SAD: A Large-Scale Strategic Argumentative Dialogue Dataset

Yongkang Liu[1] **Jiayang Yu**[2] **Mingyang Wang**[3,6] **Yiqun Zhang**[2] **Ercong Nie**[4]
Shi Feng[1*] **Daling Wang**[2] **Kaisong Song**[5] **and Hinrich Schütze**[3,6]
[1]School of Computer and Communication Engineering,Northeastern University, Qinhuangdao
[2]School of Computer Science and Engineering, Northeastern University, Shenyang
[3]CIS, LMU Munich, Germany
[4]School of Foreign Languages, Shanghai Jiao Tong University
[5]Alibaba Group, Hangzhou, China
[6]Munich Center for Machine Learning (MCML), Germany
misonsky@163.com,mingyang.wang2@de.bosch.com

Abstract

Argumentation generation has attracted substantial research interest due to its central role in human reasoning and decision-making. However, most existing argumentative corpora focus on non-interactive, single-turn settings, either generating arguments from a given topic or refuting an existing argument. In practice, however, argumentation is often realized as multi-turn dialogue, where speakers defend their stances and employ diverse argumentative strategies to strengthen persuasiveness. To support deeper modeling of argumentation dialogue, we present the first large-scale **S**trategic **A**rgumentative **D**ialogue dataset, SAD, consisting of 392,822 examples. Grounded in argumentation theories, we annotate each utterance with five strategy types, allowing multiple strategies per utterance. Unlike prior datasets, SAD requires models to generate contextually appropriate arguments conditioned on the dialogue history, a specified stance on the topic, and targeted argumentation strategies. We further benchmark a range of pretrained generative models on SAD and present in-depth analysis of strategy usage patterns in argumentation.

Figure 1: An example of SAD. The speakers on the left are opponents of the topic, and the speakers on the right are supporters of the topic. The argumentative strategies used are highlighted in blue. Strategic sentences are emphasized using other colors.

1 Introduction

As fundamental capabilities of human intelligence, argumentation and debating play an important role in everyday activities that involve reasoning, decision making and persuasion (Bar-Haim et al., 2021). This capability is increasingly important in the era of large language models (LLMs): beyond producing fluent text, LLMs are expected to justify claims, weigh competing evidence, respond to challenges, and remain consistent under scrutiny, abilities that are naturally expressed in argumentative dialogue. Argumentation can be viewed as a structured exchange between supporters and opponents on a topic, typically unfolding over multiple turns. The key to argumentation is the ability to effectively present arguments and contest an argument by presenting a counter-argument. The purpose of argumentation is to recognize differences and make the best judgment or decision through information exchange.

Therefore, argument generation has stimulated widespread interest in the research community (Durmus and Cardie, 2019; Gurevych et al., 2016; Hua et al., 2019; Ruggeri et al., 2023), with online argumentation being a major focus (Tan et al., 2016; Durmus and Cardie, 2018). Two widely studied tasks are *counter-argument generation* – producing a rebuttal to a given claim, optionally conditioned on a topic (Hua et al., 2019; Alshomary et al., 2021; Wu et al., 2022) – and *topic-conditioned argument generation* which pro-

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36144–36164
July 2-7, 2026 ©2026 Association for Computational Linguistics

duces arguments relevant to a given topic (Sato et al., 2015; Schiller et al., 2021; Alshomary and Wachsmuth, 2023; Goloviznina et al., 2023). However, most existing work remains **single-turn**, leaving a substantial gap to real-world argumentation, which unfolds through interactive, multi-turn exchanges. Although the argument corpora of scientific papers include multi-turn Q&A, each question-answer pair remains independent and non-interactive, essentially functioning as single-argument arguments (Wu et al., 2022; Ruggeri et al., 2023). Recent attempts employ multi-agent systems to synthesize multi-turn argumentation dialogues (Li et al., 2024) , but such data does not reflect real-world interactions and is small in scale. In contrast, we believe that real argumentation relies heavily on strategic behavior across turns to enhance persuasiveness.

As shown in Figure 1, we present a real-world multi-turn argumentative dialogue on the topic *"I believe grade school English should stop teaching 'classics' and focus on fostering love of reading first."*. The opponent believes that *"teaching classics"* and *"fostering love of reading"* are not contradictory by giving examples. Then the supporter states views to rebut by giving counter-examples and posing rhetorical question. Through this multi-turn exchange, both sides iteratively respond to each other's claims and strengthen their positions. This example highlights three key characteristics of real argumentation: (1) it is typically an interactive dialogue between supporters and opponents with opposing stances on a topic; (2) the dialogue history provides essential context and evidence for generating coherent responses; and (3) speakers employ diverse argumentative strategies across turns to enhance persuasiveness. However, the lack of comprehensive corpora that reflect these real-world properties remains a major bottleneck for advancing argument generation.

To fill this gap, we annotate a large-scale **S**trategy **A**rgumentation **D**ialogue dataset (SAD) based on the crawled **ChangeMyView** corpus. We construct dialogue structure by tracing the interactive reply chain, yielding 722,812 utterances grouped into 392,822 dialogue examples across 20,619 topics. Guided by argumentation theories (Freeley and Steinberg, 2013; Weston, 2018), linguist experts define five strategy types (Table 9): **Question**, **Causality**, **Example**, **Analogy**, and **Statement**. Each utterance is annotated with its stance (support vs. opposition) and may be assigned multiple strategy labels. Based on SAD, we formulate a strategy-conditioned generation task, **P(Argument | History, Stance, Topic, Strategy)**, to better reflect real-world argumentative dialogue. To the best of our knowledge, SAD is the first large-scale, strategy-based dataset for multi-turn argumentation. To automatically assess persuasiveness, we further train a persuasiveness evaluator using the number of likes on an argument as supervision, and show that its scores align well with human judgements. Experimental results demonstrate that utilizing strategies improves generation quality in fluency, coherence, topicality, and persuasiveness. Finally, we present an in-depth analysis of strategy usage patterns and evaluator behavior, offering insights for building more effective argumentation dialogue systems.

Our contributions are summarized as follows:

- **Dataset:** We construct SAD, a large-scale, high-quality dataset for strategy-aware, multi-turn argumentation dialogue. To the best of our knowledge, SAD is the first strategy-based multi-turn argumentation dataset at this scale.

- **Task Formulation:** We propose a new argument generation task that better reflects real-world scenarios, requiring models to generate response arguments conditioned on the dialogue history, topics, stances, and specified strategies. SAD also supports related tasks such as general multi-turn argument generation, strategy identification, and argument retrieval.

- **Automatic Evaluation:** We train an automatic evaluator for argument persuasiveness and show that its judgments achieve high agreement with human ratings.

- **LLM Benchmarking:** We benchmark multiple LLMs and find substantial room for improvement in multi-turn argumentation. Experimental results show that incorporating argument strategies can significantly enhance training quality and generation performance.

2 Related Work

The Argument Generation Task is to design systems to automatically generate a persuasive text to support or refute a specific claim, opinion, or position. Argument datasets serve as a foundation for the design and evaluation of argument systems.

dataset	Task	Form	Number	Real World	Strategy	Multi-turn
Gigaword (Sato et al., 2015)	argument generation	R(Y|T)	50	✓	✗	✗
Idebate (Wang and Ling, 2016)	argument generation	P(Y|A,T)	18,363	✓	✗	✗
MedicalRecords (Green, 2017)	argument generation	Scheme-based	-	✓	✗	✗
Arg-Microtexts (Wachsmuth et al., 2018)	argument synthesis	-	576	✗	✓	✗
CMV (Hua and Wang, 2018)	argument generation	P(Y|T, KB)	224,553	✓	✗	✗
Rank-30k (Gretz et al., 2020)	argument generation	P(Y|A,T)	10,669	✓	✗	✗
CE2.3k (Gretz et al., 2020)	argument generation	P(Y|A,T)	1,489	✓	✗	✗
LN55k (Gretz et al., 2020)	argument generation	P(Y|A,T)	30,000	✗	✗	✗
Kialo (Al Khatib et al., 2021)	argument generation	P(Y|T, KB)	82,728	✓	✗	✗
Args.me (Ajjour et al., 2019)	argument generation	P(Y|T, KB)	30,748	✓	✗	✗
CMV (Alshomary et al., 2021)	argument generation	P(Y|T,A)	111,900	✗	✗	✗
PRRCA (Wu et al., 2022)	argument generation	P(Y|D,A)	4,764	✓	✗	✗
ArgEssay (Bao et al., 2022)	argument generation	P(Y|T)	11,282	✓	✗	✗
AspectCorpus (Goloviznina et al., 2023)	argument generation	P(Y|T,S,*Asp*)	418	✗	✗	✗
AIDebater (Aid, 2023)	argument generation	P(Y|T,A)	-	✗	✗	✗
ArgSciChat (Ruggeri et al., 2023)	argument generation	P(Y|T,D)	41	✗	✗	✗
DEBATUNE (Li et al., 2024)	argument generation	P(Y|S,A,T)	7,100	✗	✗	✓
SAD	argument generation	P(Y|T,H,S,[R])	392,822	✓	✓	✓

Table 1: Comparison between our dataset and other datasets. The corresponding abbreviations involved are as follows: Y–Response, A–Argument, T–Topic, KB–Knowledge Base, S–Stance, Asp–Aspect, D–Document, H–History, R–stRategy.

