33.33	33.07	43.48	0.2515

Table 6: Results for Automated Essay Grading. † indicates that the model input combines the original essay with annotated information (including argument components and relations).

leverage argumentative information for unlocking the full potential of LLMs in automated essay grading. (Additional investigations on the impact of incorporating solely argument components or relations information are detailed in Appendix B.3.)

Notably, GPT-4 performed exceptionally well under 0-shot conditions, even surpassing the fine-tuned Qwen model. However, its performance decreased significantly in 1-shot and 3-shot settings. This phenomenon not only highlights LLMs' sensitivity to input examples but also underscores their misalignment with real-world educational scenarios in high school argumentative writing assessment. These findings suggest that specific task design requires a comprehensive consideration of domain-specific characteristics, task complexity, data scale, and model capabilities.

5 Discussion

This study investigates the importance of fine-grained annotation in enhancing argument comprehension. We explore the impact of writing quality on *Argument Component Prediction* (Task 1) and *Relation Prediction* (Task 2). Moreover, we employ the learning analytics method ENA (Epistemic Network Analysis) (Shaffer et al., 2016) to compare the differences in horizontal discourse relations and vertical argument relations between high- and low-quality essays, aiming to visualize and provide interpretable insights into discourse relations and argumentative features (see Appendix C for a detailed explanation of ENA). The grouping criteria are described in Table 1, with the ratio of high- and low-quality data being approximately 7:13.

Model	Grade	Task 1				Task 2		
		$P(\%)$	$R(\%)$	Micro-F_1	Macro-F_1	Micro-F_1	Macro-F_1	Pos.-F_1
RoBERTa	High	**47.45**	**53.33**	**50.21**	**36.48**	68.74	13.82	27.88
	Low	41.05	45.79	43.29	26.58	**69.86**	**16.07**	**29.10**
ChatGLM	High	56.96	58.06	57.51	**52.98**	65.65	29.06	39.63
	Low	**59.63**	**60.88**	**60.25**	50.42	**68.34**	**33.84**	**44.14**
GPT-4	High	**39.46**	**39.19**	**39.32**	**24.00**	2.53	**5.90**	26.37
	Low	27.72	28.43	28.07	22.82	**2.72**	4.04	**29.02**

Table 7: Performance of various models on Task 1 and Task 2 at different levels of writing quality.

5.1 The Impact of Writing Quality on Argument Component Prediction and Relation Prediction

According to Table 7, while the performances varies between different writing levels, certain patterns are evident. **In Task 1**, the models performed better on high-quality essays. Specifically, RoBERTa and GPT-4 significantly outperformed on high-quality essays compared to low-quality ones, while ChatGLM achieved comparable results across both. This suggests that high-quality essays typically feature clearer argument structures, enhancing the models' ability to identify argument components. It also indicates that LLMs, leveraging their extensive knowledge, can partially mitigate the impact of writing quality differences on argument component prediction. In contrast, **Task 2** yields markedly different results. In most cases, the models performed significantly better on relation identification in low-quality essays than in high-quality ones, suggesting that the complex and diverse relations and structures in high-quality argumentative writing pose greater challenges to the models' predictive capabilities.

These findings suggest that the impact of writing quality on model performance varies depending on task type and difficulty, underscoring the importance of considering writing proficiency in argument component prediction and relation prediction tasks. This echoes the result in Section 4.4.3 that fine-grained argument information can assist in predicting writing proficiency, collectively revealing the intricate relationships among writing quality, argument information, and model performance.

5.2 The Relationship between Argumentative Features and Discourse Relations

As shown in Figure 3, the ENA result reveals significant differences in the use of discourse and argument relations between high- and low-quality essays. **High-quality essays** are more likely to use *Concession* and *Progression* discourse relations, closely integrating *Positive*, *Example*, *De-*

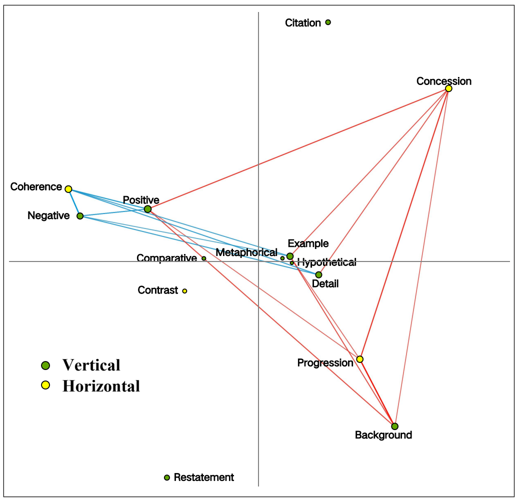


Figure 3: ENA networks of discourse and argument relations in high- (red) and low-quality essays (blue).

