CEDAR: A Chinese Evaluation Dataset for Computational Argumentation

Tian Lan, Jiang Li, Rong Yan*, Feilong Bao, Weihua Wang, Guanglai Gao, Xiangdong Su*
[1] College of Computer Science, Inner Mongolia University, China
[2] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
[3] Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, China
velikayascarlet@gmail.com, cssxd@imu.edu.cn

Abstract

Computational argumentation has received increasing attention in recent years. However, existing debate datasets neglect some important labels for argument mining, generation, and evaluation. Meanwhile, the lack of comprehensively annotated Chinese oral debate datasets hinders progress in this field. To address these gaps, we introduce a comprehensive **C**hinese **E**valuation **D**ataset for Computational **Ar**gumentation, named **CEDAR**. Compared to previous datasets, CEDAR includes the essential labels of computational argumentation (claim, stance, evidence) and five additional crucial labels: rhetorical figures, debater roles, modal words, utterance time, and debate results. Moreover, it offers complete transcripts of each debate, including speeches from the Pro and Con sides. Thus, the proposed CEDAR not only supports common argument mining and generation tasks, but also provides resources for rhetorical figure detection, argument quality evaluation, and debate result prediction. This dataset covers 600 debates about 318 topics from Chinese debate competitions. Besides providing a dataset for research, we conduct experiments on common computational argument tasks and a novel task (rhetorical figure detection), in which we also evaluate LLMs. The experimental results highlight the challenging nature of the dataset. Our corpus is available at https://github.com/VelikayaScarlet/CEDAR.

1 Introduction

Language models have demonstrated strong performance across diverse fields (Wang et al., 2024; Lan et al., 2025; Zhou et al., 2025; Zhang et al., 2025; Duo et al., 2026; Wei et al., 2026). Consequently, a growing body of research has begun to focus on capabilities in computational argumentation. Computational argumentation aims to identify argument components, analyze their relationships, and generate arguments automatically (Cabrio and Villata, 2018; Lin et al., 2023b) and plays a promising role in various fields, such as linguistics, law, and education. It includes argument mining and argument generation. Argument mining (AM) aims to identify and extract arguments automatically, whose 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). Computational argumentation poses challenges since it requires complex logic, emotional tones, strong reasoning, and rhetorical skills. Researchers have devoted effort to both argument mining and generation, including paragraph-level counter-argument generation (Hua and Wang, 2018; Alshomary et al., 2021; Alshomary and Wachsmuth, 2023), sentence-level counter-argument generation (Lin et al., 2023b), and argumentative dialogue generation and summarization (Le et al., 2018; Zhao et al., 2023).

Despite significant advancements in relevant tasks, research progress has been impeded by the following two factors. First, existing debate datasets overlook some important labels for argument mining, generating, and evaluation, such as rhetorical figures, modal words, utterance time, and debate results. Second, the lack of comprehensively annotated Chinese oral debate datasets hinders progress in this field, although numerous argument mining datasets have been developed for English (Levy et al., 2014; Rinott et al., 2015; Aharoni et al., 2014; Bar-Haim et al., 2017; Cheng et al., 2022; Visser et al., 2020; Hautli-Janisz et al., 2022; Haddadan et al., 2019; Walker et al., 2012).

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5247–5269

Datasets	Dataset Properties										
	Content Type	Topic	Claims	Pro/Con Stance	Evidence with Multi-Types	Rhetoric Figures	Debate Results	Argument Pairs	Debater Role & Time	Spoken	Lg.
CDCD	Wikipedia articles	✓	✓	✗	✗	✗	-	✗	-	✗	En
CDED	Wikipedia articles	✓	✓	✗	✓	✗	-	✗	-	✗	En
Claims and Evidence	Wikipedia articles	✓	✓	✗	✗	✗	-	✗	-	✗	En
IBM Stance Classification	Wikipedia articles	✓	✓	✓	✗	✗	-	✗	-	✗	En
IAM	Wikipedia articles	✓	✓	✓	✗	✗	-	✗	-	✗	En
ORCHID	real-world debate	✓	✓	✓	✗	✗	✗	✗	✗	✓	Zh
IBM 2019	real-world debate	✓	✓	✗	✗	✗	✗	✗	✗	✓	En
US2016	real-world debate	✓	✓	✓	✗	✗	✗	✗	✗	✓	En
QT30	real-world debate	✓	✓	✓	✓	✗	✗	✗	✗	✓	En
USElecDeb60To16	real-world debate	✓	✓	✓	✓	✗	✗	✗	✗	✓	En
IAC	real-world debate	✓	✓	✓	✗	✓	✗	✗	✗	✓	En
CEDAR (Ours)	real-world debate	✓	✓	✓	✓	✓	✓	✓	✓	✓	Zh

