Towards Computational Argumentation: A Comprehensive Chinese Debate Dataset

Anonymous ACL submission

Abstract

Computational argumentation, encompassing argument mining and generation, is attracting increasing attention. However, the lack of comprehensively annotated non-English oral debate datasets hinders progress in this field. Additionally, existing debate datasets neglect some important labels for argument mining, generation, and evaluation. To address these gaps, we introduce a comprehensive Chinese debate dataset named HARMONIC. Compared to previous datasets, HARMONIC includes the basic labels of computational argumentation (claim, stance, evidence), as well as 5 additional crucial labels: rhetorical figures, debater roles, modal words, utterance time, and debate results. Moreover, HARMONIC offers complete utterances of each debate, including speeches of the Pro and Con sides. Thus, our dataset supports not only common argument mining and generation tasks, but also provides resources for rhetorical figure detection, argument quality evaluation, and debate result prediction. This dataset is derived from transcripts of classical Chinese debate competitions covering 318 topics. More than 70,000 utterances have been manually annotated and thoroughly checked. Besides providing a dataset for research, we conduct experiments on common computational argument tasks and a novel task (rhetorical figure detection), in which LLMs are also evaluated. The experimental results highlight the challenging nature of the dataset. Our corpus will be made publicly available.

1 Introduction

Computational argumentation aims to identify argument components, analyze the relationships among them, and generate arguments automatically (Cabrio and Villata, 2018; Lin et al., 2023b). This approach plays a promising role in various fields, such as linguistics, law, and education. Computational argumentation mainly includes argument mining and argument generation. Argument mining (AM) aims to identify and extract arguments automatically. Its sub-tasks include claim extraction (Levy et al., 2014; Aharoni et al., 2014), stance extraction (Bar-Haim et al., 2017; Cheng et al., 2022; Hardalov et al., 2022; Zhao et al., 2023), evidence extraction (Rinott et al., 2015; Singh et al., 2019) and classification (Afrin et al., 2020). Argument generation automatically generates a statement that either supports or opposes the stance of the original argument (Toulmin, 2003; Schiller et al., 2021; Lin et al., 2023b; Hua et al., 2019).

The distinctive characteristics of debate dialogues pose significant challenges in argument mining and generation. First, the intricate logical relationships, emotional inclinations, and persuasive styles within the discourse significantly complicate this task. Second, the inherent noise in oral debates, characterized by verbal fillers, repetitions and grammatical errors, further complicates mining tasks. Third, dynamically mining critical points for argument generation and generating rebuttals to opponents' arguments requires robust logical reasoning and the adept use of rhetorical figures. This process also demands a keen emotional awareness, distinguishing it from the more straightforward tasks in other domains.

Researchers have devoted considerable effort to both argument mining and generation. Some studies focus on paragraph-level counter-argument generation (Hua and Wang, 2018; Alshomary et al., 2021b; Alshomary and Wachsmuth, 2023). Similarly, some researchers study sentence-level counter-argument generation(Lin et al., 2023b). Additionally, generating argumentative dialogue and summarization has also been explored (Le et al., 2018; Zhao et al., 2023).

Despite significant advancements in relevant tasks, the progress of research has been impeded by the limited availability of debate datasets. While the English language domain has seen a substantial

Datasets	Dataset Properties											
	Topic	Claims	Pro/Con Stance	Evidence with Multi-Types	Rhetoric Figures	Debate Results	Modal Words	Debater Role & Time	Spoken Lg.	Lg.	Content Type	Modal
CDCD	✓	✓	✗	✗	✗	-	✗	-	✗	En	Wikipedia articles	text
CDED	✓	✓	✗	✓	✗	-	✗	-	✗	En	Wikipedia articles	text
IBM Evidence Type	✓	✓	✗	✗	✗	-	✗	-	✗	En	Wikipedia articles	text
Claim Polarity Dataset	✓	✓	✓	✗	✗	-	✗	-	✗	En	Wikipedia articles	text
IAM	✓	✓	✓	✗	✗	-	✗	-	✗	En	Wikipedia articles	text
ORCHID	✓	✓	✓	✗	✗	✗	✗	✗	✓	Zh	real-world debate	text
IBM 2019	✓	✓	✗	✗	✗	✗	✗	✗	✓	En	real-world debate	text and audio
HARMONIC (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	Zh	real-world debate	text and audio

Table 1: Comparison among argument mining datasets, where "Claims" and "Evidence(Premise)" represent basic argument mining elements, "Rhetorical figures" stands for essential persuasive strategies in the debating process, "Debate Results" represents the result of each debate competition (Pro/Con), "Dialogue Role" indicates the speakers' fine-grained role (e.g., "Speaker 1" for the pro side, while "Speaker 4" for the con side, "Host" means the moderator). Spoken Language: when checked, it indicates spoken language; when crossed, it indicates written language. Lg. denotes language: En for English and Zh for Chinese. Some works are marked with "-" because their data sources do not originate from debate competitions, hence lacking the items.

Datasets	Dataset Properties								
	Argument Generation	Argument Summarization	Counter Speech Generation	Debate Results	Dialogue Role & Time	Spoken Lg.	Lg.	Content Type	Modal
CounterArguGen	✓	✗	✗	-	-	✗	En	online forum	text
ConcluGen	✗	✓	✗	-	-	✗	En	online forum and debate corpora	text
DebateSum	✗	✓	✗	✗	✗	✗	En	debate corpora	text
CounterSpeechGeneraion	✗	✗	✓	-	-	✓	En	real-world debate	text
ORCHID	✗	✓	✗	✗	✗	✓	Zh	real-world debate	text
HARMONIC (Ours)	✓	✗	✓	✓	✓	✓	Zh	real-world debate	text and audio

Table 2: Comparison among argument generation datasets. The meanings of the dataset properties are the same as Table 1.

number of argument mining datasets (Levy et al., 2014; Rinott et al., 2015; Aharoni et al., 2014; Bar-Haim et al., 2017; Cheng et al., 2022), corpora for Chinese computational argumentation remain in their infancy. Only (Zhao et al., 2023) proposes a Chinese debate dataset for stance detection and argumentative dialogue summarization. Moreover, existing debate datasets overlook some important labels for argument mining, generating, and evaluation, such as rhetorical figures, modal words, utterance time and debate results. Thus, there is a clear need for more Chinese debate benchmarks with fine-grained labels.

To this end, we present a comprehensive oral Chinese debate dataset HARMONIC (**H**arnessing **A**rgumentation and **R**hetoric in **M**andarin: **O**bserve, **N**otate, **I**nteract, **C**omprehensive) with more fine-grained labels to advance research. We compare the proposed dataset HARMONIC with previous debate datasets in Table 1 and Table 2 based on their task applicability. As shown in Table 1, unlike previous argument mining datasets, HARMONIC includes rhetorical figures, debate results, modal words, debater role, and utterance time. In addition, HARMONIC includes annotations for the topic, claim, pro and con-side stance, and evidence. HARMONIC is a multi-modal dataset that includes both text and audio from real-world debate. Table 2 presents a comparison of argument generation datasets. HARMONIC offers more labels and supports a wider range of argument generation tasks.

To our knowledge, HARMONIC is the first comprehensive Chinese debate dataset for argument mining and generation. It also supports the detection of rhetorical figures in debates. The annotated debate results can be used to assess argument quality. HARMONIC contains 40K annotated arguments and 23K rhetorical figure utterances in Mandarin. We use an automatic speech recognition tool to transcribe original videos, followed by manual annotation and verification. HARMONIC is available in multiple formats, tailored for different tasks. To better study debates, we we conducted experiments on common computational argument tasks and a novel task—rhetorical figure detection. The results reveal the challenges of our proposed dataset and task.