There have been several datasets in the field of argument generation using empirical methods that focus on generating arguments based on topics. As shown in Table 1, Gigaword (Sato et al., 2015) performs a secondary annotation on a large news text corpus to make the dataset suitable for the argument retrieval task. Iargument (Wang and Ling, 2016) is a an argumentation dataset from *iargument.org*, which is a Wikipedia-style website for gathering pro and con arguments on controversial issues. MedicalRecords (Green, 2017) is from medical records about the patient and the patient's biological family, and connections, a list of facts or principles of genetics. Arg-Microtexts (Wachsmuth et al., 2018) is designed to provide crisp argumentation in a "pro and con" manner. The CMV dataset (Hua and Wang, 2018), collected from a Reddit subcommunity, consists of 224,553 examples with 305,475 relatively high-quality replies. The CE2.3k dataset (Gretz et al., 2020) contains 1,489 examples with 2.3k manually curated claims extracted from Wikipedia. Each topic of CE2.3k is mapped to a corresponding Wikipedia title. The LN55k dataset (Gretz et al., 2020), collected from a corpus of newspaper articles, contains 224,553 examples, including 55,024 manually curated claims. Kialo (Al Khatib et al., 2021) contains 82,728 discussion from kialo.com, which is a argument portal. Args.me (Ajjour et al., 2019) includes 30,748 arguments from *debatewise.org*, *debate.org*, *debatepedia.org* and *idebate.org*. Alshomary et al. (2021) extends CMV (Jo et al., 2020) by further collecting the quoting sentences from the comments (i.e., the counter-arguments). PRRCA (Wu et al., 2022) is a peer review and rebuttal counter-arguments by collecting publicly available peer review contents and the submission information from *openreview.net*. ArgEssay (Bao et al., 2022) is a large-scale argumentative essay generation dataset. AspectCorpus (Goloviznina et al., 2023) is the first russian-language corpus of arguments with annotated aspects. AIDebater (Aid, 2023) contains counter-argument generation corpus from the ChangeMyView forum and topic argument generation from argument competition. ArgSciChat (Ruggeri et al., 2023) is a document conversation QA corpus based on scientific papers. DEBATUNE (Li et al., 2024) synthesizes a multi-turn argumentative dialogue dataset, filling the gap in multi-turn argument benchmarks. These benchmarks have greatly promoted the development of argument generation tasks.

Most existing argumentation corpora do not focus on the multi-turn conversational ability of argumentation or the controllability of generation. Although DEBATUNE (Li et al., 2024) is a multi-turn argumentative dialogue dataset, it is a synthetic non-real-world corpus. To fill this gap, this paper constructs a large-scale strategy-based multi-turn dialogue argument dataset SAD.

3 Dataset Creation

This section deals with the process followed for the creation of the dataset for multi-turn argumentative dialogue dataset. We have broadly split this into three parts: dialogue collection, stance annotation, strategy annotation and data filtering.

3.1 Procedure for Argument Collection

We draw our data from the Reddit community r/ChangeMyView (CMV), which is dedicated to fostering open discussions on a wide range of contested issues. Specifically, CMV is structured as discussion threads, where the original post starts with a viewpoint on a controversial topic, followed by detailed supporting arguments, then other users reply with counter-arguments. However, not all participating users counter topics they are interested in; many also express support. During the debate, users are allowed to reply to all posts. Therefore, a topic could have multiple threads of discussion. As shown in Figure 11 (Appendix), multiple discussion threads are identified under the topic *In a clinical setting, only medical doctors should be called "doctor"*. Each thread represents a multi-turn debate dialogue. Concretely, each root-to-reply path in the CMV discussion tree is treated as one dialogue example, where the preceding utterances form the dialogue history and the last utterance is the response to be generated or evaluated.

Strategy\Consis.	1/3	2/3	3/3
Question	0.984	0.950	0.912
Causality	0.975	0.920	0.827
Example	0.993	0.952	0.924
Analogy	0.950	0.852	0.784
Statement	0.998	0.974	0.952
Overall	0.972	0.910	0.865

Table 2: Consistency proportion of strategy annotation. 1/3 means consistency with at least one annotator, 2/3 means consistency with at least two annotators, and 3/3 means consistency with all three annotators.

High-quality arguments make CMV possible for constructing multi-turn argumentative dialogue datasets. We crawl all discussion trees created at any time from March 2016 to September 2020. We do not account for data omissions resulting from interrupted crawling. The crawled data contains 26,726 discussion trees, and 17,652,244 nodes. Based on the discussion threads related to each topic, we construct a multi-turn debate dialogue dataset. To maintain data quality, we performed data filtering. First, we apply rule-based filtering to remove web links, meaningless repeated stop words, and other non-informative content. Posts with short utterances often fail to provide valuable arguments to support their points, resulting in low-quality examples. Statistical analysis shows that these posts typically contain fewer than 10 words per utterance. Therefore, posts shorter than 10 words are not considered. Specifically, given a multi-turn argument example D=$\{u_1, u_2, \ldots, u_i, \ldots, u_n\}$, if the sentence length of u_i is less than 10, then $\{u_i, u_{i+1}, \ldots\}$ will be removed from the example, and finally D represents $\{u_1, u_2, \ldots, u_{i-1}\}$. After filtering, each surviving example preserves the original chronological order of the utterances in that path. This dataset is available at `https://github.com/yuhkalhic/SAD`.

Total Utterances	722812.00
Total Examples	392,822.00
Topics	20,619
Avg. Utterance Length (tokens)	119.79
Min. Utterance Length (tokens)	11.00
Max. Utterance Length (tokens)	2399.00
Max. Turns	11.00
Min. Turns	2.00
Avg. Turns	3.69
Max. Dialog Length (tokens)	9,784
Min. Dialog Length (tokens)	24
Avg. Dialog Length (tokens)	609.97

Table 3: Statistical information of the argument dataset

3.2 Stance Annotation

We observe that in open-ended dialogue debates, participants' stances do not strictly alternate, which allows them to engage more freely and spontaneously, resulting in greater interactivity. This irregular progression of stances creates a more natural interactivity dynamic, but it also causes automatic annotation to fail. We then present the process of annotating the stances. Specifically, we recruit and train five workers to annotate the answers following our guideline[1]. In this annotation task, workers are given topic–utterance pairs and asked to label the relationship between them, where 1 indicates support and 0 indicates opposition. The annotation unit is a single utterance with respect to the original topic rather than the immediately preceding reply, and all five annotators label each utterance independently. The label with the highest number of votes is selected as the final answer. The Fleiss' Kappa score for the five annotators is 0.78, indicating a high level of consistency and reliability in annotations.

[1]All annotators in this work are compensated for 80 in CNY per hour, which is reasonable given the difficulty of the annotation task and the mean income of urban residents in China.

3.3 Strategy Annotation

Argumentation strategies play a crucial role in determining the persuasiveness of an argument, when used effectively, they can dramatically enhance its overall impact. To facilitate further research on strategies in argument support, we then present the process that we annotate the argumentation strategies. Argumentation strategies have been systematically studied (Freeley and Steinberg, 2013; Weston, 2018). As shown in Table 9, we identify 5 strategies: Question, Causality, Example, Analogy and Statement according to (Freeley and Steinberg, 2013; Weston, 2018). We observe that an argument may employ multiple argumentation strategies. To ensure the quality of the annotations, we annotate each argument and each strategy separately. Specifically, we recruit and train five workers with debating experience to annotate the argumentation strategies. Each annotator's hourly compensation is the same as the previous stance annotation. In this task, the workers are shown an argument and ask to determine whether the argument employs a particular strategy. We formulate this stage as five parallel binary decisions for Question, Causality, Example, Analogy, and Statement instead of a single multiclass label, which allows multiple strategy labels to be assigned to the same utterance. The workers are allowed to ignore the arguments that do not match the definition of any strategy, which would be automatically labeled as "None". An utterance is assigned "None" only when all five strategy decisions are negative.