Table 1: Comparison among debate and argument mining datasets. Claims and Evidence: the basic argument mining elements; Rhetorical figures: the essential persuasive strategies in the debating process; Debate Results: the result of each debate competition (Pro/Con); Dialogue Role: 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/Written: the content is spoken or written language; Lg.: the language used in the dataset (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	Datset Properties							
	Content Type	Argument Generation	Argument Summarization	Counter Speech Generation	Debate Results	Spoken	Lg.	Dialogue Role & Timestamp
CounterArguGen	online forum	✓	✗	✗	✗	✗	En	✗
ConcluGen	online forum and debate corpora	✗	✓	✗	✗	✗	En	✗
DebateSum	debate corpora	✗	✓	✗	✗	✗	En	✗
CounterSpeechGeneraion	real-world debate	✗	✗	✓	✗	✓	En	✗
ORCHID	real-world debate	✗	✓	✗	✗	✓	Zh	✗
Microtexts	written or generated texts	✓	✓	✗	✗	✗	De/En	✗
RedditChangeMyView	social media posts	✓	✓	✗	✗	✗	En	✗
CounterArgumentGeneration	social media posts	✓	✓	✗	✗	✗	En	✗
UKP-Corpus	social media posts	✓	✓	✗	✗	✗	En	✗
ArgTersely	social media posts	✓	✓	✗	✗	✗	En	✗
ArgSciChat	scientific texts	✓	✓	✗	✗	✗	En	✗
CMV	social media posts	✓	✓	✗	✗	✗	En	✗
CEDAR (Ours)	real-world debate	✓	✓	✓	✓	✓	Zh	✗

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

Only Zhao et al. (2023) proposes a Chinese debate dataset for stance classification and argumentative dialogue summarization. As a result, there is an obvious need for more Chinese debate benchmarks with fine-grained labels.

To this end, we present a comprehensive oral Chinese debate dataset CEDAR with more fine-grained labels to advance research. We compare the proposed dataset CEDAR with previous debate datasets in Table 1 and Table 2 based on their properties and task applicability. As shown in Table 1, unlike previous argument mining datasets, CEDAR includes rhetorical figures, debate results, modal words, debater role, and utterance time. In addition, CEDAR includes annotations for the topic, claim, pro and con side stance, and evidence. CEDAR is a multi-modal dataset that includes both text and audio from real-world debates. CEDAR offers more labels and supports a wider range of argument generation tasks. Table 12 in **Appendix A** presents a statistics comparison of argument mining datasets. The scale of annotated elements such as claims, claim stance, and evidence in CEDAR is more extensive than previous argument mining datasets. Lastly, our corpus contains labeled Chinese rhetoric types, which can be used for argument mining and quality evaluation. Table 13 in **Appendix A** presents the comparison of the overall statistics among existing argument generation datasets.

To our knowledge, CEDAR is the first comprehensive Chinese debate dataset for argument mining and generation. It also supports rhetorical figures detection in debates. The annotated debate results can be used to assess argument quality. CEDAR contains 600 debates, 318 topics and over 250K sentences in Mandarin. We use an automatic speech recognition tool to transcribe the original videos, followed by manual annotation and verification. CEDAR is provided in JSON format

designed to support various tasks. The dataset includes fields such as debate, claim, evidence, and rhetoric. In particular, the debate field is a list of key-value pairs, each representing a debater's name and their corresponding utterances (for example, a Pro side debater labeled A1 or a Con side debater labeled B3). Depending on the research goal, users can extract content from the relevant fields to address different tasks. To better study debates, 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.

It is worth noting that, although the proposed dataset CEDAR uses online debate source as ORCHID, there are many significant differences between them, as shown in Table 1. First, the annotation properties in CEDAR are far more than ORCHID. That is, the proposed CEDAR provides 8 properties, while ORCHID only provides 3 properties. Second, the additional fine-grained properties in CEDAR make it applicable to more argument mining tasks and other NLP tasks related to debate. Third, CEDAR provides the text and audio of each debate, which can be used for multi-modal investigation. Forth, part of debates are different in these two datasets, and the processing method and the results are far different.

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

2 Related Works

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 Alshomary et al. (2022) contains argumentative texts and ethical considerations. Zhao et al. (2023) proposed a dataset for argumentative dialogue summarization. Ruggeri et al. (2023) presented a dataset of argumentative dialogues in scientific papers. Lin et al. (2023b) presented a dataset for sentence-level counter-argument task. Roush and Balaji (2020) developed 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. Jo et al. (2020); Lin et al. (2023a) released datasets based on ChangeMyView forum.

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 provided 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 prose for figures recognition. We have summarized and compared them in Table 11.

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

Counter Speech Datasets Argument mining from speech has been extensively 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 transcribed texts. Chen et al. (2023) proposed 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 18 in **Appendix E** for details). The videos consist of standard debate conversations in Chinese Mandarin. The standard format of Chinese debate competitions consists of two teams: the affirmative and the negative, each with four speakers. We excluded debates where the number of speakers was unbalanced or insufficient. Then, we employed the IFLYTEK Automatic Speech Recognition (ASR) technique to obtain raw transcriptions from the audio recordings. Subsequently, a stringent filtering process was conducted to discard recordings with unclear audio, hate speech, or irrational political content. The original audio was retained as a reference to assist annotators in accurately labeling speaker roles in the following steps.

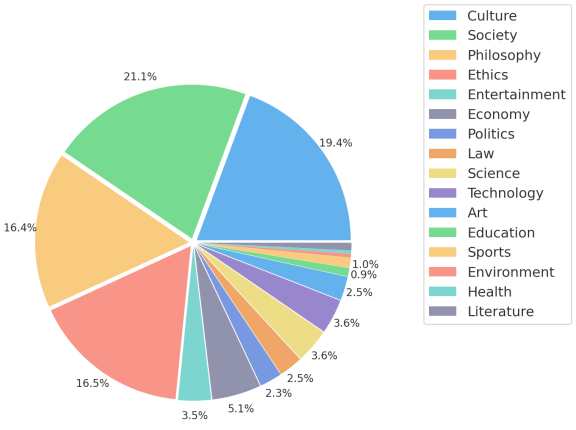

Figure 1: The situation of topic types in the CEDAR. Competitions are divided into a total of 16 types. Each competition may encompass multiple types.

