CEAMC: Corpus and Empirical Study of Argument Analysis in Education via LLMs

Anonymous EMNLP submission

Abstract

This paper introduces the **C**hinese **E**ssay **A**rgument **M**ining **C**orpus (**CEAMC**), a comprehensive dataset for fine-grained argument analysis. Existing argument types in education remain simplistic and isolated, failing to encapsulate complete argument information. Originating from authentic examination settings, CEAMC transcends previous simple representations by conducting multi-level delineation of argument components, thus capturing the subtle nuances of argumentation in the real world and meeting the needs of complex and diverse argumentative scenarios. Our contributions include the development of the CEAMC, the establishment of baselines for further research, and an in-depth exploration of the performance of Large Language Models (LLMs) on CEAMC. The results indicate that our CEAMC can serve as a challenging benchmark for the development of argument analysis in the field of education.[1]

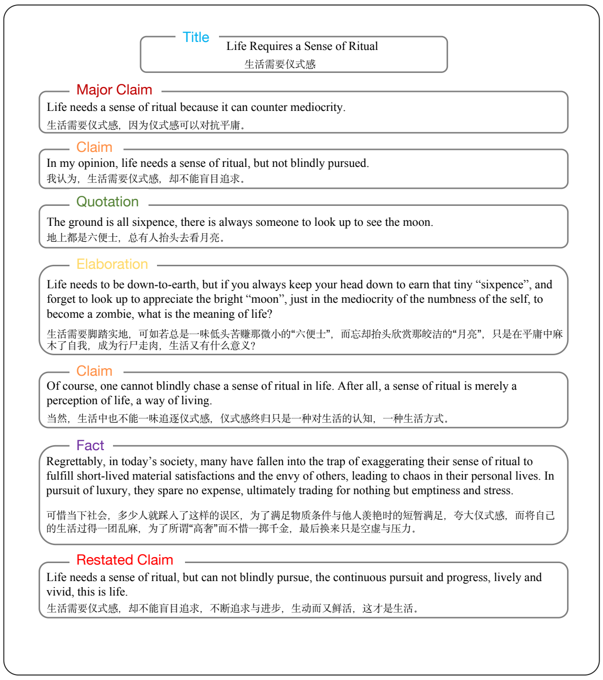

Figure 1: An excerpt from an argumentative essay in CEAMC.

1 Introduction

Argument mining (AM) aims to automatically identify and extract the structure of inference and reasoning expressed as arguments presented in natural language (Lippi and Torroni, 2016). Due to its significance, it has been widely incorporated into various natural language processing (NLP) tasks, such as argument evaluation (Ruiz-Dolz et al., 2023), fallacy detection (Goffredo et al., 2023) and text generation (Zhao et al., 2023; Lin et al., 2023).

With the surge in argumentative texts and advancements in NLP technology, AM has been developed in various domains, such as court decisions (Teng and Chao, 2021; Habernal et al., 2023), political debates (Menini et al., 2018; Goffredo et al., 2023), scientific literature (Si et al., 2022; Liu et al., 2023a), social web (Habernal and Gurevych, 2017; Gupta et al., 2021), and online comments (Park and Cardie, 2018; Scheibenzuber et al., 2023). These efforts have introduced various annotation schemes and datasets in conjunction with domain specificity, significantly advancing argumentation research.

However, existing datasets struggle to fulfill the needs for argument analysis in education. **Primarily**, current research either focuses on high-quality argument scenarios, such as legal texts (Habernal et al., 2023), and peer reviews(Purkayastha et al., 2023), where the argumentative texts are logically rigorous, highly professional, and persuasive. Alternatively, it targets online scenarios like social media (Lin et al., 2023) and online writing (Song et al., 2021), where argumentative texts tend to be more fragmented and colloquial. These corpora exhibit significant differences in argument quality, textual traits, and writing styles compared to argumentative essays in educational settings, ne-

[1] We will make the corpus and related code available for research.

cessitating datasets that can reflect the unique complexity and nature of educational writing. **Furthermore**, there remains a considerable discrepancy between the argument studies conducted by NLP researchers and the analysis of argumentative essays by teachers. Computational approaches typically simplify arguments into generic major claims, claims and premises (Stab and Gurevych, 2017; Wambsganss and Niklaus, 2022), which fall short of reflecting the realities of educational argumentation. In fact, argumentative essays in education usually encompass a rich variety of argument types, which is crucial for gaining insight into argument structures and support strategies. **Lastly**, the scarcity and limited diversity of Chinese argument mining datasets have somewhat constrained advancements in this field.

To address the shortcomings of existing research, we introduce the **C**hinese **E**ssay **A**rgument **M**ining **C**orpus (**CEAMC**). The corpus is derived from authentic high school examination scenarios, and as illustrated in Figure 1, each argumentative essay undergoes meticulous annotation. The CEAMC addresses key limitations in prior work: **firstly**, it bridges the gap between current corpora in fulfilling the needs of argument analysis in education. Considering the pivotal role of argumentation in K12 education, we have curated a corpus of argumentative essays from high school examination scenarios, covering a variety of topics, qualities, and rich argumentative information, which adequately reflects the complexity and uniqueness of educational argumentation scenarios and can provide a more reliable basis for argumentation assessment and instruction. **Secondly**, it overcomes the issue of simplified argument types prevalent in previous studies. By deeply integrating argument mining research with educational practice, it provides 4 coarse-grained and 10 fine-grained argument component types, which can adeptly capture the nuances of real-world argument texts and facilitate a thorough and comprehensive analysis of argumentation. **Lastly**, by providing a diverse dataset for Chinese argument mining and conducting comprehensive experimental analyses, CEAMC stimulates progress in this area.

Our contributions are summarised as follows:

- We develop CEAMC, the currently most comprehensive Chinese dataset for evidence-based argument mining, including detailed annotations of arguments based on student argumentative essays, which not only provides a valuable data resource for AM but also facilitates the advancement of intelligent education.

- We conduct extensive experiments on CEAMC, comparing the performance of current mainstream methods, benchmarking argument component detection task against our dataset, and providing a reference point for future research.

- To further explore the domain adaptation of LLMs on CEAMC, we test a range of LLMs under various methods including Supervised Fine-Tuning (SFT), In-context Learning (ICL), and Chain of Thought (CoT), showing that the proposed dataset can serve as a challenging benchmark for the development of argument component detection in education.

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 the identification of 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. Several studies have extended the types of argument components from the perspective of sentence function. For example, Kennard et al. (2022) focused on review and rebuttal texts and presented the various sentence types such as *request*, *social* and *structuring* for a more exhaustive understanding. Additionally, research in different domains has further classified argument types based on evidence attributes, such as *news*, *expert*, and *blog* in social media (Addawood and Bashir, 2016); *policy*, *value*, and *testimony* in online comments (Niculae et al., 2017); and *case*, *expert*, and *research* in English Wikipedia (Guo et al., 2023). Concerning argument relations, researchers also adapt additional relation types from Rhetorical Structure Theory (Mann and Thompson, 1988) such as detail, sequence (Kirschner et al., 2015), semantically same (Lauscher et al., 2018), by-means, info-required and info-optional (Accuosto et al., 2021), which hold significant value in scientific literature. These studies have enriched argument schemes and facilitated a holistic comprehension of argument structures. However, they primarily focus on high-quality argument domains or online scenarios, where the corpora differ significantly in professionalism, argument traits, and writing style compared to the educational domain, as well as the highly domain-specific of the annotation schemes, making it difficult to apply to educational argumentation.