tail, and *Background*, which presents a logically progressive argumentation. This suggests that high-quality prefer to directly support claims using positive examples, progressively developing the argument through background information and detailed elaboration, while employing concessive relations to enhance depth and critical reasoning. In contrast, **low-quality essays** primarily focus on *Coherence*, closely combining *Positive*, *Negative*, *Example*, and *Detail*. This suggests that low-quality essays rely on basic parallel reasoning, such as balancing positive and negative argumentation, and providing support through examples and details.

Overall, high-quality essays demonstrate more critical and hierarchical use of discourse relations, incorporating rich argument relations to effectively enhance the persuasiveness of the reasoning. Conversely, low-quality essays tend to rely on simple and straightforward parallel logic, limiting the depth and effectiveness of argumentation. This finding further validates the result in Section 5.1, namely that the complex and diverse relations and structures in high-quality argumentative writing pose greater challenges to the analytical capabilities of models. Therefore, future research could enhance model performance in argumentation analysis by integrating writing proficiency with fine-grained argumentative features, thereby providing more interpretative and comprehensive support for intelligent writing education.

6 Conclusion

In this paper, we propose an innovative relation annotation scheme to characterize the argument strategies within the CEAMC (Ren et al., 2024). It integrates argument and discourse relations, covering 14 fine-grained relation types from both vertical and horizontal dimensions, thereby overcoming the simplicity and monotony of argument relations in previous studies. We conducted experimental analyses on three tasks, and the results revealed significant differences between PLMs and LLMs across different tasks, indicating that specific tasks require comprehensive consideration of factors such as domain specificity, task complexity, data scale, and model capabilities. Furthermore, additional discussions highlight the importance of fine-grained annotations for a comprehensive understanding of argumentation, emphasizing the need for multi-dimensional argument analysis.

Limitations

The limitations of our research include:

- **Data Scale** While our dataset already contains a comprehensive representation of types, it remains limited in size. The diversity and complexity of argumentation imply that the larger the dataset, the more comprehensive its coverage of these phenomena. Consequently, the current size of our dataset might limit the performance and generalization of models trained on it.

- **Manual Annotation** Our dataset relies significantly on manual annotations by linguistic experts. Nonetheless, due to the labor-intensive and time-consuming nature of this process, there are inevitable limitations on the volume of annotated data. Further, the inherent subjectivity of manual annotation might lead to potential inconsistencies and bias in the annotated labels.

- **Model Exploration** While this study provides a comprehensive analysis of argumentation capabilities in current PLMs and LLMs, with a multidimensional investigation of argument components, relations, and quality assessment, future work needs to incorporate state-of-the-art approaches to advance both argument mining and automated essay evaluation.

Acknowledgments

We appreciate the support from National Natural Science Foundation of China with the Main

Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Fundamental Research Funds for the Central Universities (2024QKT004), Pudong New Area Science & Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004), Shanghai Municipal Education Commission with the Artificial Intelligence-Powered Research Paradigm Reform and Discipline Leap Plan (2024AI02004), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Pablo Accuosto, Mariana Neves, and Horacio Saggion. 2021. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36.* CEUR Workshop Proceedings.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 1–17. Springer.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ramy Elmoazen, Mohammed Saqr, Matti Tedre, and Laura Hirsto. 2022. A systematic literature review of empirical research on epistemic network analysis in education. *IEEE Access*, 10:17330–17348.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Kamila Górska, John Lawrence, and Chris Reed. 2024. Forecast2023: A forecast and reasoning corpus of argumentation structures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7395–7405.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. Ova+: An argument analysis interface. In *Computational models of argument*, pages 463–464. IOS Press.

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.

Iman Jundi, Neele Falk, Eva Maria Vecchi, and Gabriella Lapesa. 2023. Node placement in argument maps: Modeling unidirectional relations in high & low-resource scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5854–5876, Toronto, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. Disapere: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1234–1249.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuan-Jing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Huadai Liu, Xu Wenqiang, Xuan Lin, Jingjing Huo, Hong Chen, and Zhou Zhao. 2024. Antcritic: Argument mining for free-form and visually-rich financial comments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1306–1317.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995.