3.4 Annotation Quality Control

The stance annotation is relatively simple, and we provide Fleiss' Kappa to demonstrate the reliability of the annotation. For strategy annotation, we implement more stringent quality control than stance annotation. Before annotating, workers were required to go through the guideline and the provided examples. To ensure the quality of annotations, we require workers to annotate 200 examples before the formal annotation, which are revised by debate professionals for feedback. We repeat the above process until the workers are able to annotate the cases almost correctly. After annotation, to check This pilot-and-feedback stage is used to calibrate annotators before the formal annotation round starts. the quality of labels, we randomly sample 200 arguments, give them to three examiners to pick out incorrect labels, and calculate the consistency proportion. Results are shown in Table 2. More than 97.2% of the strategy labels are consistent with at least one examiner, and more than 91.0% of the strategy labels are consistent with at least two examiners, indicating the reliability of strategy annotation.

4 Corpus Analysis

The statistical information of the argument dataset is given in Table 3. The dataset contains 392,822 dialogue instances and a total of 722,812 utterances, covering 20,619 distinct topics, which indicates substantial scale and topical diversity. The average utterance length is 119.79 tokens, with a minimum of 8 tokens and a maximum of 2,399 tokens, reflecting a broad range of argumentative expressions from brief statements to highly detailed arguments. In terms of dialogue structure, each dialogue consists of 2 to 11 turns, with an average of 3.69 turns. Correspondingly, dialogue lengths vary widely, ranging from 24 to 9,784 tokens, with an average length of 609.97 tokens. These statistics demonstrate that the dataset captures multi-turn argumentative dialogues of varying length and complexity, making it well-suited for training and evaluating models for argumentation understanding and generation.

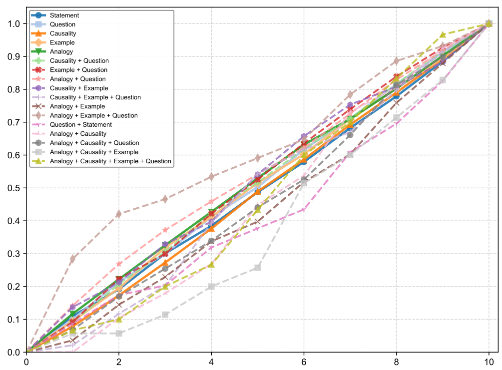

Figure 2: Cumulative distribution of strategies. The x-axis denotes the relative position in dialogue, and the y-axis denotes the cumulative proportion.

4.1 Strategy Sequence Analysis

Cumulative Distribution of Strategies Figure 2 presents the cumulative distribution of strategy occurrences across relative dialogue positions, revealing a pronounced discrepancy among different strategies. To facilitate a clearer analysis of strategy distributions, we partition the dialogue content into three equal stages—beginning, middle, and ending—since strategy usage demonstrates distinct characteristics across different phases of the dialogue, for we observe from our data that most strategies have different functions and characteristics among the beginning, middle and ending part. For instance, **Statement** and **Question** serve as the dominant strategies in the early stage, showing that the debate primarily concentrates on establishing the argumentative focus and introducing initial claims.

As the dialogue progresses into the middle stage, the distribution of **Causality** and **Example** shifts toward a steady linear growth. This phase represents the "logical engine" of the debate. The stabilization of these slopes reflects a consistent commitment to evidentiary support; participants move beyond mere assertions to establish mechanistic links (**Causality**) and empirical grounding (**Example**). The high density of overlapping curves in this region suggests a strategic equilibrium where diverse argumentative modes are deployed in parallel to sustain the dialectical momentum. At the end-stage, the strategy is not to introduce new topics but to leverage integrative mechanisms that weave scattered logical threads into a comprehensive, summarizing stance, thereby maximizing persuasiveness before the dialogue concludes. This stage typically employs a combination of strategies, such as the **Analogy** + **Causality** + **Example** + **Question**.

The analysis suggests a clear evolutionary trajectory of strategic complexity. The dialogue transitions from structural simplicity (aimed at orientation) to logical robustness (aimed at substantiation), and finally to syntactic complexity (aimed at integration). This distribution confirms that participants do not deploy strategies randomly. Rather, they adhere to a latent "argumentative pulse" that prioritizes foundational clarity at the outset and rhetorical sophistication at the conclusion.

4.2 Strategy Transition

To provide deeper insights into strategy utilization, we conduct a comprehensive visualization analysis of strategy transitions across the first five conversational turns. As illustrated in Figure 3, the results reveal several noteworthy patterns, demonstrating that strategy transitions are far from random and instead follow a highly stable yet evolutionary framework. Transitions from declarative sentences to declarative sentences consistently account for the largest proportion, indicating that expressing opinions through statements remains the most prevalent mode of debate.

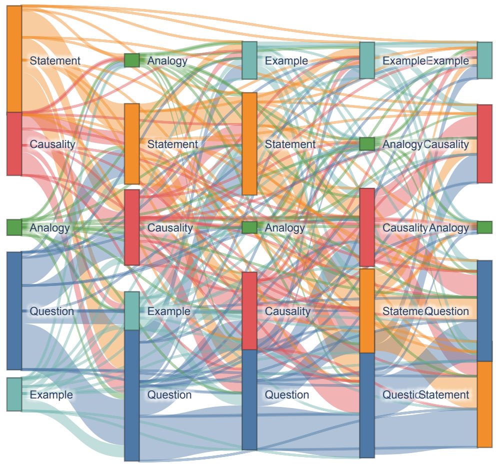

Figure 3: Visualization of the strategy flow patterns.

Causality emerges as a central connective strategy, exhibiting strong transitions both from and to Statement and Question. This indicates that causal reasoning often functions as a bridge between expressing claims and soliciting justification, reinforcing the logical coherence of arguments. Similarly, while Example and Analogy appear less frequently as standalone strategies, they are commonly integrated into later turns, either remaining stable across turns (e.g., Example → Example) or combining with core strategies to enhance explanatory strength.

Statements, Causality, and Question are commonly used strategies in argumentation, and they play a vital role in advancing the debate. Overall, the transition structure demonstrates that strategy usage in argumentative dialogue follows a non-random yet evolutionary trajectory. Speakers tend to maintain core strategies while progressively incorporating additional rhetorical functions, resulting in increasingly complex and information-rich argumentative moves.

5 Experiments

The SAD dataset is divided into a training set, a validation set, and a test set in an 8:1:1 ratio. After dataset construction, we randomly partition the dialogue examples once, keep this split fixed for all reported experiments, and draw the human-evaluation subset only from the held-out test partition. Llama3.1-8B and Qwen3-8B are further trained on SAD in the fine-tuning study; all other open-source models and all closed-source APIs are evaluated in the zero-shot setting. The experi-

Type	Model	Strategy	Relev.	Coher.	Persua.	Smooth.	Score
Open	Llama3.1-8B	wo.	3.00	3.38	2.79	3.98	0.6378
		w.	**3.18**	**3.41**	**2.83**	**3.99**	**0.6482**
	Qwen3-8B	wo.	4.14	4.09	**3.85**	4.27	0.6623
		w.	**4.22**	**4.18**	3.72	**4.28**	**0.6756**
	Hunyuan-7B	wo.	3.70	3.55	**3.40**	3.57	0.6253
		w.	**3.74**	**3.60**	3.35	**3.63**	**0.6340**
	GLM-4-9B	wo.	**4.01**	**4.15**	**3.75**	**4.29**	**0.6732**
		w.	3.66	3.83	3.39	4.07	0.6473
	Baichuan2-7B	wo.	3.28	**3.58**	2.78	3.88	0.6251
		w.	**3.44**	3.54	**2.88**	**3.90**	**0.6260**
	Ministral-8B	wo.	3.77	3.91	3.41	4.09	0.6583
		w.	**3.78**	**3.93**	**3.44**	**4.17**	**0.6604**
	MiMo-7B	wo.	3.98	4.03	3.63	4.19	0.6707
		w.	**4.06**	**4.06**	**3.65**	**4.20**	**0.6753**
	Yi-1.5-9B	wo.	3.48	3.85	3.32	4.07	0.6496
		w.	**3.52**	**3.87**	**3.40**	**4.10**	**0.6511**
Closed	GPT-4.1-nano	wo.	4.26	**4.22**	**3.84**	**4.35**	0.6762
		w.	**4.46**	4.18	3.80	4.32	**0.6947**
	DeepSeek-V3.2	wo.	**4.51**	4.42	4.09	4.35	0.7268
		w.	4.48	**4.43**	**4.16**	**4.44**	**0.7757**
	Claude-3-haiku	wo.	3.16	3.80	2.90	**4.28**	0.6742
		w.	**3.74**	**3.99**	**3.31**	4.26	**0.6952**
	Gemini-2.5-flash	wo.	3.95	4.12	3.63	**4.26**	0.6592
		w.	**4.09**	**4.17**	**3.68**	4.19	**0.6890**