3.2 ASR Transcription and Filtering

We utilized the widely recognized iFlytek ASR (Automatic Speech Recognition) tool, which efficiently converts Chinese speech to text. Then, annotators manually corrected the transcripts for lexical errors or incorrect punctuation. We used speech transcription tools on audio files to obtain the raw transcription texts. Next, we removed irrelevant content, such as accidental interruptions and advertisements. For the debates whose speakers are not presented in the camera, it is difficult to distinguish the debater's role. We discarded 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 corrected 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 in annotating the properties of the debate. 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 B** for more annotation details and **Appendix C** 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 anno-

tated various elements such as claims, stances for claims, evidence, pro/con speeches, rhetorical figure types, modal words, debate results, and timestamps.

Post-Annotation Check: We checked and corrected all annotations to ensure their accuracy and consistency.

We established elaborate annotation guidelines to ensure high quality. To ensure consistency, we recruited 20 native Chinese speakers holding at least a bachelor's degree to act as annotators. They underwent a three-day training course, including demonstrations, online documentation, video recordings, and Q&A sessions. Then, they followed guidelines to extract topics, assign roles, and annotate various aspects of the debates. 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. A custom platform was introduced, which supports span and relation labeling for claims, stances, evidence, and rhetoric to support the annotation. Two annotators jointly completed 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, and manually categorize the topics into 16 types, including culture, society, philosophy, ethics, entertainment, economy, politics, law, science, technology, art, education, sports, environment, health, literature. Figure 1 shows the statistics of the topic 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 corrected any missed or incorrectly labeled utterances. Also, we ensure that every utterance segment has the correct speaking role labeled manually. 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 C** 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 labeled 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 data set are classified into case, expert, research, explanation, etc. It ensures a comprehensive analysis of evidence in debates. We manually labeled the evidence with types for 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 the 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 group has 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 of 0.66, indicating a substantial agreement. After a thorough review, consensus was reached by both groups, and senior annotators resolved any discrepancies. See **Appendix D** for more details.

3.5 Dataset Overview

We have a total of 600 debates about 318 topics, with over 250K sentences (non-argumentative components included). The average length of each debate is 20,158 Chinese characters. 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. More statistical information can be found in **Table 12** in **Appendix A**. Compared to previous argument datasets, we propose a more comprehensive Chinese debate dataset with more fine-grained labels. This dataset can support the research of computational argumentation in Chinese.

Type	Ours
Total debates	600
Total topics	318
Total claim sentences	8,251
Total evidence sentences	5,173
Total rhetorical figures sentence	1,126
Total sentences	251,726
Avg. debate words length	20,158
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 CEDAR. Lengths of sentences are measured in Chinese characters.

4 Experiment

4.1 Tasks

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

Rhetorical Figure Detection We are the first to evaluate rhetorical figure detection on the debate dataset. Given a sequence of words that forms a figurative unit U (Zhu et al., 2022), 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 C**. Moreover, eight examples are shown in Table 16. 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 Let $D = \{d_i\}$ be a set of debates, where each debate d_i consists of multiple utterances. The goal is to automatically extract claims $\{c_j\}$ from these debates, where each claim c_j is a sentence or sentence. This is a central task in argument mining, essential for identifying and analyzing the main arguments within a debate.

Stance Classification 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). Among the 8,251 claim sentences, support claims and contest claims account for 66% and 34%, respectively.

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).

Category	Train	Dev	Test	Total
debates	480	60	60	600
claim sentences	6,600	825	826	8,251
evidence sentences	4,139	517	517	5,173
rhetorical figure sentences	901	113	112	1,126
argument pairs	2,906	362	363	3,631
Pro utterances	32,745	4,092	4,092	40,929
Con utterances	32,854	4,106	4,106	41,066

Table 4: The statistics about train/dev/test sets. In Chinese debates, the "Pro" category contains statements from the four affirmative debaters, and the "Con" category contains those from the four negative debaters. These two categories are prepared for counter speech generation tasks.

4.2 Experimental Setting

We split the dataset by **unique topic** (8:1:1) to ensure that all debates associated with a specific topic are isolated within the same split. This strategy tests the model on entirely unseen topics, thereby evaluating its true generalization ability. We treat all tasks except counter speech generation as sentence classification.To mitigate class imbalance and increase task difficulty, we apply negative sampling with a 1:10 positive-to-negative ratio. Specifically, negative samples are strictly selected from within the same debate as the positive instances. This strategy prevents the model from relying on surface-level keyword matching, forcing it to acquire a deep understanding of logical structures to distinguish similar discourses.

We evaluate two types of models, fine-tuned PLMs (BERT, RoBERTa) and LLMs (GPT-3.5/4/4o, GLM4). All of them have strong performance in Chinese and multilingual NLP tasks, respectively. We evaluate them in zero-shot and 5-shot settings. All results are averaged over three runs. Detailed hyper-parameters, hardware configurations, and prompt templates are provided in **Appendix F**.

4.3 Metrics

Rhetorical Figure Detection We use both pre-trained models and LLMs. To handle label imbalance, we apply negative sampling. We use Accuracy and Macro F_1 as evaluation metrics.

Claim Extraction We construct a balanced dataset by pairing claims with selected negative samples. We report Macro and Claim F_1 of the experiment.