In summary, our contributions are threefold: (1) We construct HARMONIC, the first comprehensive oral Chinese debate dataset that supports both argument mining and generation; (2) HARMONIC

includes 5 additional crucial labels and can support more computational argument tasks; (3) We conduct comprehensive experiments on common computational argument tasks and a novel task (rhetorical figure detection) on HARMONIC, in which LLMs are also evaluated.

2 Related Work

Argument Mining and Related Datasets Argument mining facilitates the understanding of argument structures and argumentation strategies. Various tasks have been explored. (Levy et al., 2014) introduced a dataset annotated with claims and proposed context-dependent claim detection (CDCD). (Bar-Haim et al., 2017; Chen et al., 2019) further annotated claim stances and addressed context-relevant claim stance classification (CSC). (Rinott et al., 2015) proposed context-dependent evidence detection (CDED) and a dataset. Furthermore, (Cheng et al., 2022) introduced a dataset from online forums. (Stab and Gurevych, 2014) annotated argument components and relationships in essays, creating a dataset with 402 articles. Additionally, (Peldszus, 2014) developed an argumentation dataset comprising short texts.

Argument Generation and Related Datasets (Peldszus and Stede, 2016) released an annotated corpus of argumentative microtexts. (Hua and Wang, 2018; Hua et al., 2019) created datasets for persuasive argument generation on Reddit ChangeMyView. The dataset by (Alshomary et al., 2022) contains argumentative texts and corresponding ethical considerations. (Zhao et al., 2023) proposed a dataset for argumentative dialogue summarization. (Ruggeri et al., 2023) proposed a dataset of argumentative dialogues on scientific papers. (Lin et al., 2023b) presented a dataset for sentence-level counter-argument task. (Roush and Balaji, 2020) proposed an argument mining and summarization dataset. (Syed et al., 2021) introduced the task of generating informative conclusions from argumentative text. (Ruggeri et al., 2023) proposed a dataset of argumentative dialogues on scientific papers.

Rhetorical Figures and Related Datasets (Lawrence et al., 2017) explored the utility of rhetorical figures for argument mining. (Chen et al., 2021) created a Chinese dataset for joint rhetoric and emotion identification problems. (Bothwell et al., 2023) proposed a rhetorical parallelism detection task and provides a dataset in Latin and Chinese. Moreover, previous research often emphasizes the literary aspect of figures, (Zhu et al., 2022) built a Chinese corpus from novels and proses for figures recognition. We have summarized and compared them in Table 15.

Publicly available argumentation datasets with rhetorical annotations are scarce, especially for non-English languages. (Lawrence et al., 2017) created their corpus from BBC radio program. The study aims to identify English rhetoric, such as anadiplosis, epanaphora. In contrast, HARMONIC is an annotated rhetorical figure dataset, with Chinese linguistic characteristics used to categorize rhetorical figures. As a spoken corpus, HARMONIC serves as a valuable and rare supplement to Chinese linguistic resources.

Counter Speech Datasets Argument mining from speech has been previously explored. (Mirkin et al., 2018) presented a task for machine listening comprehension in argumentation. (Lavee et al., 2019) proposed a corpus containing speeches associated with their corresponding audio files and include transcribed texts. (Chen et al., 2023) propose a counter speech generation task, which requires understanding the argumentative structures in the supporting speech and generating a counter-speech that opposes the proposition. To our knowledge, no Chinese dataset is currently available for this task.

3 Dataset Creation

3.1 Data Collection

We observed that most Chinese debate competitions do not have official texts but are presented in video recordings. Therefore, we use classical Chinese debate videos as primary sources. We first collected 980 Chinese classical debate videos from major video platforms where the organizers regularly release competitions (see Table 17 in **Appendix D** for details). The videos consist of standard debate conversations in Chinese Mandarin. After that, we filtered out unclear audio, hate speech, and political speechs through both automated tools and manual review. Those with non-standard Chinese debate formats are also abandoned. Then, we use the automatic speech recognition (ASR) technique to obtain raw transcription. Audio recordings are transcribed into text. The original audio is a reference for annotators to label speaker roles in subsequent steps. Following a meticulous filtering process, we retain 600 high-quality debates for further annotation.

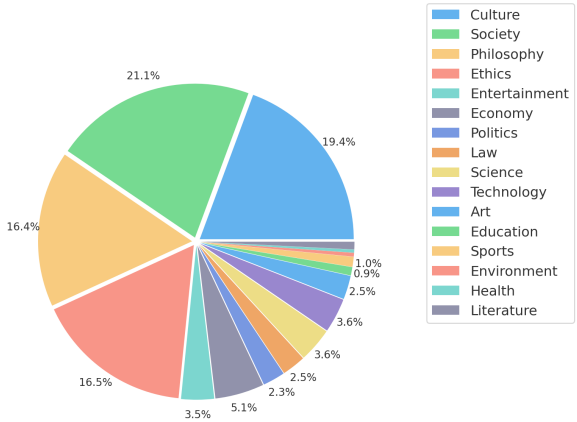

Figure 1: The situation of topic types in the HARMONIC. Competitions are divided into a total of 16 domains. Each competition may encompass multiple domains.

3.2 ASR Transcription and Filtering

We utilize the commercialized automatic speech recognition (ASR) system to obtain the initial machine transcriptions. Then, annotators manually correct the transcripts for lexical errors or incorrect punctuation. First, we use speech transcription tools on audio files to obtain the raw transcription texts. Next, we remove irrelevant content, such as accidental interruptions and advertisements. For the debates whose speakers are not given the camera, made it difficult to distinguish the debater role. We discard these debates. For the selected debates, all speakers' roles are marked, such as hosts, debaters, and judges (from A1 to A4, B1 to B4; A represents the pro side, while B represents the con side). Introductions and topic statements from the debate competitions were retained for further use. The annotators also manually checked the ASR transcription errors. They correct the disfluent content, grammar errors, missing or misplaced characters, and punctuation errors. Finally, we obtained 600 high-quality transcribed texts.

3.3 Manual Annotation

After ASR Transcription, we annotated the labels for the debate texts. All the annotators are well-trained. They extracted and categorized topics assigned participant roles based on the original video, and conducted comprehensive annotation and review, including claims, stances, evidence, rhetorical figurative units, modal words, debate results, and timestamps for each utterance. In addition to creating datasets in suitable formats for various tasks, we provide audio files from debate competitions for research in audio argument mining. See **Appendix A** for more annotation details and **Appendix B** for examples of labels. The annotation process is as follows.

- Training and Q&A Session: Annotators participated in a three-day training and Q&A session to ensure their understanding and practical application of the annotation manual.

- Topic Extraction: Annotators extracted and classified topics. For example, "Is AI-generated art considered art?" is categorized under Art & Technology, as illustrated in Figure 1.

- Role Annotation: Annotators assigned participant roles to the speakers based on the original videos.

- Comprehensive Annotation: Annotators annotate various elements such as claims, stances for claims, evidence, pro/con speeches, rhetorical figure types, modal words, debate results, and timestamps.

- Post-Annotation Check: A thorough review and correction process is conducted to ensure accuracy and consistency.