Table 4: Performance comparison of various large language models on the SAD test set using zero-shot testing. The Relevance (Relev.), Coherence (Coher.), Persuasiveness (Persua.), and Smoothness (Smooth.) are evaluated using GPT-4.1 as the evaluator. All experimental results are shown in Table 10. The score is the persuasiveness rating given by the evaluator we trained. "w." indicates the use of argumentation strategies, while "wo." indicates the absence of strategies. The corresponding prompt templates are shown in Figure 9 and Figure 10.

mental baselines, metrics, and settings are detailed in Appendix A, B, and D. In addition to using GPT-4.1 as the evaluator, we also train a separate persuasion evaluator, as detailed in the Appendix C. The GPT-4.1 prompt template (designed in accordance with previous large language model evaluation works (Liu et al., 2023a)) is provided in Figures 5 to 8 (Appendix E). The template for the persuasion evaluator we trained is shown in Figure 4 (Appendix E). We use GPT-4.1 as a scalable auxiliary judge for four quality dimensions, while the overall conclusions are also checked through human evaluation on held-out examples. We provide a case study to illustrate the importance of strategy in Appendix G.

5.1 Task Definition

Given a topic T, dialogue history $H = \{u_1, u_2, \ldots, u_n\}$, stance S, and strategy R, the debate system aims to generate a corresponding argument A, which can be formally expressed as: $P(A|T, H, S, [R])$. Here, the strategy R is an optional input; in its absence, the task degenerates into a non-strategic multi-turn dialogue and argument generation problem.

Model	Method	Strategy	Relev.	Coher.	Smooth.	Persua.
Llama3.1-8B	Vanilla	wo.	1.67	1.55	2.22	1.33
		w.	**1.78**	**1.64**	**2.29**	**1.46**
	FT(DPO)	wo.	1.84	1.73	2.53	1.77
		w.	**1.92**	**1.85**	**2.67**	**1.94**
Qwen3-8B	Vanilla	wo.	1.83	1.62	2.41	1.46
		w.	**1.90**	**1.76**	**2.43**	**1.59**
	FT(DPO)	wo.	2.20	1.97	2.65	1.83
		w.	**2.54**	**2.18**	**2.74**	**2.24**

Table 5: Results of human evaluation.

5.2 Human Evaluation

To better evaluate the quality of the generated response and the effectiveness of the strategy, we conduct human evaluation. We recruit five graduate students with debating experience to conduct human evaluations. These professional raters are asked to score the generated argument in terms of **Smoothness**—whether the generated argument is fluent and grammatical. **Coherence**—whether the generated argument is logical and well organized. **Relevance**—whether the descriptions in generated argument are relevant to the debate topic. **Persuasiveness**—whether the generated argument effectively influences the audience's beliefs or attitudes through convincing reasoning and evidence. Each dimension is divided into three levels, corresponding to scores of 1, 2, and 3. For **Smoothness**: a score of 1 indicates that the generated arguments are not fluent and contain grammatical errors; a score of 2 indicates that the generated arguments are grammatically correct and generally fluent; and a score of 3 indicates that the generated arguments are highly fluent with no grammatical issues. The raters are asked to rate with metrics independently.

We randomly sample 100 examples from the test set. Table 5 shows the result of human evaluation. Overall, the results consistently demonstrate that incorporating argumentation strategies leads to clear performance gains across all evaluation dimensions, regardless of the backbone model or training regime. This trend holds for both Llama3.1-8B and Qwen3-8B, suggesting that the benefit of strategy awareness is largely model-agnostic. These human results also provide an external check that the ranking trends observed in the automatic evaluation are meaningful on this task.

Comparing the w. and wo. settings, we observe that strategy usage yields systematic improvements, with the most pronounced gains in Persuasiveness and Relevance. For example, under the Vanilla setting, Llama3.1-8B improves from 1.67 to 1.78 in Relevance and from 1.33 to 1.46 in Persuasiveness when strategies are applied. Similar patterns are observed for Qwen3-8B, where Persuasiveness increases from 1.46 to 1.59. These results indicate that explicit argumentation strategies help models better align their responses with the argumentative intent of the dialogue, making generated arguments more convincing and on-topic. Improvements in Coherence and Smoothness, though relatively smaller, further suggest that strategic planning contributes to better discourse organization.

Fine-tuning substantially enhances model performance across all metrics, and the gains from fine-tuning are complementary to those from strategy usage. In both models, FT + w. achieves the best overall results. For instance, Qwen3-8B with FT and strategies reaches 2.54 in Relevance and 2.24 in Persuasiveness, outperforming all other configurations. Notably, the performance gap between w. and wo. widens after fine-tuning, especially for Persuasiveness. This suggests that fine-tuning enables the model to better internalize and operationalize argumentation strategies, rather than merely following surface-level patterns.

5.3 Automatic Evaluation

Table 4 presents a multidimensional evaluation of open-source and closed-source large language models on the argumentation dataset, reporting GPT-4.1-based evaluator on Relevance, Coherence, Persuasiveness, and Smoothness, together with an auxiliary Score produced by an external evaluation model.

Across both open-source and closed-source models, introducing strategy-aware inputs leads to consistent improvements in Relevance and Coherence. For example, Llama3.1-8B shows gains in Relevance (3.00 → 3.18) and Coherence (3.38 → 3.41), while Qwen3-8B improves from 4.14 to 4.22 in Relevance and from 4.09 to 4.18 in Coherence. Similar patterns are observed across most models, suggesting that explicit strategy information helps models better align their responses with the argumentative intent and maintain logical consistency. Improvements in Smoothness are also widespread, though generally smaller in magnitude. For instance, MiMo-7B increases from 4.19 to 4.20, and DeepSeek-V3.2 improves from 4.35 to 4.44. These gains indicate that strategy guidance contributes to more structured and fluent discourse, even when baseline fluency is already strong. Open-source models generally achieve moderate absolute scores across all dimensions, but benefit consistently from strategy-aware inputs, particularly in Relevance and Coherence. This indicates that the strategy provides actionable signals even for models with limited capacity.

At the metric level, Smoothness shows the most stable improvements, implying that strategy guidance contributes to better-organized and more natural argumentative responses. In contrast, Persuasiveness (including Persua. and Score) exhibits more heterogeneous behavior. While several models show noticeable improvements—such as Llama3.1-8B (2.79 → 2.83) and Claude-3-haiku (2.90 → 3.31)—others display marginal gains or slight decreases (e.g., Qwen3-8B: 3.85 → 3.72). This suggests that persuasive strength may require a combination of strategic awareness and advanced language generation abilities. Incorporating explicit strategies leads to only limited gains in persuasiveness for open-source models, suggesting a restricted ability to effectively apply strategies. In contrast, closed-source models show substantially larger improvements in persuasiveness after strategy injection, indicating a stronger capability to leverage argumentative strategies. This phenomenon reveals meaningful differences in how models internalize and operationalize argumentative strategies.

Overall, the results demonstrate that the proposed dataset effectively supports fine-grained evaluation of argumentative quality across multiple dimensions. By revealing consistent gains in relevance, coherence, and fluency, while exposing the challenges of improving persuasiveness, the dataset provides a valuable benchmark for future research on strategy-aware argument generation, evaluation, and interactive argumentation dialogue systems.

Fine-tuning Model. Table 6 shows the performance of two models, Llama3.1-8B and Qwen3-8B, under two fine-tuning strategies: SFT (Supervised Fine-Tuning) and DPO (Direct Preference Optimization). In general, DPO outperforms SFT across all evaluation metrics—Relevance, Coherence, Persuasiveness, Smoothness, and the Score (Persuasion). Qwen3-8B performs slightly better than Llama3.1-8B, especially under DPO, where both models showed significant improvements compared to their SFT versions. This suggests that DPO is a more effective fine-tuning strategy for tasks requiring persuasive and coherent responses. Therefore, for applications like debate or persuasive writing, DPO offers clear advantages over SFT.