Model	Setting	Acc.	Macro F_1
BERT-base-Chinese	fine-tuning	**0.9888**	**0.7927**
chinese-roberta-wwm-ext		0.9754	0.3883
GLM4	0-shot	0.1395	0.0108
GPT-3.5 Turbo		0.0627	0.0029
GPT-4o		0.2024	0.1422
GPT-4 Turbo		**0.2306**	**0.1994**
GLM4	5-shot	**0.6421**	**0.2125**
GPT-3.5 Turbo		0.2280	0.0056
GPT-4o		0.2106	0.1694
GPT-4 Turbo		0.2617	0.1842

Table 5: Results of rhetorical figure detection on CEDAR. **Bold** scores indicate the best performance in their group.

Stance Classification We report accuracy, overall F_1 as evaluation metrics.

Evidence Classification Input sequences concatenate the topic and claim, paired with evidence candidates to predict relevance. We use Accuracy and Macro F_1 as metrics.

Counter Speech Generation For the Counter Speech Generation task, we employ different evaluation strategies for LLMs. Specifically, we utilize n-gram overlap metrics, such as BERTScore and ROUGE-L, while adopting an LLM-as-a-judge framework. We select DeepSeek-V3 as the judge due to its superior performance in Chinese language processing. As for the LLM-as-a-judge, the evaluation is conducted from three perspectives: *Logical Coherence*, *Rebuttal Precision*, and *Persuasiveness and Eloquence*. Each dimension is rated on a 5-point scale. The final score is calculated by averaging these three ratings and normalizing the result to a 1-point scale. The prompt templates used for this evaluation are provided in Table 24 and Table 25.

5 Results and Discussion

5.1 Rhetorical Figure Detection

Table 5 presents our results using BERT-base-Chinese and chinese-roberta-wwm-ext. BERT-base-Chinese achieved 0.9888 accuracy and a 0.7927 macro-F1. In contrast, zero-shot GLM4 and GPT-3.5 Turbo performed poorly. However, with just five examples (5-shot), GLM4 improved to 0.6421 accuracy and 0.2125 macro-F1, underscoring the importance of targeted training data for fine-grained rhetorical classification. We also provide a case study of Rhetorical Figure Detection in **Appendix G**.

Model	Setting	Macro F_1	Claim F_1
BERT-base-Chinese	fine-tuning	0.8664	**0.7495**
chinese-roberta-wwm-ext		**0.9473**	0.5902
GLM4	0-shot	0.0343	0.1056
GPT-3.5 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
GPT-3.5 Turbo		0.3211	**0.4632**
GPT-4o		0.4058	0.3095
GPT-4 Turbo		**0.5191**	0.4470

Table 6: Results of claim extraction on CEDAR. **Bold** scores indicate the best performance in their group.

Model	Setting	Acc.	F_1
BERT-base-Chinese	fine-tuning	**0.6307**	0.6577
chinese-roberta-wwm-ext		0.5051	**0.6712**
GLM4	0-shot	0.0370	0.0192
GPT-3.5 Turbo		0.3989	0.0977
GPT-4o		**0.4421**	0.3364
GPT-4 Turbo		0.4219	**0.3827**
GLM4	5-shot	0.1465	0.0486
GPT-3.5 Turbo		0.4018	0.0960
GPT-4o		0.4942	**0.4389**
GPT-4 Turbo		**0.5401**	0.3880

Table 7: Results of stance classification on CEDAR. **Bold** scores indicate the best performance in their group.

Model	Setting	Acc.	F_1
BERT-base-Chinese	fine-tuning	0.9813	0.7752
chinese-roberta-wwm-ext		0.9091	0.1587
GLM4	0-shot	0.2315	0.0159
GPT-3.5 Turbo		0.5263	0.0236
GPT-4o		0.5393	**0.1608**
GPT-4 Turbo		**0.5417**	0.1220
GLM4	5-shot	0.5471	0.0168
GPT-3.5 Turbo		0.6043	0.0760
GPT-4o		**0.6413**	**0.2116**
GPT-4 Turbo		0.5701	0.1694

Table 8: Results of evidence detection performance on this dataset. **Bold** scores indicate the best performance in their group.

5.2 Claim Extraction

Table 6 presents the claim extraction results. BERT-base-Chinese achieved strong performance, demonstrating effective claim identification. In contrast, although chinese-roberta-wwm-ext attained a higher Macro F1 (0.9473), its lower Claim F1 (0.5902) suggests difficulty in precise claim detection. Large models performed poorly in zero-shot settings except GPT-4 Turbo. Surprisingly, few shot learning reduced the GPT-4 Turbo's performance. This decline is likely because the introduction of few-shot exemplars imposed specific semantic patterns that conflicted with the diverse linguistic expressions of claims, thereby misleading the models' judgment.

5.3 Stance Classification

Table 7 compares model performance under different settings. BERT-base-Chinese excelled after fine-tuning (Accuracy: 0.6307, F1: 0.6577), while chinese-roberta-wwm-ext achieved a higher F1-score (0.6712) but lower accuracy (0.5051). LLMs, particularly GLM4, struggled under zero-shot and few-shot conditions, with accuracy and F1 dropping to 0.0370 and 0.0192, respectively, indicating significant limitations in low-data scenarios.These findings reinforce the effectiveness of fine-tuned models over large pre-trained ones in resource-constrained tasks. The accuracy-F1 trade-off in chinese-roberta-wwm-ext suggests it captures certain classification aspects better but lacks overall prediction stability. Meanwhile, GLM4's poor performance highlights the necessity of fine-tuning large models in structured prediction tasks. The strong performance of the GPT family in Stance Classification may perhaps be attributed to the following reasons: (1) GPT family is more robust in understanding complex instructions, which leads to the GPT series outperforming GLM-4 in most tasks (except for claim extraction). (2) GPT family may have been exposed to more debate-related data during training, although most of it is in English. Its strong transfer learning ability allows it to apply this knowledge to Chinese tasks effectively.