We established elaborate annotation guidelines to ensure high quality. To ensure consistency, annotators underwent a three-day training course, including demonstrations, online documentation, video records, and Q&A sessions. They followed guidelines to extract topics, assign roles, and annotate various aspects of the debates. The accuracy and consistency of the annotations were thoroughly verified.

We built an annotation platform that supports span and relation labels to facilitate the annotation process. Annotators select labels for claims, stances, evidence, and rhetoric and establish relationships between arguments.

The annotation process consists of two stages. The first involves recruiting 20 native Chinese speakers holding at least a bachelor's degree. They follow the established guidelines to annotate several examples.

In the second stage, ten individuals who demonstrated superior annotation effectiveness and efficiency were retained for the formal annotation phase. Annotation and manual evaluation were

conducted, with the authors serving as senior annotators. The formal phase involves two rounds, with two annotators jointly completing each debate. Manual evaluation addressed label accuracy, span reasonableness, and resolved disagreements. Below are the explanations of different annotation elements.

Topics and Topic Types We use video titles to extract debate topics. As shown in Figure 1, we manually categorize the topics into 16 types.

Debater Role and Utterance Time Utilizing automatic speech recognition, speeches are automatically segmented according to roles. However, some speakers' utterances or roles do not match the truth. Therefore, we conduct manual verification to ensure consistency with the facts. Annotators will correct any missed or incorrectly labeled utterances. Also, we ensure that the manually labeled speaking role for each utterance segment is correct. Each role is accompanied by a labeled timestamp.

Rhetorical Figures We categorize Chinese rhetorical figures in our corpus into nine types: such as Metaphor/Simile, Hyperbole, Contrast, etc. See **Appendix B** for more details.

Dialogues Corpus We also provide complete debate dialogues with role labels and speech timestamps, which can be used to generate discourse-level arguments.

Claim and Stance We manually label claims by detecting the claims given the topics. Afterward, for each claim, we label the stance by identifying the position of the extracted claim about the specified topic.

Evidence Types Inspired by (Guo et al., 2023), evidence types in the dataset are categorized into Case, Expert, Research, Explanation, and Others. It ensures a comprehensive analysis of evidence in debates. We manually label the evidence with types given the claims.

Debate Results We annotated each debate's winning side (Pro, Con, or Draw), with most competitions having a clear victor. It is helpful for further research.

Chinese Modal Words We also labeled Chinese modal words of debates to facilitate the understanding of nuances in tone, certainty, and attitude.

Overall, our manual annotation follows a well-defined process, with trained annotators adhering to guidelines and undergoing thorough checks to ensure accuracy, consistency, and quality.

3.4 Quality Control

The annotation is based on factual grounding, as the accuracy of topic extraction, speaker's role (Pro First Speaker, Cons Third Speaker, Moderator, etc.), and utterance labeling can be confirmed by referencing the original videos, ensuring a consensus on the veracity. To ensure dataset quality, we applied specific criteria for filtering out videos. Inappropriate speeches were identified and addressed using content risk control tools, including hate speech and irrational political discourse. Additionally, the incomplete debates were excluded. Poor-quality audios that could lead to incorrect recognition were also removed. Inaudible content was eliminated as manual post-correction was not feasible for such recordings. Sessions deviated from standard Chinese debate formats and were abandoned. To maintain a high-quality corpus, we focused on exemplary sessions from official competitions, especially those from the semifinal and final rounds. Eventually, we retained 600 debates out of the initially collected 980 videos.

The annotation process was divided into two rounds and assigned to two independent groups. Each consists of 5 annotators. Consensus was reached among all members within two groups for all instances. Five annotators were randomly selected to review the correctness and overall consistency of post-corrected translations meticulously. Cohen's Kappa coefficient (Cohen, 1960) (Cohen, 1960) was calculated to assess the inter-annotator agreement between two annotation groups. We obtained a Kappa coefficient close to 0.7, indicating a substantial level of agreement. After a thorough review, consensus was reached by both groups, and senior annotators resolved any discrepancies. See **Appendix C** for more details.

3.5 Dataset Overview

We conducted a statistical analysis of this dataset. Table 3 presents an overview of the statistics for HARMONIC. We have a total of 600 debates, with over 250K sentences. The average length of each debate is approximately 20,158 sentences. Additionally, we split the dataset into train/dev/test sets, as shown in Table 4. We generally divide the dataset into an 8:1:1 ratio. Among them, there are over 60K evidence sentences and more than 20K rhetorical figures sentences. More statistical information can be found in **Table 16** in Appendix B. Compared to previous argument datasets, we pro-

pose a more comprehensive Chinese debate dataset with more fine-grained labels. This dataset can support the research of computational argumentation in Chinese.

	Ours
Total debates	600
Total unique topics	318
Total sentences	251,726
Total claim sentences	15,359
Total evidence sentences	32,173
Total rhetorical figure sentences	23,260
Avg. debate length	20,158.2
Avg. sentences per debate	419.5
Avg. length of sentences	48
Avg. length of claim sentences	77.4
Avg. length of evidence sentences	72.4
Avg. length of rhetorical figure sentences	37.2

Table 3: Overview statistics for HARMONIC. Lengths of sentences are measured in Chinese characters.

Category	Train	Dev	Test	Total
competitions	480	60	60	600
sents	202,550	23,457	25,719	251,726
claims	5,719	554	635	6,908
evidence	32,173	25,394	3,133	60,700
rhetoric sentences	15,964	2,092	2,070	20,126
support claims	9,617	1,106	1,368	12,091
contest claims	2,637	285	313	3,235
Pro utterances	-	-	-	40,929
Con utterances	-	-	-	41,066
Host utterances	-	-	-	11,868

Table 4: Dataset statistics split into train/dev/test sets.

4 Experiment

4.1 Tasks

To evaluate the existing models on the proposed HARMONIC, we conduct experiments on the following computational argumentation tasks.

Rhetorical Figure Detection We are the first to evaluate rhetorical figure detection on debate dataset. Given a sequence of words that forms a figurative unit (Zhu et al., 2022) U, consisting of n words $\{w_i\}_{i=1}^{n}$, along with the contextual words C that surround it, comprising m words $\{w_j\}_{j=1}^{m}$, our goal is to categorize the expression U into its corresponding rhetorical figure. We define nine Chinese rhetorical figures in our dataset as mentioned in **Appendix B**. Moreover, eight examples are shown in Table 12. Detecting rhetorical figures benefits debates and computational argumentation by enhancing argument quality and response persuasiveness. It improves methods' ability to generate persuasive texts, aids sophisticated argumentation mining, and supports education by enhancing writing and argumentative skills.

Claim Extraction Following (Levy et al., 2014), our task is defined as: Let sentence T be a debate topic, and let $A = \{a_i\}_{i=1}^{m}$ be a set of associated articles, where each article a_i consists of multiple sentences. The goal is to automatically extract claims $\{c_j\}_{j=1}^{n}$ from these articles, where each claim c_j is a sentence. It is a central task in argument mining, essential for identifying and analyzing the main arguments within a debate.

Stance Classification As outlined by (Bar-Haim et al., 2017), let T be a given topic, and let $\{c_j\}_{j=1}^{n}$ be a set of claims extracted for this topic. Stance s_j represents a position toward the controversial topic T and can take the values $\{\text{support}, \text{oppose}, \text{neutral}\}$. The task is to determine whether each claim c_j supports, opposes, or remains neutral toward the topic T, which aims to assign a stance s_j to each claim c_j. Table 4 demonstrates the distribution of claims across two stances (support claims, contest claims), with 66 % supporting and 34 % contesting the topic.