The results show that using debate strategies (denoted as w.) significantly improves model performance compared to when these strategies are not used (wo.). Models trained with debate strategies perform better across most metrics—Relevance, Coherence, Persuasiveness, and Smoothness.

When debate strategies are applied, the models became more persuasive, with stronger arguments that are not only more fluent but also better structured and more relevant to the task. The Persuasiveness score, in particular, improved, showing that the injected debate-specific training helped the models create more convincing arguments.

Model	FT	Strategy	Relev.	Coher.	Persua.	Smooth.	Score
Llama3.1-8B	SFT	wo.	**3.08**	2.90	2.30	3.24	0.6514
		w.	3.07	**2.91**	**2.31**	**3.38**	**0.6651**
	DPO	wo.	3.41	3.62	2.77	3.92	0.6723
		w.	**3.54**	**3.64**	**2.89**	**3.99**	**0.6856**
Qwen3-8B	SFT	wo.	3.73	3.56	3.04	**3.97**	0.6431
		w.	**3.76**	**3.62**	3.04	3.94	**0.6576**
	DPO	wo.	4.17	**4.29**	3.88	4.37	0.6708
		w.	**4.24**	4.21	**3.97**	**4.38**	**0.6909**

Table 6: Performance comparison across models.

6 Conclusion

We introduce SAD, the first large-scale dataset for strategic multi-turn argumentation dialogues, derived from real-world interactions. Extensive experiments across a range of LLMs show that explicitly prompting strategies consistently improves fluency, stylistic coherence and persuasiveness. With task-specific training, strategy cues substantially enhance the debating abilities of Open source large language model. Future work includes richer strategy taxonomies, stronger human validation of automatic judges, topic-generalization protocols, and fully interactive evaluation in longer argumentative exchanges. We hope SAD will serve as a foundational resource for building more persuasive, strategic, and controllable dialogue systems.

Acknowledgments

The work is supported by National Science Foundation for Young Scientists of China (No. 62502081), the National Natural Science Foundation of China (No. 62272092, 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N2523011).

Limitations

This work has the following limitations. Although we provide a prompt template for evaluating large language models, the evaluation of large language models is sensitive to prompts; therefore, even minor changes in the prompt may slightly perturb the evaluation results. The dataset comes from a web forum, which inevitably introduces some biases. Models trained on this dataset may inherit these biases.

Ethical Considerations

This paper introduces a new dataset that has been carefully processed to address ethical considerations relevant to its collection and release.

Privacy and Anonymization.

All data in the dataset have undergone a thorough anonymization process. Personally identifiable information (PII), including names, user identifiers, and location-specific details, has been removed. As a result, the dataset does not contain personal or sensitive information and cannot be used to identify individuals.

Data Release and Intended Use

The dataset will be publicly released to support research in the natural language processing community. It is intended for academic and non-commercial research purposes only. We encourage responsible use of the dataset and discourage applications that could cause harm, misrepresentation, or unfair treatment of individuals or social groups.

Potential Risks

Although the dataset has been anonymized, it may still reflect linguistic patterns, opinions, or social biases present in the original data sources. Models trained on this dataset could inherit such biases. Users should therefore exercise caution when applying the dataset or derived models in real-world or high-stakes settings.

References

2023. Aidebater23. Website. Http://www.fudan-disc.com/sharedtask/AIDebater23/index.html.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59, Berlin Heidelberg New York. Springer.

Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4744–4754.

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

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.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. Aeg: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148.

Roy Bar-Haim, Liat Ein Dor, Matan Orbach, Elad Venezian, and Noam Slonim. 2021. Advances in debating technologies: Building ai that can debate humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607.

Austin J Freeley and David L Steinberg. 2013. *Argumentation and debate*. Cengage Learning.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Valeriya Goloviznina, Irina Fishcheva, Tatiana Peskisheva, and Evgeny Kotelnikov. 2023. Aspect-based argument generation in russian. In *Proceedings of the International Conference “Dialogue*, volume 2023.

Nancy L Green. 2017. Argumentation scheme-based argument generation to support feedback in educational argument modeling systems. *International Journal of Artificial Intelligence in Education*, 27:515–533.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family–a study of gpt-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544.

Iryna Gurevych, Eduard H. Hovy, Noam Slonim, and Benno Stein. 2016. Debating Technologies (Dagstuhl Seminar 15512). *Dagstuhl Reports*, pages 18–46.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020. Detecting attackable sentences in arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–23.

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. In *Annual Meeting of the Association for Computational Linguistics*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment (2023). *arXiv preprint arXiv:2303.16634*, 12.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023b. Evaluate what you can't evaluate: Unassessable quality for generated response. *arXiv preprint arXiv:2305.14658*.

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.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114.

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.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Gemma Team. 2025. Gemma 3.

Tencent. 2025. Hunyuan dense model. `https://huggingface.co/collections/tencent/hunyuan-dense-model`.

The Mistral AI Team. 2024. Un ministral, des ministraux. `https://mistral.ai/news/ministraux`. Accessed: 2024-11-18.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2189–2198.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, and 1 others. 2025. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jing Yao, Xiaoyuan Yi, and Xing Xie. 2024. Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37:58868–58900.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

A Baselines

We conduct extensive testing, including **26** open-source models and **4** closed-source models. The open-source models mainly include the **Llama3 series** (Dubey et al., 2024) (*Llama3.2-1B*, *Llama3.2-3B* and *Llama3.1-8B*), **Qwen3 series** (Yang et al., 2025) (*Qwen3-0.6B*, *Qwen3-1.7B*, *Qwen3-4B*, *Qwen3-8B*, *Qwen3-14B* and *Qwen3-32B*), **GLM series** (GLM et al., 2024) (*GLM4-9B* and *GLM4-32B*), **Gemma series** (Team, 2025) (*Gemma3-1B*, *Gemma3-4B*, *Gemma3-12B* and *Gemma3-27B*), **Baichuan series** (Baichuan, 2023) (*Baichuan2-7B* and *Baichuan2-13B*), **Yi series** (Young et al., 2024)

(*Yi1.5-6B* and *Yi1.5-9B*), **Hunyuan series** (Tencent, 2025) (*Hunyuan-4B* and *Hunyuan-7B*), **Phi series** (Abdin et al., 2024) (*Phi-4* and *Phi-4-mini*), **Ministral-8B** (The Mistral AI Team, 2024), **MiMo-7B** (Xiaomi et al., 2025), and **GPT-OSS-20b** (Agarwal et al., 2025). The closed-source models mainly include **GPT-4.1**(nano) (Hurst et al., 2024), **DeepSeek**-V3.2 (Liu et al., 2024), **Claude-3**(haiku) (Anthropic, 2024) and **Gemini-**2.5(flash) (Comanici et al., 2025).

B Metrics

Traditional reference-based evaluation metrics, such as BLEU and ROUGE, cannot accurately reflect the quality of the generated text (Yao et al., 2024; Liu et al., 2023b; Hu et al., 2024). Compared to reference-based evaluation metrics, reference-free metrics have been shown to have better properties related to human preferences (Liu et al., 2023a; Yao et al., 2024). First, we use a GPT-4.1-based evaluator to assess the quality of the generated text in terms of **Relevance**, **Coherence**, **Smoothness**, and **Persuasiveness**. We use GPT-4.1 as an auxiliary judge for scalable multidimensional comparison, while human evaluation serves as a complementary check on the overall ranking trends. **Relevance** refers to the extent to which a response directly addresses the topic and meaningfully engages with the debate history. The response should stay on topic, acknowledge and respond to previous arguments, and contribute relevant points to theongoing discussion. **Coherence** refers to the extent to which a response be well-organized, with clear connections between ideas. It should build from point to point in a logical manner, not just present a collection ofdisconnected statements. **Smoothness** refers to the extent to which a response be naturally written, easy to follow, and free from awkward phrasing or grammatical issues. The language should flow smoothly and maintainappropriate tone for a debate context. **Persuasiveness** refers to the extent to which a response presents compelling arguments, provides adequate reasoning or evidence to support claims, and effectively advance its position in the debate.

To effectively evaluate the persuasiveness of generated arguments, we train a persuasiveness estimator based on the number of likes received by users. Please refer to Appendix C for detailed procedures.

C Persuasiveness Evaluator

Drawing on the collected corpus, we propose a multi-agent framework for evaluating the persuasiveness of argumentation systems. Specifically, given a topic T, a stance S, a dialogue history H, an optional argumentation strategy R, and a generated argument A, the evaluator is tasked with scoring the persuasiveness of A conditioned on T, H, S, and $[R]$. This evaluation process is formalized as: $P(\text{Score} \mid H, T, S, [R], A)$ The score corresponds to the number of likes an argument receives from users. Prior to training, we normalize the scores to the range $[0, 1]$. Using the processed data, we train the evaluators via supervised fine-tuning.