5.4 Evidence Classification

Table 8 presents evidence detection results. BERT-base-Chinese demonstrated superior performance (Accuracy: 0.9813, F1: 0.7752), while chinese-roberta-wwm-ext showed moderate accuracy but lower F1 scores. LLMs exhibited slight F1 improvements from zero-shot to five-shot settings, suggesting the potential for enhancement through few-shot learning.

These results emphasize the impact of fine-tuning, particularly for evidence detection, where precise claim-evidence matching is crucial. Besides, The strong performance of BERT-base-Chinese confirms its adaptability, while the moderate results of chinese-roberta-wwm-ext suggest limitations in fine-grained evidence extraction. The improvements in large models with few-shot learning indicate that further adaptation is needed before they can match fine-tuned smaller models.

Stance	Model	Logical Coherence	Rebuttal Precision	Persuasiveness	Average Score
Pro	GLM-4	3.764	3.064	3.848	0.7296
	GPT-4o	**3.835**	**3.195**	**3.915**	**0.7370**
Con	GLM-4	3.806	2.923	3.711	0.6960
	GPT-4o	**3.914**	**3.167**	**3.826**	**0.7271**

Table 9: Results of counter speech generation (LLM-as-a-judge).

Stance	Model	BERTScore	ROUGEL
Pro	GLM4	0.0831	0.0013
	GPT-4o	**0.0953**	**0.0036**
Con	GLM4	0.0732	0.0032
	GPT-4o	**0.0832**	**0.0042**

Table 10: Results of counter speech generation of PLM's.

5.5 Counter Speech Generation

The experimental results are presented in two tables. The LLM-as-a-judge evaluation in Table 9 demonstrates that GPT-4o outperforms GLM-4 in logical coherence, rebuttal precision, persuasiveness, and average score across both Pro and Con stances, exhibiting superior overall performance. However, as shown in Table 10, automatic evaluation metrics such as ROUGE-L and BERTScore are relatively low. This discrepancy may be attributed to the fact that the model-generated counter speeches are significantly shorter than the reference texts (1.9k vs. 8k characters).

6 Conclusion

In this study, we introduce CEDAR, an innovative and comprehensive Chinese debate dataset specifically designed for computational argumentation tasks. This dataset significantly extends previous resources by incorporating multi-type annotations that support several key tasks, including claim extraction, stance classification, evidence identification, and counter speech generation. Additionally, CEDAR uniquely supports the task of identifying rhetorical figures within argumentative texts. This feature is particularly valuable for improving argument mining systems by providing more granular insights into the rhetorical strategies used in debates. By evaluating baseline argument mining methods on this dataset, we demonstrate its potential to advance the development of more robust argumentation models.

Limitations

It is necessary to conduct a more comprehensive evaluation of this dataset through more computational argumentation tasks. And it is worthwhile to introduce more debates into this Chinese dataset. We also intend to employ a broader range of prompt engineering techniques to experiment with the dataset for further research.

Ethics Statement

All data in this study are freely available to the public. We follow the policy of using these data without infringing on any copyright issues. All debate videos we collected were publicly released by competition organizers on publicly accessible video-sharing platforms. Manual annotation and human evaluation were carried out by collaborators within our research group. We completed manual annotation and human evaluation with the assistance of our research group students. Payment is adequate given to the participants. Their demographic information is presented in Table 26 in **Appendix I**.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), Outstanding Youth Fund Project of Inner Mongolia Autonomous Region (Grant No. 2025JQ010), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Major Science and Technology Projects of Inner Mongolia Autonomous Region (Grant No. 2025ZDSF0029), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2025YFDZ0011, 2025YFDZ0026, 2025YFSH0021, 2025YFHH0073), Hohhot Science and Technology Project (Grant No. 2023-Zhan-Zhong-1).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

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

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.

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.

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.

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.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zehua Duo, Jiang Li, Xiangdong Su, and Guanglai Gao. 2026. Flore: Integrating full lorentz group and directional offsets for effective knowledge graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20968–20976.

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

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*.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of us presidential campaign debates. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690.

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.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300.

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.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020. Detecting attackable sentences in arguments. *arXiv preprint arXiv:2010.02660*.

Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025. F^2bench: An open-ended fairness evaluation benchmark for llms with factuality considerations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046.

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.

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.

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)*.

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.

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*.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for chatgpt to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043.

Xuefeng Wei, Zhixuan Wang, Xuan Zhou, Zhi Qu, Hongyao Li, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Cartbench: Evaluating vision-language models on chinese art understanding, interpretation, and authenticity. *Preprint*, arXiv:2604.11632.

Jusheng Zhang, Kaitong Cai, Xiaoyang Guo, Sidi Liu, Qinhan Lv, Ruiqi Chen, Jing Yang, Yijia Fan, Xiaofei Sun, Jian Wang, Ziliang Chen, Liang Lin, and Keze Wang. 2025. Mm-cot:a benchmark for probing visual chain-of-thought reasoning in multimodal models. *Preprint*, arXiv:2512.08228.

Zijun Zhang. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee.

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.

Jingshi Zhou, Fei Cheng, Xiaojun Zhang, and Kaizhu Huang. 2025. A parallel corpus of chinese-english legal judgments with argumentative structure annotations. *DATA INTELLIGENCE*, 7(4):1291–1304.

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.