Evidence Classification We define the task as Let T be a concrete topic with an implicit claim and let $E = \{e_i\}_{i=1}^{m}$ be a set of candidate evidence. The task is to identify the relevant evidence $\{e_j\}_{j=1}^{n}$ and determine its type for the given topic T. We focus on candidate evidence sentences near the claims, which our annotators manually annotate. Inspired by (Guo et al., 2023)'s work, we categorize the annotated evidence types into five distinct classes: Case, Explanation, Research, Expert, and Others. These categories are used in a multi-class classification task.

Counter Speech Generation This task involves generating responsive or opposing speeches in reaction to supporting speeches, aiming to evaluate large language models' ability to comprehend and generate counter-arguments (Chen et al., 2023).

4.2 Experimental Setting

We randomly divided our dataset into training, development, and testing sets using a ratio of 8:1:1. We treat rhetorical figure detection, claim extraction, stance classification, and evidence classification as sentence classification tasks. We consider

counter speech generation as a text generation task. For the former, we conducted experiments using pre-trained models and LLMs; for the latter, we used LLMs for our experiments. See **appendix E** for the prompt templates used in all the above experiments. We use negative sampling to reduce data imbalance.

Rhetorical Figure Detection We experimented with pre-trained models and LLMs. For traditional models, we experiment on BERT, RoBERTa, and Longformer with the training set of this dataset. Models are employed to determine the relationship between the two sentences. For LLMs, we test ChatGLM and GPT3.5-turbo under the zero-shot and few-shot settings. Facing the challenge of unbalanced binary labels due to a lower frequency of some figures than often occurs. We implement negative sampling strategies in line with (Mikolov et al., 2013). For non-LLM models, we performed tests on the test set three times with different random seeds and reported the average results. For LLMs, we conduct experiments on the same test set, with the number of API calls limited to three.

Claim Extraction We report results on pre-trained models and LLMs. For the training phase, we strategically select a predetermined quantity of sentences that do not contain claims or evidence as negative instances. These are then paired with claims to form an enhanced training dataset for the task, thereby alleviating the imbalance. Given the recent exceptional achievements of LLMs in various NLP tasks, we also implement two direct prompting approaches using GPT-3.5 (OpenAI, 2022). These include Zero-shot Prompting, which employs a straightforward prompting technique with basic instructions, and Few-shot Prompting (Brown et al., 2020), incorporating a small number of examples (five in our study) with accurately categorized stances in the prompt (see **Appendix E** for detailed prompts). These methods are exclusively applied to the test set, with three iterations run and the average results reported.

Stance Classification As aforementioned, we also conduct experiments using pre-trained models and LLMs. We choose accuracy, support F_1, contest F_1 scores as automatic evaluation metrics and report them below.

Evidence Classification We choose Accuracy, Macro F_1 as automatic evaluation metrics and report. We concatenate the topic and the claim as a sentence and pair it up with the evidence candidates to predict whether it is evidence of the given claim under the specific topic.

Model	Setting	Acc.	Macro F_1
BERT-base-Chinese	fine-tuning	0.9888	0.7927
RoBERTa-base		0.9754	0.3883
Longformer-base		0.9760	0.4426
GLM4	0-shot	0.1395	0.0108
GPT3.5-turbo		0.0627	0.0029
GLM4	5-shot	0.6421	0.2125
GPT3.5-turbo		0.2280	0.0056

Table 5: Results of rhetoric classification task on this dataset.

Counter Speech Generation We utilize GPT-4o and GLM4 to generate counter speeches for both the affirmative and negative sides of the debate. We will vary the prompting methods and attempt to use Retrieval-Augmented Generation (RAG). The experiment is conducted in the zero-shot setting. **appendix F** provides more details.

4.3 Metrics

For the multi-class task of Rhetorical Figure Detection, we use accuracy (Acc.) and Macro F_1 as evaluation metrics. For the binary classification tasks of claim extraction and stance classification, as well as the multi-class task of evidence detection, we employ Macro F_1 and Claim F_1 as evaluation metrics. For the generative task of Counter Speech Generation, we use BERTScore and ROUGE-L as metrics.

5 Results and Discussion

5.1 Rhetorical Figure Detection

Table 5 shows results of the task. We observe that the BERT-base-Chinese model achieved the highest accuracy of 0.9888 and a macro F1 score of 0.7927. In contrast, the zero-shot performances of GLM4 and GPT-3.5 Turbo were substantially lower, with accuracy scores of 0.1395 and 0.0627 and macro F1 scores of 0.0108 and 0.0029, respectively. The 5-shot performance showed improvement, particularly for GLM4, which reached an accuracy of 0.6421 and a macro F1 score of 0.2125, indicating better adaptability with minimal training data.

5.2 Claim Extraction

Table 6 shows the claim extraction performance. The claim extraction results show that when BERT-

Model	Setting	Macro F_1	Claim F_1
BERT-base-Chinese	fine-tuning	0.8664	0.7495
RoBERTa-base		0.9473	0.5902
Longformer-base		0.7880	0.6004
GLM4	0-shot	0.0343	0.1056
GPT3.5-turbo		0.1875	0.1060
GPT-4o		0.2057	0.5676
GPT-4 Turbo		0.5396	0.6087
GLM4	5-shot	0.1362	0.1253
GPT3.5-turbo		0.3211	0.4632
GPT-4o		0.4058	0.3095
GPT-4 Turbo		0.5191	0.4470

Table 6: Results of claim extraction performance on this dataset.

Model	Setting	Acc.	F_1
BERT-base-Chinese	fine-tuning	0.6307	0.6577
RoBERTa-base		0.5051	0.6712
Longformer-base		0.5003	0.3337
GLM4	0-shot	0.0370	0.0192
GPT3.5-turbo		0.3989	0.0977
GLM4	5-shot	0.1465	0.0486
GPT3.5-turbo		0.4018	0.0960

Table 7: Results of stance classification performance on this dataset.

base-Chinese achieved a strong performance with a Macro F1 of 0.8664 and Claim F1 of 0.7495, indicating effective claim identification. Conversely, RoBERTa-base, despite a higher Macro F1 of 0.9473, had a lower Claim F1 of 0.5902, suggesting challenges in precise claim detection. In zero-shot settings, GLM4 and GPT-3.5-turbo performed poorly, but few-shot learning improved their performance significantly, with GPT-3.5-turbo reaching a Macro F1 of 0.3211 and Claim F1 of 0.4632.

5.3 Stance Classification

As summarized in Table 7, we compared the performance of various language models under different settings. The results indicate that BERT-base-Chinese, after fine-tuning, achieved the best performance with an accuracy of 0.6307 and an F1 score of 0.6577; RoBERTa-base slightly outperformed in terms of the F1 score, reaching 0.6712, though its accuracy was only 0.5051. In contrast, large models such as GLM4 and GPT-3.5-turbo performed poorly under zero-shot and five-shot settings, particularly GLM4, with accuracy and F1 scores of merely 0.0370 and 0.0192, respectively, highlighting the limitations of these models in low-sample tasks.

Model	Setting	Acc.	F_1
BERT-base-Chinese	fine-tuning	0.9813	0.7752
RoBERTa-base		0.9091	0.1587
Longformer-base		0.9453	0.2632
GLM4	0-shot	0.2315	0.0159
GPT3.5-turbo		0.5263	0.0236
GLM4	5-shot	0.5471	0.0168
GPT3.5-turbo		0.6043	0.0760

Table 8: Results of evidence detection performance on this dataset.