The corpus proposed by Stab and Gurevych (2014, 2017) marks the first attempt of computational argumentation in the field of education. The argumentative essays within this corpus originate from an online forum, encompassing basic three components and two relations. Building on this, Ke et al. (2018) randomly select 102 essays from the corpus to annotate argument attributes for assessing persuasiveness. Subsequently, Ke et al. (2019) design a set of more refined scoring criteria and expand their research based on the International Corpus of Learner English (ICLE) (Granger et al., 2009), which primarily consists of essays on various subjects written by university students with diverse native language backgrounds. Additionally, Song et al. (2021) define five sentence functions (i.e., introduction, thesis, main idea, evidence, elaboration, and conclusion) to evaluate the organization of essays. Recently, Wambsganss and Niklaus (2022) collect German business pitches from university lectures to assess the persuasiveness of argumentative writing. These efforts have advanced argumentation research in education. However, they all focus solely on the most basic argument types and fall far short of covering the complexity and variety of arguments in real educational scenarios, limiting their further development.

2.2 LLMs in Argument Mining

Recently, LLMs such as ChatGPT [2] have demonstrated their capabilities in various NLP tasks. In the realm of argument mining, researchers have explored the power of LLMs in stance detection (Zhao et al., 2023) and financial argument relation recognition (Otiefy and Alhamzeh, 2024). Furthermore, Chen et al. (2023) systematically evaluate the performance of LLMs in multiple computational argumentation tasks in zero-shot and few-shot settings. Mirzakhmedova et al. (2024) focus on the potential of LLMs as proxies for argument quality annotators. Currently, research on LLMs in argument mining is still in its nascent stage, and to our knowledge, there has not been a systematic exploration of LLMs in Chinese argument mining.

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

3 Corpus Construction

This section delineates the process of collection and annotation for the **C**hinese **E**ssay **A**rgument **M**ining **C**orpus (**CEAMC**), designed for extensive argument mining research.

3.1 Data Collection

For the construction of CEAMC, we collect 226 argumentative essays from high school examination scenarios. These essays range from 557 to 1,101 tokens with an average of approximately 829.82 tokens, where the writing requirement is no less than 800 tokens. Figure 2 depicts the distribution of score ranges for the selected essays, where the scores represent the comprehensive evaluations awarded by educators, and the categorization of score ranges are derived from the authoritative scoring standards.

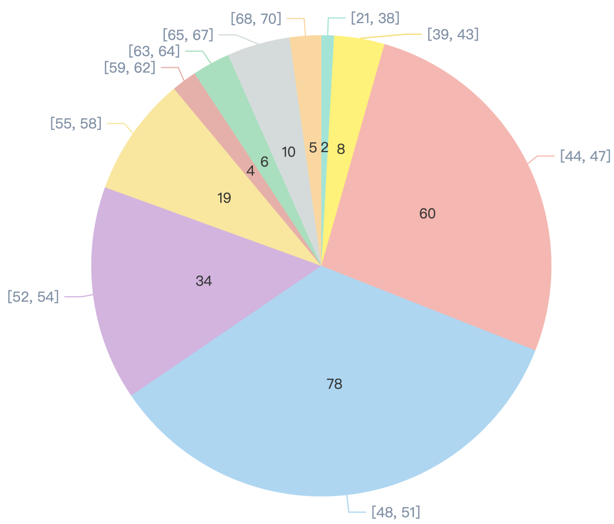

Figure 2: Distribution of score ranges in CEAMC. The internal numbers represent the number of essays in each score range, totalling 226.

We specifically chose persuasive essays from high school exams for their significance in argument mining research. On the one hand, these essays from authentic educational settings encapsulate rich argumentative information, offering a unique perspective for insightful exploration of argument strategies and structures. On the other hand, argumentative essays within an examination context can reflect the actual state of students' argumentative writing skills to a certain extent, serving as a vital resource for assessing and enhancing students' argumentation abilities. Lastly, as high school is a pivotal period for students to learn argumentative writing and develop critical thinking

(Hess and McShane, 2014), filling this data gap will aid in the progress of intelligent education.

3.2 Annotation Scheme

The classic Toulmin model of argument (Toulmin, 2003) revolves around three key elements: a claim, or the assertion to be argued for, data that provide supportive evidence (empirical or experiential) for the claim, and a warrant that explains how the data support the claim. Regarding argument relations, Stab and Gurevych (2017) attempt to distinguish them into support or attack, with the latter being in lesser quantity. However, Wambsganss and Niklaus (2022) did not find any attack relation in 200 business pitches. Additionally, Song et al. (2021) did not mark the relations in Chinese argumentative essays, implying subtly that there exists a supportive relation between evidence and claim.

Taking into account argument mining research with educational practice, we focus on the argument types in argumentative essays by defining and categorizing them in detail to meet the needs of complex argumentation. Following previous studies (Song et al., 2021; Kennard et al., 2022; Guo et al., 2023), we annotate at the sentence level, not only to avoid the propagation of argument detection errors, but also because of high probability of aligning argument units with sentence boundaries.

In CEAMC, we define 4 coarse and 10 fine-grained argument types, as follows:

Assertion Assertions are further subdivided into *major claim*, *claim* and *restated claim*. Major claim and claim are common components of argument, used to express the primary assertion and its supporting views, respectively. Restated claim typically appears at the end of paragraphs or documents to emphasize the importance of the claim or major claim, a common practice in argumentative writing.

Evidence To more comprehensively understand the sources and attributes of evidence, aiding in the assessment of an argument's persuasiveness and sufficiency, we further classify it into five types: *fact*, *anecdote*, *quotation*, *proverb*, and *axiom*.

Elaboration *Elaboration* includes the further presentation, explanation, or analysis of assertions or evidence.

Others *Others* refers to sentences that do not fit into any of the aforementioned cases.

For a detailed overview of our argument types annotation scheme and samples, please refer to Appendix A.

3.3 Annotation Process

Our annotation team consists of expert reviewers and students from the fields of linguistics and education, all of whom received training prior to commencing the annotation work. The dataset was divided into three groups for efficient and consistent annotation. The entire annotation process took three months and included detailed annotation of sentence types (i.e., argument components), with a total of 226 essays. For a detailed overview of the annotation process, please refer to Appendix B.