A Dataset Comparison

Table 12 presents the comparison of the overall statistics among existing argument mining datasets. 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. Table 11 compares rhetorical figures benchmarks.

B 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 the 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 identify and articulate the topics from the debates into clear, concise position statements.

Role Annotation: This involves annotating the roles of 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 anno-

Datasets	Dataset Properties			
	Rhetoric Types	Content Type	Spoken Language	Language
MM2012c	8	social media radio	✓	En
ChineseRhetoricalEmotion	5	literary works, textbooks, microblog, websites	✗	Zh
ASP	1	sermons and exam essays	✗	Latin/Zh
ConFiguRe	12	novel and proses	✗	Zh
Ours	**9**	**debate competitions**	✓	**Zh**

Table 11: Comparison between various rhetorical figures benchmarks.

# Datasets	# Topics	# Debates	# Claims	# Support	#Contest	# Evidence	# Rhetorical Figure	Content Domain	Language	Spoken Corpus
CDCD	32	326	976	-	-	-	-	Wikipedia articles	En	✗
CDED	39	274	1,734	-	-	3,057	-	Wikipedia articles	En	✗
Claims and Evidence	33	586	1,392	-	-	1291	-	Wikipedia articles	En	✗
IBM Stance Classification	55	-	2,394	1,324	1,070	-	-	Wikipedia articles	En	✗
IAM	123	1,010	4,890	2,613	2,277	9384	-	online forum articles	En	✗
IAC	10	390,704	-	-	-	282,478	-	articles	En	✗
ORCHID	476	1,218	-	-	-	-	-	debate competitions	Zh	✓
IBM 2019	200	400	2,440	-	-	-	-	Wikipedia articles	En	✓
US2016	2,754	823	623	-	-	433	-	debate	En	✓
QT30	30	30	6,181	5205	976	-	-	debate	En	✓
USElecDeb60To16	39	39	16,087	-	-	13,434	-	debate	En	✓
CEDAR (Ours)	318	600	6,251	5,001	1,250	5,173	1,126	debate competitions	Zh	✓

Table 12: Comparison of overall statistics between existing argument mining datasets. Topics represent unique topics in the corpus.

# Datasets	# Argumment	# Speech
CounterArguGen	111,900	-
ConcluGen	136,996	-
DebateSum	187,386	-
ORCHID	-	1,600
CounterArgumentGeneration	287,152	-
UKP-Corpus	5,032	-
ArgTersely	31, 197	-
CEDAR(Ours)	6,251	600

Table 13: Comparison of overall statistics between existing argument generation datasets. Topics represent unique topics in the corpus.

tations. 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 annotation accuracy and consistency.

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, other 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).

C 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 14: A simplified example from one debate.

As illustrated in Table 14, for each debate, our dataset consists of no less than the crucial elements: the debate topic, the roles of the speakers

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 Explanation	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 15: 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 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 Hypophora	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 Rhetorical Q.	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 16: 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.

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 17: 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.

(host and different debaters), and their utterances labeled with claims, stance (Pro/Con), evidence (Table 15), rhetorical figures (Table 16), 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 17 presents examples of debate topics and their types. Besides, Table 16 introduces a concrete rhetorical figure 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.

- 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.

- Hypophora : A figure of speech where the speaker poses a question and immediately answers it. It is used to guide the audience through a logical argument and to preemptively address potential objections. In debates, debaters use hypophora to control the flow of the discussion and ensure their reasoning is clear and persuasive.

- Rhetorical Question : A figure of speech in the form of a question asked to make a point rather than to elicit an answer, as the answer is already implied. In debates, rhetorical questions are used to challenge the opponent, highlight absurdities, or create a strong emotional impact, especially when the speaker wants to reinforce their viewpoint without needing an explicit response.

- 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 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 above-mentioned 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 Research), comprehensive explanations of claims (Evidence Type Explanation), and any evidence that

does not conform to the categories above (Evidence Type Others).

D Quality Control & Inconsistent Treatment

We filtered the videos according to the following criteria to ensure the dataset's quality. First, content risk control tools were applied to all speeches to avoid 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 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 Cohen's Kappa coefficient (Cohen, 1960) to assess the inter-annotator agreement between these two collective annotation groups. The resulting Kappa coefficient was 0.66, indicating a good agreement level between the two annotator groups. After a thorough review round, both reached 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.

E Data Source

Table 18 presents the data sources used for dataset construction.

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 18: 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.

F Details of Experimental Setup

We split our dataset into training, development, and testing sets in a topic-stratified 8:1:1 ratio. All tasks excepted counter speech generation are treated as sentence classification tasks. To address the class imbalance, we apply negative sampling (Mikolov et al., 2013) to the training set, maintaining a positive-to-negative ratio of 1:10.

For pre-trained models, we fine-tune BERT-base-chinese (Devlin, 2018) and chinese-roberta-wwm-ext (Cui et al., 2020) with a batch size of 16, a learning rate of 2e-5, and the AdamW optimizer (Zhang, 2018) (weight decay 0.01). The maximum sequence length is 128, and training runs for 10 epochs on a V100 GPU, selecting the best checkpoint based on development set performance. Each experiment is repeated three times with different random seeds, and results are averaged.

For LLMs, we use GLM4 (GLM et al., 2024), GPT-3.5 Turbo (OpenAI, 2022) and GPT-4o and GPT-4 Turbo (Achiam et al., 2023). We report the average performance over three independent API runs to ensure stability. In the few-shot setting, we provide five fixed labeled examples. Prompt templates are in **Appendix H**.