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)*.

Roosmaryn Pilgram and Lotte van Poppel. 2021. The strategic use of metaphor in argumentation. *The language of argumentation*, pages 191–212.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Yupei Ren, Hongyi Wu, Zhaoguang Long, Shangqing Zhao, Xinyi Zhou, Zheqin Yin, Xinlin Zhuang, Xiaopeng Bai, and Man Lan. 2024. Ceamc: Corpus and empirical study of argument analysis in education via llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6949–6966.

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of dialam-2024: Argument mining in natural language dialogues. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 83–92.

Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024. Darius: A comprehensive learner corpus for argument mining in german-language essays. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4356–4367.

David Williamson Shaffer, Wesley Collier, and Andrew R Ruis. 2016. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3):9–45.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2021. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3875–3881.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Maja Stahl, Nadine Michel, Sebastian Kilsbach, Julian Schmidtke, Sara Rezat, and Henning Wachsmuth. 2024. A school student essay corpus for analyzing interactions of argumentative structure and quality. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2661–2674.

Sophie Vanbelle. 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81(2):399–410.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Appendix

A More Details of Corpus

A.1 Annotation Scheme and Samples

By deeply integrating argument relations with discourse relations, we propose 14 fine-grained relation types from both vertical and horizontal dimensions, thereby capturing the intricate interplay between argument components for a thorough understanding of argument structure. Detailed definitions and examples of argument relations in the vertical dimension are presented in Table 8 and discourse relations in the horizontal dimension are presented in Table 9.

A.2 Confusion Matrix of Relation Annotation

Figure 4 showcases the confusion matrix heatmap results of the relation type annotations, with a detailed analysis as follows:

High-Distinction Types *Restatement*, *Example*, and *Citation* argumentation demonstrate the best performance with clear semantic boundaries. This stems from the explicit binding and mapping between these argument relations and argument components, ensuring high certainty in annotation once the associated components are identified.

High-Confusion Types

- **Argument Relations** Significant confusion exists between *Positive and Negative* argumentation. Key reasons include: (1) the indirect reasoning nature of Negative argumentation increases identification difficulty; (2) their low frequency further reduces annotation accuracy. Additionally, *Detail* relations exhibit widespread confusion, attributable to the multifunctionality of Elaboration components (e.g., supporting claims, detailing points, or providing logical transitions).
- **Discourse Relations** Bidirectional confusion occurs between *Progression and Coherence* relations, while *Contrast and Concession* relations also show notable overlap. Potential reasons include: (1) unclear or disorganized student argumentation complicating logical relation judgment; (2) the lack of connectives posing additional challenges.
- **Special Case** *Metaphorical* argumentation exhibits a high proportion of unlabeled samples (no_relation), likely due to students' nonstandard usage, hindering recognition.

B More Details of Experiments

B.1 Automated Essay Grading Prompt

In the automated essay grading task, to investigate the impact of fine-grained annotation information on model performance, we incorporate argument components and relations as joint inputs in addition to the original essays. Specifically, this refers to encoding relevant argument components and relations information as prompt inputs to the model using sentence IDs (see Figure 5 for example).

B.2 Fine-grained Results of Relation Prediction

To further evaluate the models' prediction performance across different relation categories, we conducted a fine-grained analysis of relation prediction tasks, with the results presented in Table 10. Notably, under the SFT setting, LLMs outperformed PLMs on most metrics, demonstrating certain outcomes even in low-resource relation types such as *Metaphorical*, *Progression*, *Contrast*, and *Concession*. This highlights the superior capability of LLMs in relation prediction tasks, particularly when handling imbalanced data distributions and low-resource scenarios. In contrast, GPT-4 performed poorly in the ICL setting, indicating that the intricate interactions and high-level abstractions involved in argument components within relation prediction tasks pose significant challenges to the model's learning process.