3.4 Inner Annotator Agreements

To evaluate the reliability of the argument component annotations, we follow the approach of Kennard et al. (2022) and Cheng et al. (2022), using Cohen's kappa to computed the Inter-Annotator Agreement (IAA). A total of 4,726 sentences are labeled and the average Cohen's kappa is 75.62% between the three groups of annotators, which is a reasonable and relatively high agreement considering the annotation complexity (Cheng et al., 2022; Kennard et al., 2022) . Further details on IAA calculation can be found in Appendix C.

Coarse	Fine-grained	# Freq.	# AvgTok.	% of Total
Assertion (1,013)	Major Claim	232	36.69	4.91%
	Claim	583	32.39	12.34%
	Restated Claim	198	32.05	4.19%
Evidence (1,124)	Fact	882	52.37	18.66%
	Anecdote	20	49.65	0.42%
	Quotation	205	36.91	4.34%
	Proverb	9	30.89	0.19%
	Axiom	8	47.00	0.17%
Elaboration (2,535)	-	2,535	38.42	53.64%
Others (54)	-	54	19.13	1.14%
Total	-	4,726	39.69	100.00%

Table 1: Distribution and average tokens of annotated argument types. # Freq. and # AvgTok. denote the frequency and average token of each type, respectively.

3.5 Data Statistics and Analysis

The final corpus consists of 226 Chinese argumentative essays containing 4,726 sentences, and the distribution of argument types is shown in Table 1. *Elaboration* is the most frequent argument type (with 2,535 instances), consistent with the typical requirements of argumentative essay writing, where extensive elaboration is often used to clarify the viewpoint or the evidence supporting their argument. In stark contrast, the evidence subcategories, especially *proverb* and *axiom*, account for fewer than 10 instances each, indicating a relative scarcity of argumentative resources among students.

Dataset	Lg.	Domain	# Doc.	# Sent.	# AvgSent.	# AvgTok.
Niculae et al. (2017)	En	Online Forum (comment)	731	3,800	5.20	120.38
Fergadis et al. (2021)	En	Scientific Literature (abstract)	1,000	12,374	12.37	263.25
Cheng et al. (2022)	En	English Wikipedia (article)	1,010	69,666	68.98	1451.95
Stab and Gurevych (2014)	En	Online Forum (essay)*	90	1,673	18.59	387.97
Stab and Gurevych (2017)	En	Online Forum (essay)*	402	7,116	17.70	366.35
Ke et al. (2018)	En	Online Forum (essay)*	102	1,462	14.33	240.37
Song et al. (2021)	Zh	Online Forum (essay)*	1,220	32,433	26.58	558.27
Wambsganss and Niklaus (2022)	De	University Lecture (business pitch)*	200	3,207	16.04	309.82
CEAMC	**Zh**	**High School Examination (essay)***	**226**	**4,726**	**20.91**	**829.82**

Table 2: Comparison between CEAMC and other datasets, the upper section represents data from online platforms, while the lower section indicates data from real-world scenarios. * denotes the educational domain corpus. Lg. denotes language: En for English, Zh for Chinese, and De for German. # Doc. and # Sent. denote the total number of documents and sentences. # AvgSent. and # AvgTok. denote the average sentences and tokens of each essay.

Furthermore, Table 2 illustrates the comparison between CEAMC and argumentation datasets from other domains and sources. It is evident that, excluding Wikipedia articles, the context of CEAMC (i.e., # AvgTok.) is significantly longer compared to existing datasets, especially when contrasted with similar argumentative essay corpora. Although CEAMC contains fewer essays than some online corpora, its richness in sentences and longer textual content partially compensates for the lower quantity. Additionally, collecting a large amount of high-quality data in real-life scenarios poses significant challenges.

Fine-grained	Train Num (Prec.)	Dev Num (Prec.)	Test Num (Prec.)
Major Claim	184 (4.92%)	25 (4.98%)	23 (4.78%)
Claim	460 (12.29%)	64 (12.75%)	59 (12.27%)
Restated Claim	157 (4.19%)	18 (3.59%)	23 (4.78%)
Fact	728 (19.45%)	66 (13.15%)	88 (18.30%)
Anecdote	14 (0.37%)	4 (0.80%)	2 (0.42%)
Quotation	152 (4.06%)	29 (5.78%)	24 (4.99%)
Proverb	7 (0.19%)	1 (0.20%)	1 (0.21%)
Axiom	6 (0.16%)	1 (0.20%)	1 (0.21%)
Elaboration	2,000 (53.43%)	284 (56.57%)	251 (52.18%)
Others	35 (0.94%)	10 (1.99%)	9 (1.87%)

Table 3: Data split statistics for benchmark testing. Train/Dev/Test Num (Perc.) denotes the count and percentage of each type in the train/dev/test set.

4 Experiments

Having constructed CEAMC, we conduct an empirical study to benchmark the performances of some existing methods on on the task of argument component detection against our dataset. To address this task, we split our data as summarized in Table 3, a total of 226 labelled argumentative essays are split by roughly 8:1:1. To avoid excessive variance, we manually adjust the randomized splits to ensure diversity balance of data.

4.1 Task

Argument component detection aims to identify argument units and determine their argument types. As described in Section 3.2, our data is annotated at the sentence level, so we formulate the argument component detection task as a sentence-level classification problem, aimed at recognising fine-grained argument types in argumentative essays.

4.2 Experiment Setup

As shown in Table 3, argument types are highly imbalanced. Hence, The task is a 10-way classification with imbalanced data, each sentence consisting one single category label. In line with Liu et al. (2023b), we employ F_1 score for each argument component category and their Macro-F_1 to measure the performance. Additionally, considering the significant imbalance of CEAMC, we also report the Micro-F_1 results.

Supervised Fine-Tuning (SFT) We experiment on three well-established pretrained language models (PLMs): *BERT* (Kenton and Toutanova, 2019), *RoBERTa* (Liu et al., 2019), and *Longformer*(Beltagy et al., 2020). Specifically, we implement BERT-Base-Chinese, which is pre-trained on Chinese corpora and captures rich semantic and syntactic information. As for RoBERTa, we use Chinese-RoBERTa-wwm-ext (Cui et al., 2021), a Chinese pre-trained BERT with whole word masking. Given the lengthy context of CEAMC, we employ Longformer due to its ability to capture contextual information from long input texts.

Given the recent unparalleled achievements of autoregressive LLMs in various NLP tasks, we also evaluate the performance of a range of different open-source Chinese LLMs on CEAMC using

Model	Assertion			Evidence					Elaboration	Others	Macro-F_1	Micro-F_1
	Major Claim	Claim	Restated Claim	Fact	Anecdote	Quotation	Proverb	Axiom				
BERT	44.44	36.19	48.89	71.90	0.00	74.42	0.00	0.00	78.23	36.36	39.04	69.02
RoBERTa	41.03	49.48	29.41	85.23	0.00	75.56	0.00	0.00	**81.65**	36.36	39.87	74.43
Longformer	37.50	32.38	27.78	50.00	0.00	52.63	0.00	0.00	71.11	0.00	27.14	59.04
Baichuan2-7B	44.90	52.43	**55.00**	**85.26**	0.00	78.05	**66.67**	0.00	80.93	31.58	49.48	74.43
ChatGLM3-6B	50.00	52.63	44.44	73.74	0.00	68.18	0.00	0.00	77.01	0.00	36.60	69.23
Qwen1.5-7B	**51.06**	**55.46**	52.00	83.06	**100.00**	**79.07**	**66.67**	0.00	81.07	**61.54**	**62.99**	**74.64**

Table 4: Performance of various models on the fine-grained argument component detection task in SFT setting. Displayed are the F_1 scores (%) of each type, with the best results in **bold** and the second best results underlined.

instruction-tuning with the LoRA technique (Hu et al., 2021). Specifically, we utilize *Baichuan2-7B* (Yang et al., 2023), *ChatGLM3-6B* (Du et al., 2022), and *Qwen1.5-7B* (Bai et al., 2023). We conduct experiments using the recommended hyperparameter settings for all LLMs.