5.4 Evidence Classification

Results We randomly choose one negative sample for each sentence that presents evidence. As presented in Table 8, results indicate that the BERT-base-Chinese model with fine-tuning achieves superior performance in evidence detection tasks, with an accuracy of 0.9813 and an F1-score of 0.7752. In contrast, the RoBERTa-base and Longformer-base models demonstrate moderate accuracy values but exhibit lower F1 scores. Notably, the large models, GLM4 and GPT3.5-turbo, show varying performance across different conditions, with a slight improvement in F1-scores from 0-shot to 5-shot settings, underscoring the potential for enhanced performance through further refinement in few-shot learning.

5.5 Counter Speech Generation

We extracted the speeches from both the pro and con sides of each debate, constructing lengthy speech texts. Expanding the reply content and using more detailed prompts significantly improved the BERTScore for the generated rebuttals, suggesting that models require more comprehensive instructions to generate effective counter speech. The preliminary experimental results are listed in Table 23 below. See **Appendix F** for more details.

6 Conclusion

In this work, we present a novel and comprehensive Chinese debate dataset for computational argumentation. Our dataset incorporates meticulously annotated multi-type labels to support related tasks. Besides, we explore the task of identifying Chinese rhetorical figures from argumentative texts. We also evaluate several baseline argument mining methods on the dataset.

Limitations

Despite our efforts, the dataset might contain potential biases introduced by annotators. Such as uneven topic distribution, differences in debaters' backgrounds. Utterances are made by students from higher education institutions or renowned Chinese debaters, and most of these debaters are from East Asia. Further research should consider data preprocessing and debiasing methods to mitigate these effects. We intend to employ a broader range of prompt engineering techniques to experiment with the dataset in the future. Additionally, we plan to expand the scale of the debate competition and explore a more comprehensive array of tasks suitable for utilizing the dataset.

Ethics Statement

All data in this study are freely available to the public. We follow the policy of using these data without infringing any copyright issues. All debate videos we collected were publicly released by competition organizers on publicly accessible video-sharing platforms. The manual annotation and human evaluation were carried out by collaborators within our research group. We completed the manual annotation and human evaluation with the assistance of our research group students. Payment is adequate given the participants. Their demographic information is presented in Table 9.

Demographic Characteristics	Value
Total Participants	20
Age	[23, 30]
Sex (Female/ Male)	8 / 12
Mandarin Chinese Proficiency	all native
Education	postgraduate

Table 9: The demographic information of the annotators. All evaluators have at least a postgraduate-level education.

References

Tazin Afrin, Elaine Wang, Diane Litman, Lindsay C Matsumura, and Richard Correnti. 2020. Annotation and classification of evidence and reasoning revisions in argumentative writing. *ACL 2020*, page 75.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8782–8797. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021a. Argument undermining: Counter-argument generation by attacking weak premises. *arXiv preprint arXiv:2105.11752*.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021b. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967, Dubrovnik, Croatia. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Stephen Bothwell, Justin DeBenedetto, Theresa Crnkovich, Hildegund Müller, and David Chiang. 2023. Introducing rhetorical parallelism detection: A new task with datasets, metrics, and baselines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5007–5039, Singapore. Association for Computational Linguistics.

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.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

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*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of NAACL-HLT*, pages 542–557.

Xin Chen, Zhen Hai, Deyu Li, Suge Wang, and Dian Wang. 2021. Jointly identifying rhetoric and implicit emotions via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1429–1434, Online. Association for Computational Linguistics.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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. *arXiv preprint arXiv:2305.19902*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis-and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672.

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

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Shachar Mirkin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, et al. 2019. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. *arXiv preprint arXiv:1907.11889*.

John Lawrence, Jacky Visser, and Chris Reed. 2017. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8(3):289–310.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. *EMNLP 2018*, page 121.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuan-Jing Huang, and Zhongyu Wei. 2023a. 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.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023b. 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, Singapore. Association for Computational Linguistics.

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.

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724.

OpenAI. 2022. Gpt-3.5-turbo-16k.

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: 1st European Conference on Argumentation (ECA 16)*. College Publications.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization

dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396.

Keshav Singh, Paul Reisert, Naoya Inoue, Pride Kavumba, and Kentaro Inui. 2019. Improving evidence detection by leveraging warrants. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 57–62.

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.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

Dawei Zhu, Qiusi Zhan, Zhejian Zhou, Yifan Song, Jiebin Zhang, and Sujian Li. 2022. ConFiguRe: Exploring discourse-level Chinese figures of speech. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3374–3385, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Annotation Guidelines

Firstly, each annotator received the previously mentioned annotation manual, underwent training with live demonstrations, and was provided with online documentation for annotation, demonstration videos, transcribed texts, and links to original audio and video on cloud storage. This ensured annotators' engagement and consistency. Annotations were mainly fact-based, and the authenticity of speech roles could be verified through the original videos, thereby gaining validation and recognition.

To effectively carry out annotation, a comprehensive set of annotation guidelines has been established, encompassing several pivotal phases:

Training Phase: The annotators undergo an intensive three-day training session to become proficient in the guidelines of the annotation manual. This training includes interactive live demonstrations, access to online documentation, and a variety of resources such as demonstration videos, transcribed texts, and links to the original audio and video materials.

Topic Extraction and Restatement: During this phase, annotators are tasked with identifying and articulating the topics from the debates into clear, concise position statements.

Role Annotation: This involves annotating the roles of the debate participants (for instance, Pro First Speaker, Con Third Speaker) based on their speeches. This annotation is cross-referenced with the original video recordings to ensure accuracy and context relevance.

Comprehensive Annotation: In this stage, various elements are annotated, including arguments, types of evidence, the relationships between arguments and evidence, modal words, rhetorical devices, competition outcomes, and temporal annotations. Evidence is meticulously categorized into Case, Expert, Research, Explanation, and Others for precise identification.

Post-Annotation Quality Control: This crucial step involves meticulous reviews and corrections to guarantee the accuracy and consistency of the annotations.

The samples used in our experiment consist of these five types of evidence and other debate speech sentences. The case refers to instances drawn from real-life events, examples, or occurrences among the evidence types. Expert type encompasses opinions or statements from authoritative professionals, scholars, or official organizations. Research includes findings or conclusions derived from scientific studies, statistical reports, or surveys. Explanation pertains to comprehensive elaborations or clarifications of the claim, including its reasons or impacts. Lastly, others type captures any evidence that does not fit the categories above.

Evidence types are explicitly defined to include real-world examples (Evidence Type Case), opinions from experts or authoritative bodies (Evidence Type Expert), findings from scientific research or surveys (Evidence Type Research), comprehensive explanations of claims (Evidence Type Explanation), and any evidence that does not conform to the categories above (Evidence Type Others).

B Label Examples

Category	Example
Topic	AI绘画是艺术吗?
Pro/Con Major Claim	AI绘画是艺术。/ AI绘画不算艺术。
Debater Utterance	**Pro-side Speaker 1**: "首先，AI有能力达到满足人类审美价值的要件。" **Con-side Speaker 2**: "您方所谓的AI绘画中，人类参与的部分大概是多少？" **Pro-side Speaker 1**: "说实话，AI目前完全没有办法脱离人类独自进行创作。" **A Few Rounds Later...** **Con-side Speaker 1**: "整个AI绘画过程不过是经验和随机性的堆叠，以及人为设置的对目标值的奖励。" **A Few Rounds Later...** **Pro-side Speaker 3**: "我们不仅认为要有审美价值，同样我们也认可艺术确实要传达创作者的意图。"

Table 10: A simplified example from one debate.