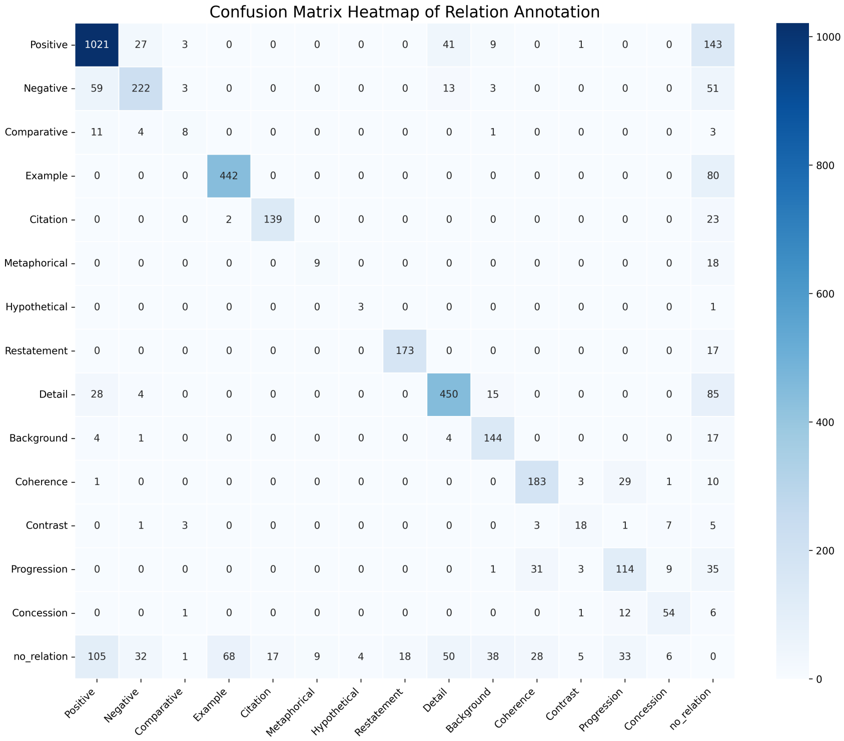


Figure 4: Confusion matrix heatmap of relation annotation.

B.3 Supplementary Results on Automated Essay Grading

To further investigate the role of different annotation information,we supplemented the comparison experiments for combining only argument components or relations as input, with results shown in Table 11.

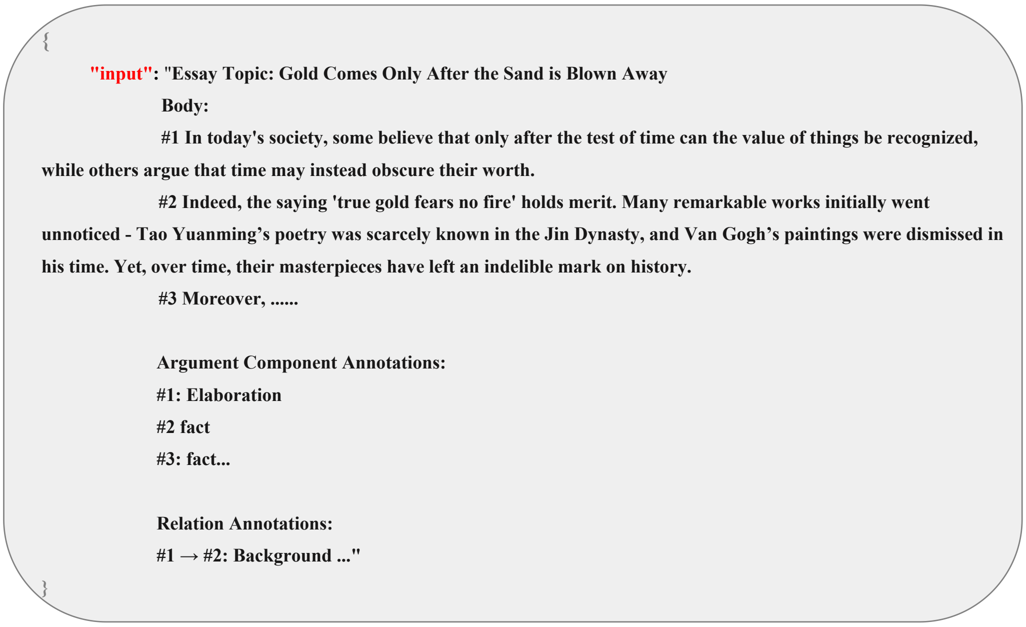
The experimental results are generally consistent with the main findings of the paper, while also revealing several new insights. First, the results reaffirm the superiority and stability of PLM in automated essay evaluation. Longformer achieved the best performance, demonstrating significant improvements across all evaluation metrics under three argument annotation input conditions. Second, LLMs exhibited notable sensitivity to input information and model-specific variations: Qwen achieved consistent QWK improvements under all three input settings (AC, Re, and Both), ChatGLM exhibited performance improvements under the Both annotation setting, while DeepSeek experienced significant declines in all scenarios.