In-Context Learning (ICL) We introduce two direct prompting methods: *Zero-shot Learning*, a direct prompting method with minimal instructions and *Few-shot Learning* (Brown et al., 2020), which adds a few correctly categorized samples to the prompt (see Appendix D.1 for complete prompts). We directly call the closed-source APIs of each model, including OpenAI's ChatGPT[2] (i.e., GPT-3.5-turbo and GPT-4-turbo), qwen-turbo[3], glm-3-turbo[4], and Baichuan2-Turbo[5] for comparison. The reason for choosing closed-source models of Chinese LLMs is their markedly superior foundational performance compared to the corresponding open-source models, thereby enabling a more precise investigation into the boundaries of Chinese LLMs on CEAMC, as well as facilitating a more in-depth comparison with GPT. Only the test set is used, and we run 3 times and report the average results.

Chain of Thought (CoT) We introduce the CoT prompting strategy to generate intermediate reasoning steps (Wei et al., 2022), aiming to explore the capabilities of LLMs in simulating the human process of step-by-step argument analysis (see Appendix D.2 for complete prompt). The models and settings used here are consistent with those in ICL.

4.3 Implementation Details

For PLMs, we adopt AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate of $2e^{-5}$ to update the model parameters, and set batch size to 8. For open-source LLMs, we employ LoRA with the LoRA rank of 8 and the dropout rate of 0.1 across all training sessions. Training configurations include the learning rate of $5e^{-5}$ and the batch size of 2. In addition, we implemented a Cosine learning rate scheduler without the inclusion of warm-up steps and enable mixed precision training (fp16) to enhance training efficiency and stability. In the ICL setting, given that context length of LLMs and each essay is relatively lengthy, we choose 0-shot, 1-shot, 2-shot, and 3-shot configurations. For the same reasons, during the training of BERT and RoBERTa models, argumentative essays are divided into two or three parts based on sequence length and paragraph structure as input; while for Longformer and LLMs, the maximum input length was set to 1200 tokens. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.4 Results and Analysis

4.4.1 Experiments of SFT

Tables 4 displays the performance of various models on the argument component detection task under the SFT setting. Our findings are as follows.

Firstly, it is evident that the performance of LLMs far surpasses that of PLMs, both in overall Macro-F_1 and various argument types F_1 scores, indicating the exceptional capability of LLMs in recognizing argument types, especially in handling imbalanced and low-resource data. This is attributed to the rich knowledge and powerful learning ability of LLMs, and it further confirms the scaling laws (Kaplan et al., 2020), that is, larger models will perform better.

Secondly, within the realm of open-source LLMs, Qwen1.5-7B demonstrates the best performance, followed closely by Baichuan2-7B, while ChatGLM3-6B notably falls short of its counterparts. This is primarily due to differences among the models in identifying low-resource categories. The ChatGLM3-6B model fails to recognize all scarce-sample argument types (including *Anecdote*,

[3] https://github.com/QwenLM/Qwen
[4] https://github.com/THUDM/ChatGLM3
[5] https://github.com/baichuan-inc/Baichuan2

Model	Setting	Assertion			Evidence					Elaboration	Others	Macro-F_1	Micro-F_1
		Major Claim	Claim	Restated Claim	Fact	Anecdote	Quotation	Proverb	Axiom				
Baichuan2-turbo	0-*shot*	31.75	15.58	22.22	61.87	23.53	76.60	50.00	22.22	59.04	12.50	37.53	47.40
	1-*shot*	45.27	27.09	42.53	59.90	15.00	68.19	34.52	35.56	71.98	11.11	41.11	60.22
	2-*shot*	28.72	28.92	46.90	63.02	0.00	74.88	57.78	33.33	75.40	21.01	43.00	63.34
	3-*shot*	34.29	31.78	49.28	65.69	0.00	75.00	**66.67**	0.00	**76.40**	**36.36**	**43.65**	**63.90**
Glm-3-turbo	0-*shot*	12.95	27.66	38.10	54.55	28.57	61.54	40.00	20.00	46.77	22.22	35.24	40.12
	1-*shot*	39.95	27.96	28.22	**68.22**	24.34	64.18	11.11	26.30	71.59	11.85	37.37	60.43
	2-*shot*	34.72	17.75	10.56	63.91	11.11	66.39	55.56	44.44	74.79	29.90	40.91	62.44
	3-*shot*	31.75	18.82	14.81	60.87	0.00	71.79	50.00	0.00	72.54	33.33	35.39	60.91
Qwen-turbo	0-*shot*	30.43	24.32	25.32	60.81	**36.36**	62.22	25.00	11.11	24.85	0.00	30.04	32.22
	1-*shot*	29.66	28.46	28.45	61.47	3.70	59.97	38.33	21.30	40.69	0.00	31.20	39.71
	2-*shot*	23.47	30.69	31.90	56.32	6.84	62.14	37.78	**45.08**	44.39	9.52	34.81	40.91
	3-*shot*	16.67	29.07	27.91	47.62	10.53	46.51	40.00	25.00	50.71	0.00	29.40	40.33
GPT-3.5-turbo	0-*shot*	13.16	23.26	13.56	58.38	0.00	61.11	22.22	0.00	31.52	0.00	22.37	32.22
	1-*shot*	22.23	16.93	7.41	50.07	0.00	55.01	32.38	0.00	67.61	0.00	25.16	53.57
	2-*shot*	11.29	20.01	18.97	50.78	16.92	55.56	26.80	0.00	65.52	20.00	28.59	51.49
	3-*shot*	8.51	24.72	19.35	43.75	25.00	54.05	28.57	0.00	68.01	00.00	27.20	53.85
GPT-4-turbo	0-*shot*	38.10	**40.38**	**51.43**	56.93	15.38	**80.95**	33.33	0.00	69.31	19.35	40.52	58.00
	1-*shot*	**55.91**	33.37	51.03	48.72	14.71	76.34	31.19	0.00	74.95	26.51	41.27	61.61
	2-*shot*	50.26	33.47	47.66	55.16	32.48	71.15	38.89	0.00	74.94	31.75	43.58	63.62
	3-*shot*	40.91	29.79	41.51	47.93	0.00	66.67	**66.67**	40.00	72.23	30.77	**43.65**	60.50

Table 5: Performance of various LLMs on the fine-grained argument component detection task in the ICL setting. Displayed are the F_1 scores (%) of each type, with the best results in **bold** and the second best results underlined.