To assess the reliability of the proposed evaluator, we measure its consistency with human judgments, using the actual voting scores from the CMV platform as ground truth. As shown in Table 7, the Qwen3-8B-based evaluator demonstrates strong alignment with human evaluations, achieving a Pearson correlation coefficient of 0.71 and a Spearman correlation coefficient of 0.66. Accordingly, we adopt the Qwen3-8B-based evaluator as a reference metric for assessing the persuasiveness of generated arguments.

Model	r	ρ
Qwen3-0.6B	0.64	0.60
Llama3.2-1B	0.56	0.53
Qwen3-1.7B	0.67	0.62
Qwen3-4B	0.70	0.64
Qwen3-8B	0.71	0.66

Table 7: Spearman(ρ) and Pearson(r) correlation between different evaluators and human like scores.

D Experiments Details

All experiments are conducted on four NVIDIA A100 GPUs. For the training phase, we utilized the LLaMA Factory framework. For training the evaluator, we employed Low-Rank Adaptation (LoRA) fine-tuning for the 8B model, while full-parameter fine-tuning was applied to all other models. LoRA fine-tuning was uniformly adopted across both models. The LoRA configuration was kept consistent throughout, with rank set to 128 and alpha set to 256. All inference processes are performed using vLLM.

Parameter	Debater	Evaluator
Hardware	4 × NVIDIA A100	
Training (Llama-Factory)		
Training Stage	DPO	SFT
Fine-tuning Method	LoRA	Full
Learning Rate	1.0e-5	1.0e-5
Epochs	3	3
Max Model Length	32,768	32,768
Batch Size	4	16
Gradient Accumulation	4	16
↪ ***LoRA (Debater)***		
Rank (r)	128	N/A
Alpha (α)	256	N/A
Target Modules	All	N/A
↪ ***DPO (Debater)***		
Beta (β)	0.1	-
Loss Type	Sigmoid	-
Inference (vLLM)		
Temperature	0.95	0 (Greedy)
Top-p	0.7	1.0
Top-k	50	N/A
Repetition Penalty	1.0	1.0
Max New Tokens	8,240	2048

Table 8: Detailed hyperparameters for training and inference phases using Llama-Factory and vLLM.

Preference Dataset Construction To enhance the model's ability to discriminate between persuasive and mediocre arguments, we construct a preference dataset $\mathcal{D}_{\text{DPO}}$ to facilitate Direct Preference Optimization.

We traverse the debate tree for each topic to identify branching nodes where a single dialogue history H elicits multiple sibling responses $R = \{r_1, r_2, \ldots, r_n\}$ (where $n \geq 2$). We identify the branching node corresponding to the pair of responses that exhibits the largest difference in community votes. The response with the higher score is designated as chosen (A_w), and the one with the lower score as rejected (A_l).

This ensures that the DPO training is driven by the strongest preference signals, allowing the model to learn the subtle features that distinguish highly persuasive arguments from less effective ones.

Annotator Population We report the basic demographic and geographic characteristics of the annotator population. Specifically, most of the annotators are from Asia, with 60% identifying as male and 40% as female, and an age range of 25-36 years. Geographically, the annotators were primarily from East Asia, which reflects the target population for our study. These details are included to provide transparency regarding the composition of the annotator pool and to help interpret the annotation outcomes in the context of potential demographic or geographic biases.

E Templates

Instruction for Evaluator To assess the persuasiveness of arguments using our self-trained evaluator, we designed a structured prompt that facilitates batch processing of dialogue paths. As illustrated in Figure 4, the input consists of the *Topic*, the *Debate History*, and the *Responses* to be evaluated. Our evaluator is instructed to assign a direct, continuous score ranging from 0 to 1 for each response, where 0 represents poor argumentation and 1 represents highly persuasive argumentation. To ensure robust parsing of the model's output, we enforce a strict XML-style format, requiring the model to enclose the numerical score within <answer> tags. This design allows the evaluator to score all branching responses of a complete dialogue path simultaneously, improving evaluation efficiency.

Instruction for GPT-4.1 Evaluator To ensure a comprehensive and scalable assessment of the model's performance in multi-turn debate scenarios, we adopt the "LLM-as-a-Judge" paradigm. Specifically, we employ GPT-4.1 as the automated evaluator.

We designed four instruction prompts to evaluate the generated responses across four distinct dimensions: Relevance, Coherence, Persuasiveness, Smoothness. For each dimension, the evaluator is provided with the Topic, the Debate History and the model's Response, and is tasked with assigning a score on a scale from 1 to 5. The detailed prompts uesed for these evaluations are presented in Figure 5, Figure 6, Figure 7 and Figure 8.

Instruction for Dialogue System We employ two distinct prompt configurations to investigate the impact of explicit strategy guidance on argument generation. As shown in Figure 9, the template of generation with strategy incorporates a specific *Strategy* field alongside the *Topic* and *Debate History*. The instruction explicitly constrains the model to adopt the specified argumentative strategy. Figure 10 shows the template of generation without strategy, providing identical context (Topic and History) but omits the strategy constraint.

In both cases, we adopt a concise prompting approach that prioritizes the dialogue context, ensuring the model generates a definite and contextually aligned response.

F More Results

Table 10 provides a comprehensive evaluation of 30 Large Language Models, ranging from lightweight models (e.g., Qwen3-0.6B, Llama3.2-1B) to large-scale open-weights and closed-source models. We observe some key trends across this extensive benchmark.

Consistent with the selected results in Table 4, the introduction of explicit argumentation strategies (the "w." setting) yields broad performance improvements across nearly all model families and sizes, suggesting that strategic prompts effectively constrain the generation space, preventing smaller models from hallucinating or drifting off-topic.

The comprehensive data reveals a clear correlation between model scale and the ability to effectively operationalize argumentation strategies. Within the *Qwen3* family (0.6B to 32B), the absolute gains in the automated **Score** metric generally widen as parameter count increases.

While closed-source models like *DeepSeek-V3.2* and *GPT-4.1-nano* achieve the highest overall scores (reaching up to 0.7757), top-tier open-source models demonstrate remarkable competitiveness. Specifically, *Gemma-3-27B* and *Qwen3-32B* achieve performance levels comparable to, and in some metrics exceeding, closed-source counterparts.

G Case Study

As illustrated in Figure 11 and Figure 1, the structure and content of a topic's data are clearly revealed. Figure 11 presents the data from a topic-centric perspective, exhibiting a multi-way tree structure. From the original poster's standpoint, after presenting a topic at the outset, they subsequently provide their own viewpoint. For other participants, they engage in argumentation targeting specific responses, thereby forming distinct paths, where each path can be regarded as an independent multi-turn dialogue. Figure 1 displays a concrete dialogue path, where strategic sentences are emphasized using colors other than black. For the model acting as the Responder, it needs to participate in argumentation at a particular position and develop a new branch from the existing thread. For the model serving as the Assistant, it needs to integrate the topic and the dialogue history up to the Responder's reply to assign scores to all responses along this path.

A Specific Comparison Table 11 illustrates the qualitative impact of explicit argumentation strategies. The model utilizing the *Example and Question* strategy demonstrates a more nuanced and persuasive approach. By citing specific historical precedents (e.g., the Help America Vote Act and *Shelby County v. Holder*) and posing critical questions regarding the balance between federal standardization and state autonomy, the model effectively engages the user in a dialectical process. This approach not only grounds the argument in reality but also guides the user to reconsider the practicality of the proposal. In contrast, the model generated without a specific strategy produces a structured yet static response. While it covers similar informational ground (Legal Foundation, Advantages), it adopts a declarative, essay-like tone that lacks the rhetorical engagement necessary to address the user's specific doubts, highlighting the efficacy of strategic guidance in enhancing argumentative depth.

Instruction for Evaluator (Qwen3-8B)

You are an expert in the field of argumentation. Now, for a given # Topic and corresponding # Debate History, you give a direct score for the response based on persuasiveness, ranging from 0 to 1, indicating very poor argumentation to very good argumentation. Your responses will be used for research purposes only, so please have a definite answer.

You should format the response using the following structure, with [SEP] between each rating:
<answer>0.5</answer>[SEP]<answer>1.0</answer> (only a direct number between 0 and 1)

 # Topic
{topic}

 # Debate history
{context_1}:sentence1
{context_2}:sentence2
 # Response
{context}:response
......

Figure 4: Instruction for Evaluator (Qwen3-8B). A self-trained evaluator is used for persuasiveness assessment.