These findings indicate that: (a) Smaller-parameter models can achieve satisfactory performance relying solely on essay content; (b) While the integration of fine-grained argument components and relations can enhance evaluation to some extent, the actual impact varies substantially across models; (c) Effective utilization of fine-grained argumentative information to improve LLMs performance in automated assessment remains a critical research direction.

C Concept of ENA

Epistemic Network Analysis (ENA) describes the co-occurrence structure within discourse data or any paragraph-based textual data (Shaffer et al., 2016). It has become a mainstream research method in the field of learning analytics, widely used for analyzing and modeling the relationship within collaboration, learning, and cognitive activities (Elmoazen et al., 2022). ENA helps researchers understand complex cognitive and interactive processes by visualizing the co-occurrence relationships among different elements.

In ENA, nodes represent distinct concepts, behaviors, or themes. In the context of this study, nodes represent specific **argument relations** or **discourse relations**. The edges connecting these nodes indicate the co-occurrence relationships between them, with the strength of these connections represented by the weight (or thickness) of the edges. The edge weight reflects the frequency of co-occurrence between two nodes, where a greater weight indicates a stronger relationship.



```
{
    "input": "Essay Topic: Gold Comes Only After the Sand is Blown Away
              Body:
              #1 In today's society, some believe that only after the test of time can the value of things be recognized,
while others argue that time may instead obscure their worth.
              #2 Indeed, the saying 'true gold fears no fire' holds merit. Many remarkable works initially went
unnoticed - Tao Yuanming’s poetry was scarcely known in the Jin Dynasty, and Van Gogh’s paintings were dismissed in
his time. Yet, over time, their masterpieces have left an indelible mark on history.
              #3 Moreover, ......

              Argument Component Annotations:
              #1: Elaboration
              #2 fact
              #3: fact...

              Relation Annotations:
              #1 → #2: Background ..."
}
```

Figure 5: Example of integrated argument component and relation annotations as model input.