Proverb, *Axiom*, and *Others*), leading to its lagging performance. However, *Axiom* type recognition remains a challenge for all models, reflecting the difficulties of detecting low-sample data within CEAMC. It may require additional domain knowledge or data augmentation methods to enhance model recognition of this argument type.

Finally, within the PLMs, RoBERTa performs best, followed closely by BERT, while Longformer lags far behind the other two. This may be due to the excessive context throughout the text introducing noise and negatively impacting the model's ability to distinguish sentence types. It is noteworthy that the RoBERTa model outperforms ChatGLM3-6B in composite metrics, with its Micro-F_1 even comparable to that of Qwen1.5-7B, which demonstrates the prowess of smaller models in identifying argument types, but also reflects their limitations in identifying low-resource categories.

4.4.2 Experiments of ICL

Table 5 shows the performance of various close-source LLMs on CEAMC under the ICL setting, revealing the following findings.

Firstly, it is apparent that the Baichuan2-turbo achieved the best overall results in the 3-shot setting, demonstrating its outstanding capability in Chinese argumentation. Interesting outcomes have emerged between Chinese and English LLMs in the identification of various argument types. For the recognition of *Major Claim*, *Claim*, and *Restated Claim*, GPT-4-turbo demonstrates outstanding performance, showcasing its strength in capturing conclusive or declarative statements. In contrast, for most evidence types (including *Fact*, *Anecdote*, *Proverb*, and *Axiom*), *Elaboration*, and *Others* argument types, the best results are distributed among Chinese LLMs, signifying their superiority in understanding complex Chinese information and discerning intricate details. These findings not only highlight the differences between Chinese and English LLMs, but also reflect the importance of our CEAMC in the field of Chinese argumentation.

Secondly, in the 0-shot, 1-shot, and 2-shot settings, the overall performance of LLMs progressively improves with the increase of prompt samples, reflecting that input examples can effectively enhance the model's learning in specific task. However, in the 3-shot setting, the models' performance does not improve significantly and may even decline, suggesting that the enhancement of LLMs' performance in the ICL setting is not unlimited, and that excessive examples may introduce additional noise which affects the models' ability to recognize argument types. For the F_1 scores across various argument types, no clear trend emerges, but *Anecdote* in Qwen-turbo, as well as *Claim*, *Restated Claim*, and *Quotation* in GPT-4-turbo reach optimal results with zero-shot learning (specific cases are detailed in Appendix E). This seems to confirm the sensitivity and instability of LLMs in response to prompt samples, and the acquisition of high-quality samples to enhance model performance warrants further exploration.

Finally, comparing Tables 4 and 5, it can be observed that in most cases, the open-source LLMs in

Model	Assertion			Evidence					Elaboration	Others	Macro-F_1	Micro-F_1
	Major Claim	Claim	Restated Claim	Fact	Anecdote	Quotation	Proverb	Axiom				
Baichuan2-turbo	31.75	15.58	22.22	**61.87**	23.53	76.60	**50.00**	22.22	59.04	12.50	37.53	47.40
Baichuan2-turbo$_{CoT}$	3.77	27.27	16.33	28.85	13.33	52.94	33.33	0.00	22.17	5.13	20.31	19.54
Glm-3-turbo	12.95	27.66	38.10	54.55	28.57	61.54	40.00	20.00	46.77	**22.22**	35.24	40.12
Glm-3-turbo$_{CoT}$	13.84	22.99	39.02	29.82	17.39	42.11	0.00	20.00	35.87	10.53	23.16	28.90
Qwen-turbo	30.43	24.32	25.32	60.81	**36.36**	62.22	25.00	11.11	24.85	0.00	30.04	32.22
Qwen-turbo$_{CoT}$	6.11	22.43	19.61	25.23	0.00	17.65	0.00	**28.57**	25.46	0.00	14.51	19.54
GPT-3.5-turbo	13.16	23.26	13.56	58.38	0.00	61.11	22.22	0.00	31.52	0.00	22.37	32.22
GPT-3.5-turbo$_{CoT}$	12.77	22.67	25.93	40.00	0.00	33.33	**50.00**	0.00	23.56	0.00	20.83	21.00
GPT-4-turbo	**38.10**	**40.38**	**51.43**	56.93	15.38	**80.95**	33.33	0.00	**69.31**	19.35	**40.52**	**58.00**
GPT-4-turbo$_{CoT}$	37.68	40.00	44.00	41.07	0.00	72.73	28.57	0.00	50.00	7.19	32.12	40.54

Table 6: Performance of various LLMs on the fine-grained argument component detection task in the CoT setting. Displayed are the F_1 scores (%) of each type, with the best results in **bold** and the second best results underlined.

the SFT setting significantly outperform the closed-source models in the ICL setting, despite the superior foundational capabilities of closed-source models. This highlights the strength of SFT and underscores the importance of data annotation.

4.4.3 Experiments of CoT

In Table 6, we report the performance of various LLMs under the CoT setting. It is clear that the performance significantly drops across most metrics for all LLMs, indicating that the CoT method faces considerable challenges in the task of argument type identification. This seems to suggest that LLMs struggle to mimic the human process of step-by-step argument analysis. Certainly, this is related to the generative nature of LLMs, which often generate explanatory reasons or argument summaries despite being explicitly instructed not to do so, making it difficult to accurately predict the argument type of specific sentence.

To further investigate the impact of CoT and ICL settings, we conduct ablation experiments, the results displayed in Table 10 (see Appendix F). Despite directly utilizing prompt example to guide content output under the CoT method, LLMs still face significant challenges in identifying argument types. Specifically, compared to the CoT setting, the 1-shot-CoT method significantly enhances the performance of LLMs. However, this improvement still falls short of the performance seen in the 1-shot setting and, in some cases, even inferior to the zero-shot results. This may attribute to the nuances of the Chinese language in CEAMC and the inherent complexity of argumentation.

5 Case Study

As shown in Table 11, LLMs have accumulated a considerable amount of common knowledge, demonstrating basic argument analysis capabilities, as seen in sentences **#1** and **#14**. However, this also seems to confirm the biases and hallucination of LLMs, such as in sentence **#18**, a famous *Quotation* by Voltaire, which is most often misclassified as a *Proverb* or *Fact*, attributable to the biases inherent in the pre-training corpora. It is worth noting that LLMs are unable to accurately identify the *Major Claim* and *Claims* in the vast majority of cases, and there are even cases where they are directly classified as *Restated Claim* (sentence 3 under 0-shot setting) and sentences with obvious celebrity quotes are judged as Major Claim (sentence 1 under CoT setting), suggesting that there a significant discrepancy between LLMs' understanding of argumentation and human interpretation.