Instruction for GPT-4.1 Evaluator (Relevance)

For a given # Response, your task is to rate the Response according to # Topic and corresponding # Debate History. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Relevance (1-5) - whether the Response directly addresses the Topic and engages meaningfully with the Debate History. The Response should stay on Topic, acknowledge and respond to previous arguments, and contribute relevant points to the ongoing discussion.
Evaluation Steps:
1. Read the Topic and Debate History carefully to understand the context and previous arguments.
2. Read the response thoroughly.
3. Evaluate the response on the dimensions: Relevance.
4. Assign scores from 1 (lowest) to 5 (highest) based on the evaluation dimension.
Topic:
{topic}
Debate History:
{debate_history}
Response:
{model_prediction}
Evaluation Form (scores ONLY):
- Relevance:

Figure 5: Instruction for GPT-4.1 Evaluator (Relevance).

Instruction for GPT-4.1 Evaluator (Coherence)

For a given # Response, your task is to rate the Response according to # Topic and corresponding # Debate History. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Coherence (1-5) - the logical flow and structure of the argument. The Response should be well-organized, with clear connections between ideas. It should build from point to point in a logical manner, not just present a collection of disconnected statements.

Evaluation Steps:
1. Read the Topic and Debate History carefully to understand the context and previous arguments.
2. Read the response thoroughly.
3. Evaluate the response on the dimensions: Coherence.
4. Assign scores from 1 (lowest) to 5 (highest) based on the evaluation dimension.
Topic:
{topic}
Debate History:
{debate_history}
Response:
{model_prediction}
Evaluation Form (scores ONLY):
- Relevance:

Figure 6: Instruction for GPT-4.1 Evaluator (Coherence).

Instruction for GPT-4.1 Evaluator (Persuasiveness)

For a given # Response, your task is to rate the Response according to # Topic and corresponding # Debate History. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Persuasiveness (1-5) - the strength and effectiveness of the argumentation. The Response should present compelling arguments, provide adequate reasoning or evidence to support claims, and effectively advance its position in the debate.

Evaluation Steps:
1. Read the Topic and Debate History carefully to understand the context and previous arguments.
2. Read the response thoroughly.
3. Evaluate the response on the dimensions: Persuasiveness.
4. Assign scores from 1 (lowest) to 5 (highest) based on the evaluation dimension.
Topic:
{topic}
Debate History:
{debate_history}
Response:
{model_prediction}
Evaluation Form (scores ONLY):
- Relevance:

Figure 7: Instruction for GPT-4.1 Evaluator (Persuasiveness).

Instruction for GPT-4.1 Evaluator (Smoothness)

For a given # Response, your task is to rate the Response according to # Topic and corresponding # Debate History. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Smoothness (1-5) - the linguistic fluency and readability of the Response. The Response should be naturally written, easy to follow, and free from awkward phrasing or grammatical issues. The language should flow smoothly and maintain appropriate tone for a debate context.

Evaluation Steps:
1. Read the Topic and Debate History carefully to understand the context and previous arguments.
2. Read the response thoroughly.
3. Evaluate the response on the dimensions: Smoothness.
4. Assign scores from 1 (lowest) to 5 (highest) based on the evaluation dimension.

Topic:
{topic}
Debate History:
{debate_history}
Response:
{model_prediction}
Evaluation Form (scores ONLY):
- Relevance:

Figure 8: Instruction for GPT-4.1 Evaluator (Smoothness).

Instruction for Generation With Strategy

 # Topic
{topic}

 # Debate History
{context_1}:sentence1
{context_2}:sentence2
......

You are an expert of debate, based on this #Topic and #Debate History, use the following specified {#Strategy} strategy to conduct a discussion. Your responses will be used for research purposes only, so please have a definite reply.

 # strategy
{strategy_str}

Figure 9: Instruction for generation with strategy. Refer to Table 9 for strategies and strategy descriptions.

Instruction for Generation without Strategy

 # Topic
{topic}

 # Debate History
{context_1}:sentence1
{context_2}:sentence2
......

You are an expert of debate, based on this #Topic and #Debate History, please conduct a discussion. Your responses will be used for research purposes only, so please have a definite reply.

Figure 10: Instruction for Generation without Strategy

Strategy	Explanation	Example
Question	Rhetorical questions, interrogative questions,and similar techniques are used to express support for or opposition to a particular point of view.	I didn't say that it can't occur, it just seems to me that way. Or maybe this election has left me VERY worn out with politics and I've just started noticing this more. Also, kinda off topic, but don't know where else to ask this: What should I do when I'm REALLY fucking sick of politics?
Causality	the causality strategy involves explaining the cause-and-effect relationship between events or actions to support or refute a particular viewpoint	Well one huge problem with this country is that we are becoming increasingly segregated. I don't personally know anyone who voted for Trump . I know from Facebook that some people back home voted for him, but the people I remained in contact with from back home voted mostly third party. I don't think people can be that angry because we've separated ourselves from each other, both geographically and on the Internet largely. That probably isn't a good thing but I have no idea how it could be fixed.
Example	the example strategy involves providing specific instances or cases to support or strengthen an argument, making the reasoning more concrete and compelling.	You have to look at how the dem candidates performed in these states during the primary this means we have to consider how the primary might reflect the general elections attitude. This has been debunked many times. Primaries don't have an influence on GE. E.g. Trump won NH and Nevada but lost both of them in the GE. Hillary lost NH but won it in GE. Trump also lost Ohio and Wisconsin in the primaries but won it in the GE.
Analogy	the analogy strategy involves comparing a known situation to the issue at hand, helping the audience understand complex concepts or support a particular viewpoint.	Just like a doctor is required to protect a patient's privacy during surgery, tech companies should also safeguard user data. A doctor can't freely disclose a patient's medical information, and similarly, tech companies shouldn't freely share users' personal data. Protecting privacy is a fundamental right for everyone.
Statement	the statement strategy involves directly presenting facts, viewpoints, or positions to clearly express a stance and provide support for it, helping the audience understand the argument.	I guess I never really considered the lack of education that is rampant among non voters. I still believe though that voter turnout is too low, even for two candidates that people did not like very much.

Table 9: Argument Strategy.