G A Case Study of Rhetorical Figure Detection

When evaluating model performance on the CEDAR dataset, we typically instruct models to provide direct predictions without requiring explanatory analysis. However, to more clearly illustrate how the evaluated model performs on the rhetorical figure detection task, we include examples where the model is asked to justify its choice.

It can be observed that in Table 19, when given the statement "Is it just that we are avoiding the problem?", GPT-3.5-turbo correctly identifies it as a rhetorical question, recognizing that it implies a pointed suggestion rather than seeking an answer. In contrast, GLM4 incorrectly labels it as a contrast, misclassifying the query as a contrast based on perceived thematic oppositions and overemphasizing implied oppositions not explicitly expressed. This demonstrates that some models struggle to distinguish surface rhetorical forms from deeper semantic inferences, underscoring the nuanced reasoning required by CEDAR.

H Prompt Templates in LLM Experiments

Rhetorical Figure Detection
Table 20 presents rhetorical figure Detection prompts.

Claim Extraction
Table 21 presents claim extraction prompts.

Stance Classification
Table 22 shows stance classification prompts.

Evidence Classification
Table 23 shows Evidence classification prompts.

Counter Speech Generation
Table 24 and 25 shows claim extraction prompts.

I Demographic Characteristics

See Table 26 for details.

Model	Predicted Label	Explanation
GPT-3.5 Turbo	反　问(Rhetorical Question)	Identified as a rhetorical question because it makes a pointed suggestion to provoke thought rather than seeking an answer.
GLM-4	对比(Contrast)	Interpreted as a contrast by juxtaposing active problem-solving with passive avoidance to emphasize responsibility.

Table 19: Model response comparison on rhetorical device identification for the statement "Is it just that we are avoiding the problem?"

Method	Prompt (Chinese Original)	Prompt (English Translation)
0-shot	给定辩题和辩论语句，判断该语句使用了哪种修辞手法，或者不含修辞。你只能输出以下类别之一：其他, 反问, 排比, 引用, 夸张, 设问, 对比, 对偶, 或O。 不要输出无关内容。 辩题：{topic} 语句：{statement} 输出：	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: Others, Rhetorical Question, Parallelism, Quotation, Exaggeration, Hypophora, Contrast, Antithesis, or O (No rhetorical device). No irrelevant content should be output. Debate Topic: {topic} Statement: {statement} Output:
5-shot	给定辩题和辩论语句，判断该语句使用了哪种修辞手法，或者不含修辞。你只能输出以下类别之一：同上类别。不要输出无关内容。 辩题：天性基因与后天环境哪个更重要 语句：是不是我们只是在回避问题呢? 输出：反问 辩题：人生追求更应看重道德成就还是事业成就 语句：我们要呼唤他们，我们要鼓励他们，我们要用爱让他把咖啡煮得更香。 输出：排比 辩题：打卡式旅游的兴起是丰富还是贫乏了旅游的文化内涵 语句：俗话说得好，一千个读者眼中有一千个哈姆雷特。 输出：引用 辩题：知识付费能否缓解年轻人的焦虑 语句：这些标题是什么？他们在说如果你现在不学，就太晚了。 输出：设问 辩题：短视频流行提升还是降低了当代人的认知能力? 语句：这就是对方想要引用的短视频。 输出：O 辩题：{topic} 语句：{statement} 输出：	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: Same as above. 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 20: Rhetorical figure detection on argumentative corpus for LLM prompting methods. The middle column shows the Chinese prompts used in experiments, and the right column provides the corresponding English translation.

Method	Original Prompt (Chinese)	Prompt (English Translation)
0-shot	给定具体的辩论主题和相关陈述，自动从文章中提取观点。判断每个陈述是相关观点(C) 还是非观点(O)。只输出C (claim) 或O (not claim)。 不要输出无关内容。 辩题：{topic} 语句：{statement} 输出：	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	给定具体的辩论主题和相关陈述，自动从文章中提取观点。判断每个陈述是相关观点(C) 还是非观点(O)。只输出C (观点) 或O (非观点)。不要输出无关内容。 示例： 辩题：个人利益和集体利益能否共存 语句：这恰恰证明了我们今天的观点，即个人利益和集体利益是可以共存的。 输出：C 辩题：个人是否有责任抵制消费主义 语句：不，你能举个例子吗? 输出：O 辩题：无情还是不忠更可悲 语句：但他也有忠诚，为什么? 输出：O 辩题：对于当代青年，加快还是减慢生活节奏更有利于成长 语句：正方认为，今天的成长只有一个意义，那就是我们需要获得更多的技能，创造更多的效益。 输出：C 辩题：艺术家需要更出世还是入世的心态 语句：感谢反方三辩，现在我邀请正方三辩进行总结。 输出：O 辩题：{topic} 语句：{statement} 输出：	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 21: Claim extraction task prompts for LLM prompting methods. The middle column shows the original Chinese prompts, and the right column provides the English translation.