As illustrated in Table 10, for each debate, our dataset consists of no less than the crucial elements: the debate topic, the roles of the speakers (host and different debaters), and their utterances labeled with claims, stance (Pro/Con), evidence (Table 11), rhetorical figures (Table 12), speech etc. Our dataset has eight debater roles: Pro-side Speaker 1 to 4, host, and Con-side Speaker 1 to 4. The annotation elements are in Chinese and at the sentence level.

More concretely, Table 13 presents examples of debate topics and their types. Besides, Table 12 introduces a concrete rhetorical figures type case defined in our dataset. The items below are detailed explanations.

- Metaphor: A metaphor is a figure of speech that identifies something as being the same as some unrelated thing for rhetorical effect, thus highlighting the similarities between the two. Commonly seen in debate language, a metaphor involves using one entity to represent another based on similarities between the two, albeit to varying degrees. Common types include similes, implicit metaphors, and allegories.

- Hyperbole: Hyperbole is an exaggerated statement or claim not meant to be taken literally but used for emphasis or rhetorical effect. Hyperbole is often employed in debates to achieve an expressive specific impact, deliberately exaggerating or diminishing aspects such as the imagery, characteristics, or magnitude of something to create a strong impression on the audience.

- Contrast: This rhetorical device involves juxtaposing two or more elements to highlight their differing characteristics, thereby emphasizing their distinctiveness. This technique involves placing two opposing or contrasting entities together in a debate. Using comparative methods to describe or illustrate them reveals their distinct features and makes the argument more memorable.

- Rhetorical Question: A rhetorical question is a figure of speech in the form of a question that is asked to make a point rather than to elicit an answer. Often, self-answered, rhetorical questions are used to spark interest and thought. Debaters might employ concise rhetorical questions to reinforce their viewpoints and create a direct impact when the atmosphere becomes subdued in debates.

- Hypophora: Hypophora is a figure of speech where the speaker poses a question and immediately answers it, often used to guide the topic or emphasize a point. More prevalent than rhetorical questions, hypophora in debates involves asking questions without providing answers. It's used in the content related to the debate topic and as a technique in argumentation.

- Quotation: This involves directly citing words from another source or person to enhance the authority of an argument or to explain a point more clearly. Debaters commonly use existing quotes, idioms, or proverbs to support their points and enhance their expressive effect.

- Antithesis: This technique uses paired symmetrical structures to express ideas, often creating correspondence in sentence structure, meaning, and syllables.

- Parallelism: Parallelism uses components in a sentence that are grammatically the same or similar in their construction, sound, meaning, or meter, used for rhythm and emphasis. A rhetorical device where similar structures of

Evidence Type	Evidence
案例 Case	爱迪生花了20年的时间做了5万多字的经验，才能够成功地发明了电芯。 Edison spent 20 years and wrote over 50,000 words on experiences before he could successfully invent a battery.
解释 Explanatinon	随着社交媒体时代的蓬勃发展，人们对于在社交媒体上展示自己旅游经历 和获取关注的需求增加，更倾向于追求照片中的视觉效果，而不是对当地 文化的深入了解，是快节奏生活的压迫感，是对人云亦云的盲目追求，而在 奔波和攀比中忽略了旅游真正的意义和文化体验。 With the vigorous development of the social media era, people's desire to showcase their travel experiences on social media and gain attention has increased. They tend to pursue visual effects in photos rather than a deep understanding of local culture. This reflects the pressure of a fast-paced life, a blind pursuit of following the crowd, and in the rush and comparison, travel's true meaning and cultural experience are overlooked.
专家 Expert	正如卢梭所言，人生而自由却无时无刻生活在社会的枷锁之中。 As Rousseau said, people are born free yet are in chains everywhere in society.
研究 Research	某大学发布的大学生调研报告中提到，71.59%大学生每月所得收入在300元以下， 大部分大学生几乎没有除了家庭以外的其他经济来源。 A university research report mentioned that 71.59% of college students have a monthly income of less than 300 yuan, with most students having almost no financial sources other than their families.

Table 11: Labeled evidence example for four main types. The English explanation in the table is a translation of Chinese. It does not exist in the dataset.

Rhetorical Figures Type	Sentences
比喻 Metaphor	钱是洪水猛兽吗？不！贫穷才是！ Money is a flood and a fierce beast? No! Poverty is!
夸张 Hyperbole	我的颜值，堪称撞脸十八家，长得像彭于晏等男神级人物。 My appearance resembles a mix of eighteen celebrities, resembling male icons like Peng Yuyan (an actor). 好看的人不会以别人为标准，每个人心里都带着一个美图秀秀在看自己，迷之自信一时爽。 Attractive people don't use others as their standard; everyone views themselves through an internal beautifying lens, enjoying a moment of unfounded confidence.
对比 Contrast	医疗不发达的时候，我们每个人生了重病去看医生，我们会面临两种结果，要么死，要么活； In times of underdeveloped medical care, facing a serious illness meant two outcomes: life or death. 今天医疗科技十分发达，它在死和活之间开辟出了第三条道路，半死不活。 Today, with advanced medical technology, a third option has emerged between life and death: a state of being half-dead.
设问 Rhetorical Question	我认为不该有，为什么？很简单，如果在什么年龄干什么事的话，我们为什么要祝老人长命百岁呢？ (Should age dictate our actions?) I don't think so. Why? It's simple. If we restrict activities by age, why wish the elderly a long life? 因为他100多岁的高龄也是时候该走了。 It suggests that at their advanced age, it's time for them to go.
反问 Hypophora	对方辩友一再不回答我方的问题，是不是默认了我方观点又不好意思承认呢？ Is the opponent's consistent avoidance of our questions an implicit agreement with our point, too embarrassed to admit it? 今天我们双方进行争论，难道仅仅只是为了输赢吗？ Are we debating today just for the sake of winning or losing?
引用 Quotation	那些历史上真正收获了豁达心态的人，杨慎，是非成败转头空；王维，行到水穷处，坐看云起时。 Those in history who truly attained a broad mindset, like Yang Shen who said 'Right and wrong, success and failure, in the end, all are nothing'; and Wang Wei who mused 'Travel to where the water ends, and sit watching the rising clouds.' 他们是在遍历了人世的沧桑，经历了繁华才悟出的道理，如果青年的他们不做加法又何来如此的境界。 These realizations came after experiencing life's vicissitudes and splendor. Without adding experiences, how could they have reached such a state?
对偶 Antithesis	读书破万卷，下笔如有神。 Reading extensively brings divine-like proficiency in writing.
排比 Parallelism	教育培养了我们的思维，塑造了我们的未来，决定了我们的命运。 Education cultivates our thinking, shapes our future, and determines our fate.

Table 12: Labeled rhetorical figures utterance example for our eight main types. Some rhetorical figures are only valid in Chinese contexts. The English explanation in the table is a translation of Chinese. It does not exist in the dataset.

phrases or sentences are used to strengthen an argument and enhance the expressiveness of language. Parallelism brings rhythm and focus to speech, making the argument more powerful and impactful.

- Others: This category includes various other rhetorical devices that do not fall into the abovementioned categories, each employed for specific persuasive or stylistic effects.