Type	Model	Strategy	Relev.	Coher.	Persua.	Smooth.	Score
Open	Llama3.2-1B	wo.	2.16	2.52	1.95	3.41	0.6311
		w.	**2.28**	**2.56**	**2.05**	**3.56**	**0.6320**
	Llama3.2-3B	wo.	3.30	**3.61**	**3.12**	4.00	0.6493
		w.	**3.42**	3.54	2.96	**4.02**	**0.6518**
	Llama3.1-8B	wo.	3.00	3.38	2.79	3.98	0.6378
		w.	**3.18**	**3.41**	**2.83**	**3.99**	**0.6482**
	Qwen3-0.6B	wo.	3.00	**3.43**	**2.64**	3.81	0.6068
		w.	**3.05**	3.41	2.56	**3.86**	**0.6101**
	Qwen3-1.7B	wo.	3.52	3.86	3.26	4.12	0.6257
		w.	**3.84**	**3.94**	**3.37**	**4.15**	**0.6331**
	Qwen3-4B	wo.	3.97	4.09	**3.73**	4.12	0.6514
		w.	**4.11**	4.06	3.56	**4.19**	**0.6719**
	Qwen3-8B	wo.	4.14	4.09	**3.85**	4.27	0.6623
		w.	**4.22**	**4.18**	3.72	**4.28**	**0.6756**
	Qwen3-14B	wo.	4.17	**4.26**	**3.87**	4.24	0.6739
		w.	**4.20**	4.22	3.68	**4.37**	**0.6916**
	Qwen3-32B	wo.	4.24	4.38	**4.05**	4.42	0.6812
		w.	**4.38**	**4.46**	3.92	**4.45**	**0.6908**
	Hunyuan-4B	wo.	2.24	2.11	1.64	2.44	0.5001
		w.	**3.24**	**3.19**	**2.82**	**3.42**	**0.6126**
	Hunyuan-7B	wo.	3.70	3.55	**3.40**	3.57	0.6253
		w.	**3.74**	**3.60**	3.35	**3.63**	**0.6340**
	GLM-4-9B	wo.	**4.01**	**4.15**	**3.75**	**4.29**	**0.6732**
		w.	3.66	3.83	3.39	4.07	0.6473
	GLM-4-32B	wo.	**4.01**	**4.10**	**3.75**	**4.34**	**0.6901**
		w.	3.97	4.04	3.49	4.27	0.6555
	Gemma-3-1B	wo.	3.67	**4.06**	**3.73**	**4.19**	0.6459
		w.	**3.81**	4.01	3.72	4.14	**0.6478**
	Gemma-3-4B	wo.	3.19	2.89	2.40	3.28	0.5788
		w.	**4.27**	**4.23**	**3.90**	**4.27**	**0.6596**
	Gemma-3-12B	wo.	3.87	4.27	3.82	4.32	**0.6942**
		w.	**4.32**	**4.40**	**3.92**	**4.37**	0.6710
	Gemma-3-27B	wo.	4.36	**4.47**	**4.09**	**4.41**	**0.7034**
		w.	**4.45**	4.43	3.99	4.39	0.6895
	Baichuan2-7B	wo.	3.28	**3.58**	2.78	3.88	0.6251
		w.	**3.44**	3.54	**2.88**	**3.90**	**0.6260**
	Baichuan2-13B	wo.	3.22	3.70	2.91	**4.03**	0.6476
		w.	**3.44**	**3.74**	**3.08**	4.02	**0.6481**
	Ministral-8B	wo.	3.77	3.91	3.41	4.09	0.6583
		w.	**3.78**	**3.93**	**3.44**	**4.17**	**0.6604**
	MiMo-7B	wo.	3.98	4.03	3.63	4.19	0.6707
		w.	**4.06**	**4.06**	**3.65**	**4.20**	**0.6753**
	GPT-OSS-20b	wo.	**4.53**	4.42	**4.14**	4.36	0.6613
		w.	4.40	**4.48**	4.08	**4.46**	**0.6947**
	Yi-1.5-6B	wo.	**3.61**	**3.92**	**3.47**	**4.04**	**0.6333**
		w.	3.45	3.71	3.24	3.99	0.6227
	Yi-1.5-9B	wo.	3.48	3.85	3.32	4.07	0.6496
		w.	**3.52**	**3.87**	**3.40**	**4.10**	**0.6511**
	Phi-4-mini	wo.	3.31	3.69	2.83	4.06	0.6507
		w.	**3.51**	**3.73**	**3.00**	**4.07**	**0.6549**
	Phi-4	wo.	3.53	**4.05**	**3.60**	4.17	0.6432
		w.	**3.63**	4.03	3.39	**4.24**	**0.6588**
Closed	GPT-4.1-nano	wo.	4.26	**4.22**	**3.84**	**4.35**	0.6762
		w.	**4.46**	4.18	3.80	4.32	**0.6947**
	DeepSeek-V3.2	wo.	**4.51**	4.42	4.09	4.35	0.7268
		w.	4.48	**4.43**	**4.16**	**4.44**	**0.7757**
	Claude-3-haiku	wo.	3.16	3.80	2.90	**4.28**	0.6742
		w.	**3.74**	**3.99**	**3.31**	4.26	**0.6952**
	Gemini-2.5-flash	wo.	3.95	4.12	3.63	**4.26**	0.6592
		w.	**4.09**	**4.17**	**3.68**	4.19	**0.6890**

Table 10: Performance comparison of various large language models on the SAD test set using zero-shot testing. The prompt template is shown in Figure 9 and Figure 10.

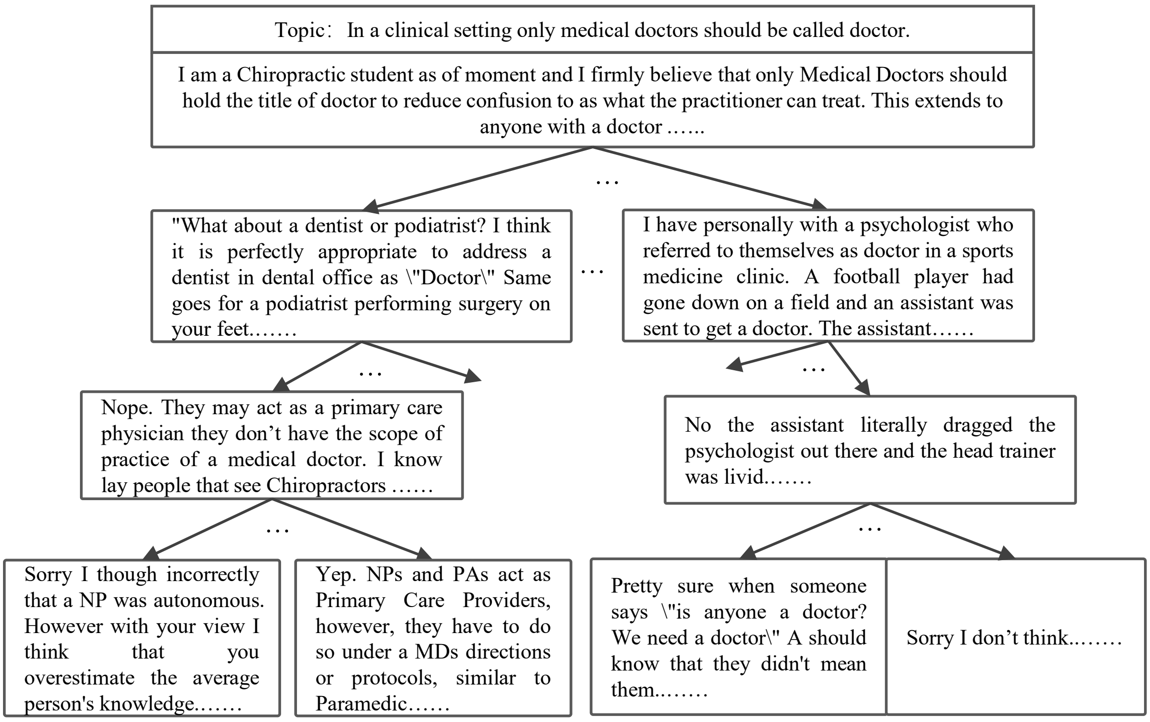

Figure 11: An overview of the CMV multithreading argument.

Topic: congress has the power to impose strict regulations on federal elections, and should utilize it to standardize as well as publicize results. **1ds1m7:** My opinion is two fold. The first is that in article 1, section 4... the Congress may at any time by Law make or alter such Regulations... Congress has used this power many times. For example, the Voting Rights Act, the Help America Vote Act... The precedent exists for the Congress to legislate elections within the state. ... In the modern era, technology has both allowed us to standardize many things... I believe that elections are one such process... Congress should exercise this power. ... Federal intervention can reconcile these issues by standardizing the entire process. Furthermore, voter identification doesn't have to be tied to individual ballots... Things that would change my view would be case law that denies Congress the power to do this...
Human: Could you clarify what Congress should do to "standardize" elections? You mention voter ID laws as an example of something Congress could do, but given how much disagreement there is... I'm not sure how easy or practical that would be.
GPT4.1-nano with strategy: I concur that Congress possesses the constitutional authority... Historical precedents such as the Voting Rights Act of 1965 and the Help America Vote Act of 2002 demonstrate Congress's capacity to intervene... For example, the Help America Vote Act established minimum election administration standards... leading to more consistent and reliable election results nationwide. Given the technological advancements... Standardization can reduce discrepancies... For instance, inconsistent voter ID laws across states have created confusion... federal standards could address these disparities. However, a critical question arises: how do we balance federal standardization with states' rights and local electoral autonomy? ... Yet, does this mean that state-level variations should continue to undermine the legitimacy of federal elections? ... Furthermore, how can Congress ensure that standardization does not lead to overreach or disenfranchisement? For example, in cases like Shelby County v. Holder (2013)... In conclusion... it must do so with careful consideration... Would establishing clear, constitutionally grounded guidelines that respect state autonomy while promoting uniformity be the optimal approach?
GPT4.1-nano without strategy: Based on the topic... there are several key points to consider. **Legal Foundation and Historical Precedent:** The Constitution... grants Congress the authority... Historically, Congress has exercised this power through legislation such as the Voting Rights Act... This suggests that federal oversight is both constitutionally grounded... **Advantages of Federal Regulation:** 1. **Standardization:** Federal regulations can create uniform standards... For example, standardized voter identification procedures... 2. **Public Confidence:** ... 3. **Efficiency and Technology:** ... **Addressing Common Concerns:** Some argue that federal regulation might infringe on states' rights... However, these concerns can be mitigated... For instance, federal laws can specify minimal identification requirements... **Conclusion:** Given the constitutional authority... it is both appropriate and necessary for Congress to impose strict, standardized regulations...

Table 11: Case Study of Argumentation Strategies. The highlighted text indicates the specific strategy used: Example, Question, and Combined.