Aspect	**Label**	**Definition**	**Example**
Stance-Based	Positive	A method that directly validates the correctness of a viewpoint by using elaboration or evidence consistent with the viewpoint to support it, emphasizing direct affirmation of the viewpoint.	Quotation: Nietzsche once said, "Every day that you do not dance is a betrayal of life." → Claim: Exploring the spiritual world is an individual's journey of self-awareness—a process of exercising subjective initiative to recognize one's own uniqueness.
	Negative	A method that indirectly proves the correctness of a viewpoint through elaboration or evidence that are contrary to the viewpoint. It emphasizes the negation of opposing viewpoints, thereby achieving the purpose of the argumentation.	Quotation: As Shakespeare said, "Without surprises, life would have no luster." → Claim: Under a certain sense of ceremony, people can become more passionate about life, helping them cherish the moment and look forward to the future.
	Comparative	A shorthand for positive and negative argumentation, is an argumentative approach that involves contrasting and comparing two items to highlight their differences, thereby making the conclusion more evident and persuasive.	Fact:Take the recent marathon as an example: many contestants did not finish the race, some even quitting midway. This occurred because one runner started accelerating early on, prompting others not to fall behind, a manifestation of tension. Conversely, those who maintained their composure and were undisturbed ended up securing better positions, illustrating the benefits brought by a sense of relaxation. → claim: In real life, we need a sense of relaxation more than tension.
Evidence-Based	Example	An argumentation method that proves a thesis through concrete, or typical examples.	Fact: The flourishing Tang Dynasty, despite its grandeur, is reduced to fleeting pages in historical records. Without ritualistic significance and the poetic brilliance of Li Bai, Du Fu, and others, how could we today appreciate the splendor of ancient Chang'an or comprehend the complex emotions embedded in phrases like 'returning to Chang'an as one's homeland'? → Claim: Ritualistic significance adds brilliance to mundane life, liberating individuals from mediocrity in that moment and infusing dull emotions with romantic yearning for beauty.
	Citation	An argumentation method that proves a thesis by using quotations or axioms.	Quotation: Nietzsche once said, "Every day that you do not dance is a betrayal of life." → Claim: Exploring the spiritual world is an individual's journey of self-awareness—a process of exercising subjective initiative to recognize one's own uniqueness.
Discourse-Based	Metaphorical	By employing metaphorical rhetoric, familiar things are used as metaphors to argue the correctness of a viewpoint. In drawing parallels between two items with similar characteristics, the artful use of metaphors often serves to better elucidate concepts, making the argument more vivid and interesting.	Elaboration: If understanding objects is likened to baking a cake, then the method of comprehension is the mold. Those who only heed the words of authoritative experts apply others' molds; thus, no matter how sweet the resulting cake is, it will not be in a shape that suits them. → Claim: A deep-rooted reliance on authoritative experts also reflects a more profound issue – a lack of fundamental methods for understanding things oneself.
	Hypothetical	Analyzing evidence from the opposite side based on hypothesis to infer its authenticity and reliability, thus robustly supporting a thesis.	Fact: The grandeur and brilliance of the Tang Dynasty, though but a fleeting mention in the annals of history, would be lost to us without the ceremonial gravitas and the exquisite verses of poets like Li Bai and Du Fu. How else could we, in the present day, glimpse the golden splendor of ancient Chang'an or grasp the myriad emotions encapsulated in the phrase "Returning to Chang'an, my homeland" ? → Claim: Ceremony adds a luster to the mundane, lifting those numbed by the monotony of daily life out of their mediocrity, infusing their arid emotions with a romantic yearning for the beautiful.
	Restatement	For argument of the type *restated claim*, its relation with the target argument (*major claim* or *claim*) is defined as restatement relation.	Restated Claim: Rituals are never unnecessary or superfluous. → Major claim: In life, rituals are just so indispensable.
	Detail	When an argument (*elaboration* type) primarily aims to further explain or analyze other content, it establishes a detail relation with the corresponding argument (*assertion* or *evidence* type).	Elaboration: Nietzsche's words actually tell us to know thyself and become thyself, which all but maps out the exploration of the spiritual world of self. → Quotation: Nietzsche once said, "Every day that you don't dance is a failure of life.
	Background	When an argument (*elaboration* type) primarily serves the function of introducing background, it constructs a background relation with the corresponding argument (*assertion* or *evidence* type).	Elaboration: It's just that is such a mode of exploration really beneficial to people's perceptions?" → Claim: This process of transformation essentially reflects the expansion of instrumental rationality and people's active abandonment of "thinking".

Table 8: A list of argument relations in the vertical dimension, their descriptions and samples. Argument component types are indicated in blue, with the argument before and after the → corresponding to the source component and target component, respectively. It is noteworthy that multiple argument relations may exist between argument-pair, and these relations can occur between different types of argument components.

Label	**Definition**	**Example**
Coherence	Describing several aspects of the same event, related events, or contrasting situations that coexist, co-occur, or oppose in meaning. These aspects can be reordered without altering the overall significance.	Fact: The idea of a commonwealth of nations, as proposed by Confucius, is also what we aspire to nowadays. → Fact: Another example is Wang Mang's seizure of power and his promulgation of a series of new measures, which were denied at the time, but in fact he referred to Western countries for these initiatives.
Progression	The subsequent argument represents an advance in scope or meaning than the preceding one, intended to emphasize a deepening, expansion, or reinforcement of logic, and the order of the arguments is usually non-interchangeable.	Claim: However, the negative impacts caused by the pursuit of rituals are not few. → Claim: Only by getting rid of the solidified idea that a sense of ritual is necessary in life can they focus on the abundance of the spiritual world and climb higher.
Contrast	Comparison and selection are made by examining the similarities or differences between two or more things, situations, or viewpoints, emphasizing the contrast between them.	Fact: We all know that Wei Liangfu improved the Kunqu opera, leaving brilliant cultural treasures for future generations, we all know that Yuan Longping broke through a technical barrier to solve the food problem in many areas, they are not precisely in the ancients and the authority of the forefathers under the influence of their own chapter? → Fact: There are great men, naturally, there are also small people, those so-called good learning in fact, "thick ancient and thin" academic molecules, those who listen to the authority of the scientific molecules do not understand the development of adaptability, which one has made achievements?
Concession	An argument posits a certain situation or viewpoint, followed by a shift where the subsequent argument presents an opposing or contrasting perspective, emphasizing the content of the latter argument.	Claim: Therefore, while inheritance is important, breakthroughs and development are also indispensable. → Claim: However, should those ideas and factors that have been tested be recognized in their entirety? No.