In terms of evidence types, these are explicitly defined to include real-world examples (Evidence Type Case), opinions from experts or authoritative bodies (Evidence Type Expert), findings from scientific research or surveys (Evidence Type Re-

Debate Topic	Topic Type
真爱是不是谎言？ Is true love a lie?	Philosophy
熬夜是现代人的病还是药？ Is staying up late a modern-day illness or a remedy?	Health
内卷是个真问题还是假问题？ Is "involution" a real issue or a false one?	Sociology
AI绘画是不是艺术？ Is AI-generated art truly art?	Art & Technology

Table 13: Examples of debate topics and their types. The English explanation in the table is a translation of Chinese. It does not exist in the dataset.

Benchmarks	Datset Properties			
	Arg Level	Content Type	Spoken Language	Language
Peldszus and Stede (2016)	paragraph	written or generated texts	✗	De/En
Hua and Wang (2018)	paragraph	social media posts	✗	En
Hua et al. (2019)	paragraph	social media posts	✗	En
Schiller et al. (2021)	sentence	social media posts	✗	En
Alshomary et al. (2021a)	paragraph	social media posts	✗	En
Alshomary and Wachsmuth (2023)	paragraph	social media posts	✗	En
Lin et al. (2023a)	sentence	social media posts	✗	En
Ruggeri et al. (2023)	paragraph	scientific texts	✗	En
Ours	sentence & discourse	debate competitions	✓	Zh

Table 14: Comparison between various argument generation corpus, where "Arg" means argument. We provide both argument mining and generation elements.

Benchmarks	Datset Properties			
	# Rhetoric Types	Content Type	Spoken Language	Language
Lawrence et al. (2017)	8	social media radio	✓	En
Chen et al. (2021)	5	literary works, textbooks, microblog, websites	✗	Zh
Bothwell et al. (2023)	1	sermons and exam essays	✗	Latin/Zh
Zhu et al. (2022)	12	novel and proses	✗	Zh
Ours	9	debate competitions	✓	Zh

Table 15: Comparison between various rhetorical figures benchmarks. The English explanation in the table is a translation of Chinese. It does not exist in the dataset.

search), comprehensive explanations of claims (Evidence Type Explanation), and any evidence that does not conform to the categories above (Evidence Type Others).

Moreover, Table 14 compares argument generation corpus. And Table 15 compares rhetorical figures benchmarks. Eventually, Table 16 is an overall statistics comparison for datasets.

C Quality Control & Inconsistent Treatment

We filtered out videos according to the following criteria to ensure the dataset's quality. First, content risk control tools were applied to speeches containing inappropriate remarks. We filtered out instances of hate speech and irrational political discourse and discarded sessions with severely inappropriate content. Thus, unqualified videos were discarded. We also excluded videos that do not contain the entirety of the debate. Besides, some poor-quality audio may result in incorrect recognition, so they were also discarded. Videos with inaudible content are also eliminated, as manual post-correction is unfeasible for these recordings. Moreover, sessions that do not adhere to standard Chinese debate formats were excluded. Finally, to maintain a high standard in the quality of the debate speeches, we chose exemplary sessions from official competitions, prioritizing outstanding performances from the semifinal and final rounds of qualifying events. Through such an elaborate filtering process, we ultimately retained 600 debates out of the initially collected 980 competitions.

During the annotation process, annotation tasks were distributed to two groups of annotators for two rounds of annotation. Each consisted of 6 candidates, and within each group, all members needed to reach a consensus on all instances, and these groups worked independently. Additionally, five annotators were randomly selected to meticulously review the correctness and overall consistency of the post-edited translations done by other annotators. We calculated the Cohen's Kappa coefficient (Cohen, 1960) (Cohen, 1960) to assess the inter-annotator agreement between these two collective annotation groups. The resulting Kappa coefficient was close to 0.6, indicating a good agreement level between the two annotators groups. After a thorough review round, both reached a consensus. The discrepancy was the verdict by senior annotators. Given that the label assignment process was based on objective facts, we achieved high consistency in the results.

D Data Source

Table 17 presents the data sources used for dataset construction.

E Prompt Templates in LLM Experiments

Rhetorical Figure Detection
Table 18 presents claim extraction prompts.

Claim Extraction
Table 19 presents claim extraction prompts.

Stance Classification
Table 20 shows claim extraction prompts.

Evidence Classification

	# Topics	# Articles	# Claims	# Support	#Contest	# Evidence	# Rhetorical Figures	Content Domain	Language	Spoken Corpus
Levy et al. (2014)	32	326	976	-	-	-	-	Wikipedia articles	En	✗
Rinott et al. (2015)	39	274	1,734	-	-	3,057	-	Wikipedia articles	En	✗
Aharoni et al. (2014)	33 (12)	586	1,392	-	-	(1291)	-	Wikipedia articles	En	✗
Bar-Haim et al. (2017)	55	-	2,394	1,324	1,070	-	-	Wikipedia articles	En	✗
Cheng et al. (2022)	123	1,010	4,890	2,613	2,277	9384	-	online forum articles	En	✗
Zhao et al. (2023)	476	1,218	-	-	-	-	-	debate competitions	Zh	✓
HARMONIC(Ours)	318	600	15,359	8,202	7,157	32,173	20,126	debate competitions	Zh	✓

Table 16: Comparison of overall statistics between existing argument mining datasets. Topics represent unique topics in the corpus. Our dataset consists of Chinese oral debates, which are relatively scarce. The scale of annotated elements such as claims, claim stance, and evidence is more extensive than previous argument mining datasets. Moreover, we annotate the result label to judge the final result of a competition. Lastly, our corpus contains labeled Chinese rhetoric types, a further step for argument mining and quality evaluation.

Debate Competition Name	Year
国际大学群英辩论会 International Varsity Debate	1993-2011
国际华语辩论邀请赛 International Chinese Debating Competition	2013-2024
华语辩论世界杯 Chinese Debate World Cup	2018-2023
世界华语辩论锦标赛 The World Mandarin Debating Championship	2018-2023
华语辩坛老友赛 Mandarin Debate Veterans Tournament	2016-2023
亚太大专华语辩论公开赛 Asia-Pacific Intervarsity Chinese Debate Tournament	2013-2023

Table 17: After careful selection, we choose the most representative debates among them. All collected debates were released by the official accounts of debate competition organizers on publicly accessible video-sharing platforms.

Table 21 presents claim extraction prompts.

Counter Speech Generation

Table 22 shows claim extraction prompts.

F Counter Speech Generation

(Chen et al., 2023) proposed the Counter Speech Generation task, aiming to test the comprehensive capabilities of models in argument mining and generation. To assess the performance of large language models (LLMs) in Counter Speech Generation using Chinese corpora, we extracted the speeches from both the Pro and Con sides of each debate, constructing lengthy speech texts. The preliminary experimental results are listed in the table 23 below.

From the experimental results, we found that the replies generated by most models are relatively short compared to the reference texts (approximately 1900 characters for the model's replies versus 8000 characters for the original text). This length discrepancy is likely one of the main reasons for the lower scores in BERTScore(Zhang* et al., 2020) and ROUGEL (Lin, 2004). BERTScore considers semantic content, and ROUGEL assesses the overlap with reference discourse. Additionally, taking GPT-4o as an example, we instructed the LLM to generate rebuttal speeches for the pro-side speeches in the same debate. We provided three additional prompts, using all rounds of the model's responses as the answer and comparing them with the actual negative speeches. Our prompting methods included providing the speeches from the pro side's first speaker and the con side's third speaker and asking the model to continue the speech. Alternatively, we provided segments from the previous speakers and asked the model to respond in the role of a specific speaker on the con side. This approach aimed to simulate a specific role within the debate, encouraging the model to generate contextually appropriate and targeted counterarguments. By expanding the reply content, we found that the BERTScore could reach up to **0.1991**. This suggests the model may require more detailed prompts to generate more effective debate content. See Appendix E for the prompt template. Additionally, we implemented a straightforward Retrieval-Augmented Generation (RAG) framework using the LangChain to enhance the model's performance. This approach integrates retrieval mechanisms with generation tasks, providing the model with more contextually relevant information during the argumentative speech generation process.