Method	Prompt (Chinese Original)	Prompt Translation (English)
0-shot	判断以下辩论语句的立场，并输出：1（正方）、-1（反方）或0（无明显立场）。 辩题：{topic} 语句：{statement} 立场：	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	判断以下辩论语句的立场，并输出：1（正方）、-1（反方）或0（无明显立场）。示例： 辩题：个人利益和集体利益能否共存 语句：这恰恰证明了我们今天的观点，即个人利益和集体利益是可以共存的。 立场：1 辩题：互联网经济是促进还是阻碍匠人精神的发展 语句：互联网经济阻碍了匠人精神的发展。 立场：-1 辩题：逆境还是顺境更有利于个人成长 语句：所以我们需要知道，无论对谁而言，逆境都是有利于我们成长的。 立场：-1 辩题：人工智能生成的艺术品是否算作艺术 语句：我们认为它可以拓展艺术的边界，我们有证据支持这一点。 立场：1 辩题：放下还是坚持更难 语句：因此，我们认为坚持更难。 立场：-1 辩题：{topic} 语句：{statement} 立场：	Determine the stance of the following debate utterance and output 1 (proposition side), -1 (opposition side), or 0 (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 **Debate Topic**: {topic} **Statement**: {statement} **Stance**:

Table 22: Stance classification task prompts for LLM prompting methods. The middle column shows the original Chinese prompts, and the right column provides the English translation. The label 0 (neutral) is utilized exclusively for filtering non-argumentative utterances, such as procedural remarks or conversational fillers, which does not convey a specific stance.

Method	Prompt (Chinese Original)	Prompt Translation (English)
0-shot	给定辩题和语句，判断该语句是否为辩论背景下支持该辩题的论据（证据）。仅输出以下类别之一：案例, 解释, 研究, 专家, 其他, 或O (非证据)。 辩题：{topic} 语句：{statement} 输出：	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: Case, Explanation, Research, Expert, Others, or O (Not Evidence). Topic: {topic} Statement: {statement} Output:
5-shot	给定辩题和语句，判断该语句是否为论据。仅输出以下类别之一：同上类别。 辩题：人在为了生存的情况下，是不是可以不择手段 语句：春秋时期，宋国被楚国包围，粮食吃光后，百姓交换孩子作为食物来生存。 输出：案例(Case) 辩题：刻骨铭心的感情更应该忘记还是铭记 语句：铭记刻骨铭心的感情，让一个人的情感体验有起伏。正如身体的生存依靠心电图的起伏一样，灵魂离不开刻骨铭心感情的定位和证明。 输出：解释(Explanation) 辩题：逆境还是顺境更有利于个人成长 语句：今天，我们说一项犯罪学研究表明，来自破碎家庭的青少年的犯罪率比正常家庭的高出三倍。 输出：研究(Research) 辩题："碰瓷式"维权该不该受法律保护 语句：中国消费者权益保护法学研究会秘书长表示，80%的维权案例针对的是毫无意义的商标标识，如杏仁和杏仁核外国翻译中的同义词。这些投诉不仅浪费司法资源，还危害市场秩序。 输出：专家(Expert) 辩题：门当户对是不是过时的婚姻观 语句：你有什么权利判断一个人选择门当户对是错误的呢? 输出：O 辩题：{topic} 语句：{statement} 输出：	Given a topic and statement, determine whether the statement is a premise (evidence) supporting the topic. Only output one: Case, Explanation, Research, Expert, Others, or O. **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. **Output**: 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, the soul cannot be separated from the positioning and proof of deeply felt emotions. **Output**: 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. **Output**: 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... These complaints not only waste judicial resources but also harm market order. **Output**: 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? **Output**: O **Debate Topic**: {topic} **Statement**: {statement} **Output**:

Table 23: Evidence classification task prompts for LLM prompting methods. The middle column shows the original Chinese prompts, and the right column provides the English translation.

Method	Prompt (Chinese Original)	Prompt Translation (English)
0-shot	针对辩论发言生成反驳发言。 辩论发言：{正方或反方中文发言内容} 反驳发言：	Generate counter speeches in response to debate speeches. Debate Speech: {Pro-side or Con-side Chinese speech content} Counter speech:

Table 24: Counter speech generation task prompts for n-gram overlap metrics. The actual prompts are conducted in Chinese (middle column), and the English translation is provided in the right column.

Category	Prompt (Chinese Original)	Prompt (English Translation)
Role	你是一位专业的辩论赛裁判，具备严密的逻辑思维能力和深厚的中文修养。你需要根据提供的"原论点"和"参考反驳"，对"待评价的反驳文本"进行客观的打分，范围为0-5。	You are a professional debate judge with rigorous logical thinking ability and profound knowledge of Chinese language and culture. You need to objectively score the "Generated Counter-speech" based on the provided "Original Claim" and "Reference". The score range is 0-5.
Evaluation Dimensions	逻辑连贯性：论证是否严密？是否存在逻辑谬误？结构是否清晰？ 反驳针对性：是否精准回击了原论点中的核心观点？ 说服力与话语质量：语言是否具有感染力？是否模拟了专业辩论的口吻？论据支撑是否充分？	**Logical Coherence:** Is the argument well-structured? Are there any logical fallacies? Is the structure clear? **Rebuttal Precision:** Does the rebuttal accurately address the core points of the original claim? **Persuasiveness and Eloquence:** Is the language persuasive and impactful? Does it simulate the tone of a professional debate? Is the argument well-supported?
Input Data	**[**原论点**]:** {original_claim} **[**参考内容**]:** {reference} **[**待评价内容**]:** {generated_cs}	**[Original Claim]:** {original_claim} **[Reference]:** {reference} **[Generated Counter-speech]:** {generated_cs}
Output Format	逻辑连贯性得分： 反驳针对性得分： 说服力与话语质量得分：	Score of Logical Coherence: Score of Rebuttal Precision: Score of Persuasiveness and Eloquence:

Table 25: Prompts used for the LLM-as-a-judge in the Counter Speech Generation task. The actual prompts were conducted in Chinese (middle column), and the English translations are provided for reference (right column).

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

Table 26: The demographic information of the annotators. All evaluators have at least a postgraduate-level education.