6 Conclusion

In this paper, we introduce the **C**hinese **E**ssay **A**rgument **M**ining **C**orpus (**CEAMC**), a richly annotated and comprehensive dataset designed to address the limitations in current argument mining research. Our dataset integrates argument mining research with educational practice, encompassing 4 coarse-grained and 10 fine-grained argument types, thereby overcoming the simplicity and monotony of argument types in previous studies. We also conduct several baselines with existing mainstream methods on our dataset, and the results demonstrate the superiority of LLMs, confirming the scaling laws. Further analysis indicates that while LLMs possess basic argument analysis capabilities, their inherent biases and hallucinations limit their developmental potential, also showcasing the significant differences between LLMs' understanding of argumentation and human interpretation. Therefore, how to further unleash LLM's argumentation skills in education and enhance their logical reasoning abilities remains to be explored.

Limitations

The limitations of our corpus 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.

Ethics Statement

All data annotators and expert reviewers have received compensation for their contributions. Additionally, we have obtained explicit consent from the essay authors and their guardians to use the essays for annotation and publication purposes. To protect the privacy of students, all essays in the dataset have been anonymized, ensuring the absence of any personally identifiable information. We express our sincere gratitude for the trust and support extended by all involved parties.

Acknowledgements

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.

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

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.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112. Association for Computational Linguistics.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

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.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021.

Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational linguistics*, 43(1):125–179.

Frederick M Hess and Michael Q McShane. 2014. *Common core meets education reform: What it all means for politics, policy, and the future of schooling*. Teachers College Press.

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*.

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*.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback ii: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3994–4004.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

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.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46.

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.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023a. Entity coreference and co-occurrence aware argument mining from biomedical literature. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 54–60.

Boyang Liu, Viktor Schlegel, Paul Thompson, Riza Theresa Batista-Navarro, and Sophia Ananiadou. 2023b. Global information-aware argument mining based on a top-down multi-turn qa model. *Information Processing & Management*, 60(5):103445.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? *ArXiv*, abs/2404.09696.

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.

Yasser Otiefy and Alaa Alhamzeh. 2024. Exploring large language models in financial argument relation identification. In *FINNLP*.

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)*.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495, Singapore. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Stella Heras, and Ana Garcia. 2023. Automatic debate evaluation with argumentation semantics and natural language argument graph networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6030–6040, Singapore. Association for Computational Linguistics.

Christian Scheibenzuber, Laurentiu-Marian Neagu, Stefan Ruseti, Benedikt Artmann, Carolin Bartsch, Montgomery Kubik, Mihai Dascalu, Stefan Trausan-Matu, and Nicolae Nistor. 2023. Dialog in the echo chamber: Fake news framing predicts emotion, argumentation and dialogic social knowledge building in subsequent online discussions. *Computers in Human Behavior*, 140:107587.

Jiasheng Si, Liu Sun, Deyu Zhou, Jie Ren, and Lin Li. 2022. Biomedical argument mining based on sequential multi-task learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):864–874.

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.

Yefei Teng and Wenhan Chao. 2021. Argumentation-driven evidence association in criminal cases. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2997–3001.

Stephen E Toulmin. 2003. *The uses of argument*. 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Xiutian Zhao, Ke Wang, and Wei Peng. 2023. Orchid: A chinese debate corpus for target-independent stance detection and argumentative dialogue summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9358–9375.

A Annotation Scheme and Samples

Combining previous studies and practical argumentation needs, we define 4 coarse and 10 fine-grained argument types, as shown in Table 7.

B Detailed Annotation Process

Our annotation process was carried out by a team composed of three undergraduates, three postgraduates from linguistics and education fields, and two expert reviewers with experience in Chinese teaching. Before the actual annotation process, the team underwent a training session and pre-annotation to familiarize themselves with the task.

To ensure efficiency and consistency, the data was divided into three groups for annotation. The initial annotation was done by the undergraduate and postgraduate students, while the expert reviewers validated and corrected their work. This process was aimed at maintaining the quality and consistency of the annotations. Furthermore, we organized weekly online discussions to address any common issues that arose during the annotation process. The discussion also served as a platform to make necessary adjustments in the annotation process.

Coarse	Fine-grained	Description	Sample
Assertion	Major Claim	The theme or thesis of an article, i.e., the most significant point that the author aims to convey and argue.	Life needs a sense of ritual because it can counter mediocrity. (生活需要仪式感，因为仪式感可以对抗平庸。)
	Claim	Supporting ideas or subsidiary claims articulated around the major claim.	In my opinion, life needs a sense of ritual, but not blindly pursued. (我认为，生活需要仪式感，却不能盲目追求。)
	Restated Claim	A restatement or rephrasing of an already stated Major Claim or Claim, for the purpose of emphasis or clarification.	Life needs a sense of ritual, but can not blindly pursue, the continuous pursuit and progress, lively and vivid, this is life. (生活需要仪式感，却不能盲目追求，不断追求与进步，生动而又鲜活，这才是生活。)
Evidence	Fact	Specific cases, generalized facts, and reliable historical events, etc.	Regrettably, in today's society, many have fallen into the trap of exaggerating their sense of ritual to fulfill short-lived material satisfactions and the envy of others, leading to chaos in their personal lives. In pursuit of luxury, they spare no expense, ultimately trading for nothing but emptiness and stress. (可惜当下社会，多少人就踩入了这样的误区，为了满足物质条件与他人羡艳时的短暂满足，夸大仪式感，而将自己的生活过得一团乱麻，为了所谓"高奢"而不惜一掷千金，最后换来只是空虚与压力。)
	Anecdote	Experiences from oneself or from friends and family.	And on our own part, we may have let our nerves get in the way of our performance in the exam or put ourselves under a lot of unnecessary stress. (而从我们自身来说，我们可能会因为紧张感而影响了考试的发挥，或让自己承担了很多不必要的压力。)
	Quotation	Citing others' writings, research, ideas or theories	The ground is all sixpence, there is always someone to look up to see the moon. (地上都是六便士，总有人抬头去看月亮。)
	Proverb	Sentences or phrases that are widely circulated among the populace, carrying educational value or reflecting social experience.	Without rules, nothing can be accomplished. (没有规矩，不成方圆。)
	Axiom	Recognized common sense or scientific axioms or laws.	In addition to this, the theoretical knowledge of science has become synonymous with authority in most cases, a simple example, no would argue that 1+1 does not equal 2. (除此之外，科学的理论知识也在大多数情况下成为权威的代名词，一个简单的例子，没有会认为1+1不等于2 。)
Elaboration	-	Explanation, analysis, or discussion of the assertion or evidence, providing detailed clarification or establishing the connection between arguments.	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? (生活需要脚踏实地，可如若总是一味低头苦赚那微小的"六便士"，而忘却抬头欣赏那皎洁的"月亮"，只是在平庸中麻木了自我，成为行尸走肉，生活又有什么意义？)
Others	-	None of the above.	May the wind guide our path. (愿风指引我们的道路。)

Table 7: A list of argument types, their descriptions and samples.