Method	Prompt
0-shot	Given a topic and statement, determine which rhetorical figure is used in the debate statement, or if there is no rhetorical unit. You must only output with one of the following: 其他(Others), 反问(Rhetorical Question), 排比(Parallelism), 引用(Quotation), 夸张(Exaggeration), 设问(Hypophora), 对比(Contrast), 对偶(Antithesis), or O (No rhetorical device). No irrelevant content should be output. Debate Topic:{topic} Statement: {statement} Output:
5-shot	Given a topic and statement, determine which rhetorical figure is used in the debate statement, or if there is no rhetorical unit. You must only output with one of the following: 其他(Others), 反问(Rhetorical Question), 排比(Parallelism), 引用(Quotation), 夸张(Exaggeration), 设问(Hypophora), 对比(Contrast), 对偶(Antithesis), or O (No rhetorical device). No irrelevant content should be output. Topic: Innate genetics vs. acquired environment importance Statement: Is it just that we are avoiding the problem? Output: Rhetorical Question Topic: Should life's pursuit focus more on moral achievements or career accomplishments? Statement: We must call on them, we must encourage them, we must use love to make him brew coffee better. Output: Parallelism Topic: Does the rise of check-in tourism enrich or diminish the cultural content of tourism? Statement: As the saying goes, there are 1000 Hamlets in the eyes of 1000 readers. Output: Quotation Topic: Can paid knowledge alleviate young people's anxiety? Statement: What are these headlines? They are saying if you don't learn now, it will be too late. Output: Hypophora Topic: Does the prevalence of short videos enhance or reduce contemporary people's cognitive abilities? Statement: This is the short video your side wants to cite. Output: O Debate Topic:{topic} Statement: {statement} Output:

Table 18: Rhetorical figure detection on argumentative corpus for LLM prompting methods. Actual prompts are organized in Chinese. The English explanation in the table is a translation of Chinese. It does not exist in the dataset.

Method	Prompt
0-shot	Given a specific debating topic and related statements, automatically extract the claims from the article. Determine whether each statement is the relevant claim(C) or not (O). Output C (claim) or O (not claim). No irrelevant content should be output. Debate Topic:{topic} Statement: {statement} Output:
5-shot	Given a specific debating topic and related statements, automatically extract the claims from the article. Determine whether each statement is the relevant claim(C) or not (O). Output C (claim) or O (not claim). No irrelevant content should be output. Examples: Topic: Can individual and group interests coexist? Statement: This exactly proves the point we are making today that individual and group interests can coexist. Output: C Topic: Should individuals have the responsibility to resist consumerism? Statement: No, can you give an example? Output: O Topic: Is it more pitiful to be unfeeling or disloyal? Statement: But he also has loyalty, why? Output: O Topic: For contemporary youth, is it more beneficial to speed up or slow down life for growth? Statement: The affirmative believes that today's growth has only one meaning, which is that we need to acquire more skills and create more benefits. Output: C Topic: Do artists need a more detached or engaged mentality? Statement: Thanks to the third debater of the opposition, I now invite the third debater of the proposition to summarize. Output: O Debate Topic:{topic} Statement: {statement} Output:

Table 19: Claim extraction task prompts for LLM prompting methods. Actual prompts are organized in Chinese.

Method	Prompt
0-shot	Determine the stance of the following debate utterance and output 1 (proposition side), -1 (opposition side), or 0 (without a clear stance). Topic: {topic} Statement: {statement} Stance:
5-shot	Determine the stance of the following debate utterance and output 1 (proposition side), -1 (opposition side), or O (without a clear stance). Examples: Topic: Can individual and group interests coexist? Statement: This exactly proves our point today that individual and group interests can coexist. Stance: 1 Topic: Will the internet economy promote or hinder the development of craftsmanship? Statement: The internet economy hinders the development of craftsmanship. Stance: -1 Topic: Is adversity or prosperity more beneficial to personal growth? Statement: So we need to know that no matter who it is for, adversity is beneficial to our growth. Stance: -1 Topic: Are AI-generated artworks considered art or not? Statement: We believe it can expand the boundaries of art, and we have evidence to support this. Stance: 1 Topic: Is it harder to give up or to persist? Statement: Therefore, we believe that persisting is harder. Stance: -1 Topic: {topic} Statement: {statement} Stance:

Table 20: Stance detection task prompts for LLM prompting methods. Actual prompts are organized in Chinese.

Method	Prompt
0-shot	Given a topic and statement, determine whether the statement is a premise (evidence) supporting the topic in the context of the debate. Only output one of the following: 案例(Case), 解释(Explanation), 研究(Research), 专家(Expert), 其他(Others), or O (Not evidence). Topic: {topic} Statement: {statement} Output:
5-shot	Given a topic and statement, determine whether the statement is a premise (evidence) supporting the topic in the context of the debate. Only output one of the following: 案例(Case), 解释(Explanation), 研究(Research), 专家(Expert), 其他(Others), or O (Not evidence). Topic: Can a person do anything to save their own life? Statement: During the Spring and Autumn period, the state of Song was besieged by the state of Chu. When food ran out, the people exchanged each other's children as food to survive. Answer: 案例(Case) Topic: Should deeply felt emotions be forgotten or remembered? Statement: Remembering deeply felt emotions gives a person's emotional experience both highs and lows. Just as the survival of the body relies on the ups and downs of an ECG (a medical test that records the electrical activity of the heart to detect heart problems.) , the soul cannot be separated from the positioning and proof of deeply felt emotions. Answer: 解释(Explanation) Topic: Is adversity or prosperity more beneficial to personal growth? Statement: Today, we say that a study in criminology shows that adolescents from broken homes have a crime rate three times higher than those from normal homes. Answer: 研究(Research) Topic: Should "racketeering" rights protection be protected by law? Statement: The Secretary-General of the China Consumer Law Research Association stated that 80% of rights protection cases target meaningless trademark signs, such as synonyms in foreign translations of almonds and apricot kernels. These complaints not only waste judicial resources but also harm market order. Answer: 专家(Expert) Topic: Is marrying someone of equal social status an outdated marriage value? Statement: What right do you have to judge whether it is wrong for someone to choose to marry someone of equal social status? Answer: O (Not evidence ingredient) Topic: {topic} Statement: {statement} Output:

Table 21: Evidence classification task prompts for LLM prompting methods. Actual prompts are organized in Chinese.

Method	Prompt
0-shot	Generate counter speeches in response to debate speeches. Debate Speech: {Pro-side or Con-side Chinese speech} Counter speech:

Table 22: Counter speech generation task prompts for LLM prompting methods. Actual prompts are organized in Chinese.

Method	Stance	BERTScore	ROUGEL
GPT-4o	Pro	0.0953	0.0036
	Con	0.0832	0.0042
GLM4	Pro	0.0831	0.0013
	Con	0.0732	0.0032
GPT-4o w/ RAG	Pro	0.1369	0.0033
	Con	0.1235	0.0034

Table 23: Results of counter speech generation.