Table 9: A list of discourse relations in the horizontal dimension, their descriptions and samples. Argument component types are indicated in blue, with the component before and after the → corresponding solely to the order in which the two arguments appear in the essay. It is noteworthy that the discourse relations between argument-pair is singular and occurs between argument components of the same type.

Model	Vertical										Horizontal				Neg.	Macro-F_1	Micro-F_1
	Posi.	Nega.	Comp.	Exam.	Cita.	Meta.	Hypo.	Rest.	Deta.	Back.	Cohe.	Prog.	Cont.	Conc.			
BERT$_{ft}$	25.52	8.57	0.00	21.67	25.00	6.90	10.34	13.60	0.00	11.11	9.05	0.00	0.00	0.00	**84.41**	14.41	**73.72**
RoBERTa$_{ft}$	28.67	7.76	0.00	26.62	27.45	8.12	6.38	22.22	0.00	11.11	16.86	2.99	0.00	0.00	81.42	15.97	69.39
DeepSeek$_{ft}$	34.39	21.94	0.00	31.14	33.50	0.00	0.00	31.20	23.31	11.35	30.25	5.16	12.41	15.27	80.16	22.00	65.14
Qwen$_{ft}$	**36.59**	23.37	6.06	34.06	35.52	41.11	**66.67**	39.44	26.29	17.68	**43.89**	7.48	12.94	10.58	79.99	**32.11**	65.32
ChatGLM$_{ft}$	36.33	**24.00**	**15.74**	**39.73**	**39.64**	**33.97**	8.33	**49.21**	**27.92**	**18.55**	40.42	**12.53**	**26.08**	**19.16**	81.26	31.52	67.11
GPT-4 $_{0-shot}$	5.96	2.25	2.63	5.06	10.59	3.45	0.00	21.98	3.10	1.09	5.71	4.40	0.81	1.63	1.07	4.65	2.64
GPT-4 $_{1-shot}$	4.49	2.30	0.00	7.25	7.41	6.90	0.00	12.59	1.80	1.89	14.29	3.68	0.98	3.49	18.76	5.72	10.61
GPT-4 $_{3-shot}$	5.49	3.48	0.00	4.96	9.02	1.92	0.00	13.89	5.52	1.11	14.29	4.16	1.53	2.25	5.70	4.89	4.97

Table 10: Performance of various models on the fine-grained Relation Prediction task. Displayed are the F_1 scores (%) of each type, with the best results in **bold**. Posi. denotes Positive, Nega. denotes Negative, Comp. denotes Comparative, Exam. denotes Example, Cita. denotes Citation, Meta. denotes Metaphorical, Hypo. denotes Hypothetical, Rest. denotes Restatement, Deta. denotes Detail, Back. denotes Background, Cohe. denotes Coherence, Prog. denotes Progression, Cont. denotes Contrast, Conc. denotes Concession, and Neg. denotes negative samples without relations.

Model	**Setting**	$P(\%)$	$R(\%)$	F_1 (%)	$Acc(\%)$	**QWK**
Longformer	Essay	70.31	70.63	70.32	69.57	0.7025
	+AC	76.87	**73.81**	**73.90**	**75.36**	**0.7597**
	+Re	74.83	73.02	73.17	73.91	0.7454
	+Both	**79.65**	**73.81**	72.71	**75.36**	0.7271
DeepSeek	Essay	**67.43**	**61.11**	**61.06**	**72.46**	**0.6721**
	+AC	51.57	45.50	46.28	60.87	0.4868
	+Re	47.45	43.65	43.44	60.87	0.5234
	+Both	50.00	43.12	42.81	60.87	0.4272
Qwen	Essay	61.94	48.15	50.96	**62.32**	0.5220
	+AC	**67.41**	48.41	52.22	60.87	0.5409
	+Re	60.98	**53.17**	**54.99**	60.87	**0.5804**
	+Both	62.39	51.32	54.29	60.87	0.5649
ChatGLM	Essay	**73.89**	69.58	**71.06**	71.01	0.7049
	+AC	62.70	66.67	64.05	68.12	0.6977
	+Re	68.63	64.81	65.68	68.12	0.6775
	+Both	71.49	**70.24**	69.94	**71.74**	**0.7244**

Table 11: Results for Automated Essay Grading. AC denotes Argument Components, Re denotes Relations, Both denotes Argument Component & Relations. The overall best results are indicated by underlining.