The entire process spanned three months, during which a total of 226 argumentative essays were annotated. This structured approach ensured a streamlined annotation process, resulting in a richly annotated corpus that can facilitate subsequent language model training and research.

C Inter-Annotator Agreement (IAA) Calculation

Our annotation team was divided into three groups, and Table 8 shows the IAA scores of different annotation groups and the average result.

Group	Cohen's kappa
1	72.71
2	77.80
3	76.35
Avg.	75.62

Table 8: Consistency analysis results showing the inter-annotator agreement (IAA) scores (in percentage) across different groups. The last row shows the average IAA scores for all groups.

D Prompt Template

D.1 ICL Prompt

In the argument component detection task, we employ both zero-shot and few-shot learning strategies. Figure 3 illustrates the prompts for the 0-shot and 1-shot settings. For the 2-shot and 3-shot prompt settings, please refer to the 1-shot example. For the essay content (i.e., [CONTENT]) in the prompt, we segment the essays into sentences and numbered them.

D.2 CoT Prompt

In the argument component detection task, we explore the impact of CoT strategy on the performance of LLMs, and Figure 4 illustrates the prompt we used.

E Cases of ICL

As shown in Table 9, case studies of Qwen-turbo and GPT-4-turbo in 0-shot and 3-shot settings. Each example corresponds to different argumentative essay, where **#id** indicates the sentence number, which is retained directly from its numbering in the respective essays.

F Comparison of ICL and CoT

For the comparative results of LLMs under ICL and CoT settings, please refer to Table 10. Note that here we only report the overall performance, i.e., the Macro-F_1 and Micro-F_1 scores.

Model	Method	Macro-F_1	Micro-F_1
Baichuan2-turbo	0-shot	37.53	47.40
	1-shot	**41.11**	**60.22**
	CoT	20.31	19.54
	1-shot-CoT	39.59	45.11
Glm-3-turbo	0-shot	35.24	40.12
	1-shot	**37.37**	**60.43**
	CoT	23.16	28.90
	1-shot-CoT	35.01	46.57
Qwen-turbo	0-shot	30.04	32.22
	1-shot	**31.20**	**39.71**
	CoT	14.51	19.54
	1-shot-CoT	28.19	38.53
GPT-3.5-turbo	0-shot	22.37	32.22
	1-shot	25.16	**53.57**
	CoT	20.83	21.00
	1-shot-CoT	**26.56**	48.23
GPT-4-turbo	0-shot	40.52	58.00
	1-shot	**41.27**	**61.61**
	CoT	32.12	40.54
	1-shot-CoT	38.94	59.25

Table 10: Comparison of various LLMs using ICL and CoT methods on CEAMC. In each section, the best results are highlighted in **bold**, and the overall best results are underlined.

G Details of the Case Study

Table 11 presents a case study on the argumentative essay *Do Not Let Your Mind Become a Racetrack*, which consists of 22 sentences. Considering the text length and data presentation, we focus on reporting the key sections.

Model	Type	Setting	Prompt Samples	Input Content	Output
Qwen-turbo	Anecdote	3-shot	**#7** People should have long-term plans, but we don't know how the future will be. Four years ago when I first started junior high school, I clamoured to take the exams of four schools and eight universities, full of passion and enthusiasm, but now I just want to live my life by the book, too. **#8** When I was a child, I used my hand as a gun and pulled the trigger towards the air, the bullets flew to nowhere, and in the summer when I was sixteen years old, when I turned back, I was hit right in the centre of my eyebrow. (**#7** 人要有长远的打算，但我们并不知道未来如何，四年前刚上初中的我叫嚣着考四校八大，充满激情与热忱，可如今我也只想按部就班过生活。**#8** 童年时将手作枪，朝着空气扣动板机，子弹不知飞去哪里，而在我十六岁那年的盛夏，回头时正中眉心。)	**#10** When others are immersed in the uncertainty and tension of a failed exam, you feel the relaxed atmosphere wrapped in the "breeze on the river and the bright moon in the mountains" and adjust your mindset to better face the next exam. **#11** When others find it difficult to sleep due to nervousness, you are able to sleep and rest properly so that you can be more active in the days ahead. (**#10** 当别人沉浸于考试失利时的无措与紧张时，你的感受"江上清风与山间明月"裹挟而来的松弛氛围，进而调整心态更好的面对下一次考试。**#11** 当他人因紧张而难以入眠时，你进入梦乡休息得当从而能更积极地奔入未来的日子。)	**#10** Anecdote **#11** Claim
			0-shot		**#10** Anecdote **#11** Anecdote
GPT-4-turbo	Quotation	3-shot	**#12** Confucius once said, "With a simple bowl of food and a gourd of drink, even in a humble lane, one can be free from sorrow if content; Yanhui would not change his joy." **#13** Liu Yuxi's notion that "a humble room is not meager" also influenced many generations to come. (**#12** 孔子曾说："一箪食，一瓢饮，在陋巷，人亦不堪其忧，回也不改其乐。" **#13** 而刘禹锡的"陋室不陋"也影响了很多后人。)	**#14** There's a quote from the People's Daily: "Stopping to rest is the best way to move forward. (**#14** 人民日报中有样一段话：停下休息是为了更好的前进。)	**#14** Elaboration
		0-shot	-		**#14** Quotation

Table 9: Cases of Qwen-turbo and GPT-4-turbo in 0-shot and 3-shot settings. **Type** indicates the argument type of the selected case. **Prompt Samples** indicates the sentences in the prompt instances that are consistent with the target output type. **Input Content** indicates the content of the case sentence.

ICL Prompt (0-shot)

Chinese Prompt:

你是教学经验丰富的高中语文教师，请你对下面这篇议论文进行解析，并对其中每个句子的论辩类型进行判断。
论辩类型包含中心论点、分论点、重申论点、事例史实、个人事例、名人名言、谚语俗语、公理规律、阐述和其他。
请注意，只输出句子编号“#id”和对应的论辩类型即可。
作文题目为[TITLE]，作文内容为[CONTENT]，输出结果为：

English Prompt:

You are an experienced high school Chinese language teacher. Please analyze the following argumentative essay and determine the argument type for each sentence.
The argument types include Major Claim, Claim, Restated Claim, Fact, Anecdote, Quotation, Proverb, Axiom, Elaboration, and Others.
Note: Only output the sentence number "#id" and the corresponding argument type.
Essay Title: [TITLE], Essay Content: [CONTENT],
Output:

ICL Prompt (1-shot)

Chinese Prompt:

你是教学经验丰富的高中语文教师，请你对下面这篇议论文进行解析，并对其中每个句子的论辩类型进行判断。
论辩类型包含中心论点、分论点、重申论点、事例史实、个人事例、名人名言、谚语俗语、公理规律、阐述和其他。
请注意，只输出句子编号“#id”和对应的论辩类型即可。
以下为一个示例：
作文题目为[TITLE]，作文内容为[CONTENT]，输出结果为[OUTPUT]
作文题目为[TITLE]，作文内容为[CONTENT]，输出结果为：

English Prompt:

You are an experienced high school Chinese language teacher. Please analyze the following argumentative essay and determine the argument type for each sentence.
The argument types include Major Claim, Claim, Restated Claim, Fact, Anecdote, Quotation, Proverb, Axiom, Elaboration, and Others.
Note: Only output the sentence number "#id" and the corresponding argument type.
Essay Title: [TITLE], Essay Content: [CONTENT],
Output: [OUTPUT]
Essay Title: [TITLE], Essay Content: [CONTENT],
Output:

Figure 3: The prompts under the ICL setting, include Chinese prompts and corresponding English translations.

CoT Prompt

Chinese Prompt:

你是教学经验丰富的高中语文教师，请你对下面这篇议论文进行解析，并对其中每个句子的论辩类型进行判断。
论辩类型包含中心论点、分论点、重申论点、事例史实、个人事例、名人名言、谚语俗语、公理规律、阐述和其他。
请逐步完成这个任务：第一步，找出议论文的中心论点句子。第二步，找出议论文的分论点句子。第三步，找出议论文的论据句子，按照事例史实、个人事例、名人名言、谚语俗语和公理规律的顺序依次输出。第四步，找出议论文中的阐述句子。第五步，找出议论文中的重申论点句子。第六步，找出议论文中非上述类型的句子，即其他类型的句子。
请注意，只输出句子编号“#id”和对应的论辩类型即可。
作文题目为[TITLE]，作文内容为[CONTENT]，输出结果为：

English Prompt:

You are an experienced high school Chinese language teacher. Please analyze the following argumentative essay and determine the argument type for each sentence.
The argument types include Major Claim, Claim, Restated Claim, Fact, Anecdote, Quotation, Proverb, Axiom, Elaboration, and Others.
Please complete this task step-by-step: step1, identify the Major Claim sentence. Step 2, identify the Claim sentences. Step 3, identify the evidence sentences and output them in the order of Fact, Anecdote, Quotation, Proverb, and Axiom. Step 4, identify the Elaboration sentences. Step 5, identify the Restated Claim sentences. Step 6, identify the sentences that are not of the above types, i.e., Others type of sentences.
Note: Only output the sentence number "#id" and the corresponding argument type.
Essay Title: [TITLE], Essay Content: [CONTENT],
Output:

Figure 4: The prompt under the CoT setting, include Chinese prompt and corresponding English translation.

Sents	SFT	0-shot	3-shot	CoT	C-1s	Human
#1 Schopenhauer once said, "Do not let yourself become a racetrack for the thoughts of others." (**#1**叔本华曾经说过：“别让自己成为别人思想的跑马场。”)	Quotation	Quotation	Quotation	Major Claim	Quotation	Quotation
#2 We all know not to rely solely on one side of a story, but when the speaker holds a special status, like an ancient sage or an expert, we often lose our footing and blindly believe. (**#2**我们都知道不可偏听偏信，但一旦对方有特殊身份的加持，如古人、专家等，我们便会乱了阵脚，盲目听信。)	Elaboration	Elaboration	Elaboration	Claim	Elaboration	Elaboration
#3 Are the sayings of the ancients, authorities, or books always correct? I think not. (**#3**古人、权威、书本所言便一定正确吗？我看未必。)	Elaboration	Restated Claim	Elaboration	Claim	Claim	Elaboration
#8 No wonder his theories were eventually refuted. (**#8** 也难怪会被推翻了。)	Elaboration	Restated Claim	Elaboration	Elaboration	Elaboration	Elaboration
#9 Authorities and books are the same in this respect. (**#9** 权威、书本亦是如此。)	Elaboration	Restated Claim	Elaboration	Elaboration	Fact	Elaboration
#10 Many self-proclaimed experts online post entirely inappropriate views, leading many to jokingly refer to experts as "brick experts"; there are good books and bad books, otherwise, why would there be so many banned books? (**#10** 网络上许多人自诩专家，发表一些完全不合适的观点，让许多人把专家笑称为“砖家”；书有好书，也有坏书，不然为何会有如此多的禁书？)	Fact	Elaboration	Elaboration	Restated Claim	Claim	Fact
#11 Therefore, even the words of the ancients, authorities, and books should be scrutinized for authenticity. (**#11** 因此，哪怕是古人、权威、书本所言，我们也应学会辨别真伪。)	Major Claim	Restated Claim	Elaboration	Claim	Claim	Claim
#12 If we blindly follow because "it has always been so," "the books say so," or "most people think," it can lead to serious and irreversible mistakes. (**#12** 若偏听偏信，就因为“自古以来”“书上说”“大多数人认为”便盲目跟从，会引起严重的、不可挽回的错误。)	Elaboration	Elaboration	Elaboration	Claim	Axiom	Elaboration
#13 Sunshine boy Liu Xuezhou faced life positively, and the misfortunes of his childhood did not dampen his enthusiasm for life, yet he was driven to end his life by the cold and cruel comments on the internet. (**#13** 阳光少年刘学洲，积极面对生活，童年生活的不幸没有打消他对生活的热忱，却被网络上冰冷残忍的字句中伤，选择了结生命。)	Fact	Anecdote	Fact	Anecdote	Anecdote	Fact
#14 A kind word can warm three winter months, while harsh words can chill someone deeper than the cold of June. (**#14** 良言一句三冬暖，恶语伤人六月寒。)	Proverb	Proverb	Proverb	Proverb	Proverb	Proverb
#15 Some people find pleasure in spreading rumors, and unfortunately, gossiping is a major interest for many, thus making false information increasingly exaggerated to the point of disbelief. (**#15** 有些人喜欢把造谣当作乐趣，更不幸的是，讨论八卦是大多人的兴趣点所在，于是虚假事情愈演愈烈，发展到让人纯望的地步。)	Fact	Elaboration	Elaboration	Elaboration	Elaboration	Fact
#18 No snowflake in an avalanche ever feels responsible. (**#18** 雪崩时，没有一片雪花是无辜的。)	Proverb	Proverb	Proverb	Proverb	Fact	Quotation
#19 We must remember that speaking and acting cautiously is the mark of a gentleman. (**#19**我们要牢记，谨言慎行才是君子作风。)	Elaboration	Restated Claim	Elaboration	Restated Claim	Quotation	Claim
#20 Do not let yourself become a racetrack for the thoughts of others, manipulated and trampled upon without even knowing. (**#20**别让自己成为别人思想的跑马场，任人摆弄践踏却仍不自知。)	Restated Claim	Restated Claim	Restated Claim	Restated Claim	Restated Claim	Claim
#22 Do not become a racetrack, do not follow the crowd, do not become a sharp blade, bloom under the sunlight. (**#22** 勿成跑马场，勿成从众者，勿成利刃，盛放在阳光下。)	Others	Restated Claim	Restated Claim	Elaboration	-	Major Claim

Table 11: Case study on the argumentative essay *Do Not Let Your Mind Become a Racetrack*. Texts highlighted in red indicate incorrect judgement. **C-1s** denotes the CoT-1-shot setting.