# Debatrix: Towards a Fine-Grained Automatic Debate Judging

**Anonymous ACL submission**

## Abstract

How can we construct an automated debate judge to assist with evaluating an extensive, fervent, multi-turn debate? This task is challenging, as judging a debate involves grappling with lengthy texts, intricate argument relationships, and multi-dimensional assessments, while current research mainly focuses on short dialogues, rarely touching upon the evaluation of an entire debate. In this paper, by leveraging Large Language Models (LLMs), we propose Debatrix, which makes the analysis and assessment of multi-turn debates more aligned with majority preferences. Specifically, Debatrix features a horizontal chronological workflow and a vertical multi-dimensional evaluation collaboration. To align with real-world debate scenarios, we introduced DebateArt and DebateCompetition benchmarks, comparing our system's performance to actual debate outcomes. The findings indicate a notable enhancement over directly using LLMs for debate evaluation.

## 1 Introduction

Debating is the formal process of gaining consensus among groups with different opinions. While some debates are cooperative and aim to solve a public issue, in many cases, such as *competitive* debates, only the policy from the winning side will be accepted (Zhang et al., 2016). Debaters in these debates must apply various strategies to persuade the audience to support their side. Developing systems like Project Debater (Slonim et al., 2021) demands thorough testing for effectiveness. Yet, assessing their performance against humans involves extensive human input. Furthermore, the common method of audience voting for debate outcomes is often unclear and subject to bias. This reinforces the importance of automatic debate evaluation.

Recently, large language models (LLM) such as ChatGPT and GPT-4 (OpenAI, 2023) have shown a strong ability to solve various downstream tasks, including text quality evaluation (Liu et al., 2023; Chiang and Lee, 2023). Referred to by Zheng et al. (2023) as *LLM-as-a-judge*, LLMs are capable of providing results more aligned to human preference than traditional metrics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). Additionally, verdicts from LLM judges are easier to interpret, as they are usually combined with generated explanations; this brings an extra advantage compared to crowdsource voting, including the audience voting method mentioned above.

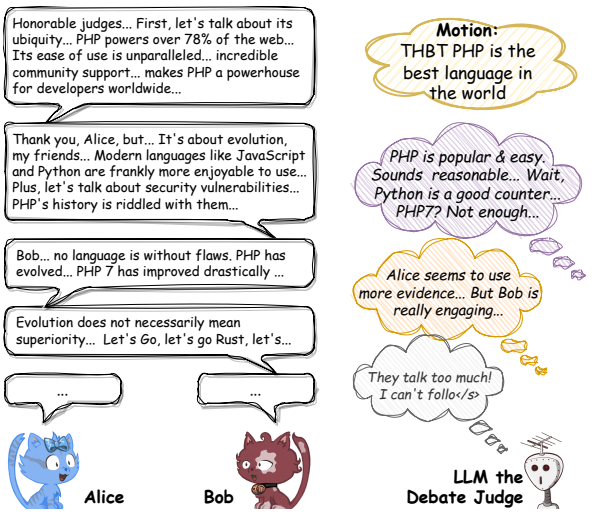


Figure 1: An LLM debate judge judging a debate between Alice and Bob. The LLM needs to understand the arguments and how they counter each other (purple bubble); the LLM also needs to evaluate the speeches in multiple dimensions (orange bubble). However, multi-round debates are often very long, detracting attention or even exceeding the context window (light gray bubble).

However, judging debates with LLMs incorporates several issues to be considered, as illustrated in Figure 1. First, evaluating long, multi-turn debates continues to be challenging, while most current research focuses on short text, such as open-question answers and user-request responses (Zhong et al., 2022; Wu et al., 2023).

Second, debate, as a field rich in logic and criti-

cal thinking, often requires a deep understanding of how arguments are organized and refuted across speeches. This makes the evaluation of debates require more professional and multi-dimensional knowledge. Finally, the quality of speeches is affected by various factors, such as argument strength, evidence reliability, and language style, demanding systematic analyses across dimensions.

To this end, we propose Debatrix, a fine-grained framework to assist LLMs in handling these challenges by breaking down debate evaluation along both *chronological* and *multi-dimensional* axes. 1) ***Chronological*** *workflow*: We instruct the LLM to analyze the debate speech by speech and use a memory system to maintain the speech stream and the analysis stream. After reviewing all speeches, the LLM then makes decisions based on these analyses. This chronological approach lets the LLM concentrate on one speech at a time and also allows it to provide feedback or decisions for each speech, each debater, and the final winner. 2) ***Multi-dimensional*** *collaboration*: Debatrix also allows LLMs to focus on a specific judging dimension, such as arguments, language, or clash, during the speech analyzing process. Each LLM agent can make comments on these specific aspects. For the overall judgment, all these individual analyses are combined into one summary, providing a systematic judgment across multiple dimensions.

For the experiments, we introduce novel debate judging benchmarks covering multi-dimensional and multi-debater scenarios, namely DebateArt and DebateCompetition. DebateArt sources from online platforms that follow competitive debate formats and have dimensional voting results, while DebateCompetition includes transcribed videos from world-class competitive debate competitions, enriching our benchmark with complex and high-quality samples. These debates follow the British Parliamentary (BP) format involving four teams (two on each side), increasing judging difficulty. On these two more challenging benchmarks, our Debatrix evaluation system achieved improvements in winner prediction accuracy both per dimension and generally, compared to the baseline of directly prompting the LLM with raw debate speeches. Furthermore, the experimental results have also proved that speech-by-speech analyses and multi-dimensional judgments help generate a more accurate final verdict.

Our contributions are as follows:

1. We propose Debatrix, a fine-grained automatic debate judging framework based on LLM but performs general or dimensional analysis speech by speech before producing the final verdict.

2. We propose a debate judging benchmark for LLMs and other autonomous debate judging systems, including multi-dimensional and multi-debater settings, which differ from simple 1v1 debate assessment.

3. We investigate how well LLMs can judge debates directly or equipped with Debatrix, enabling either chronological or dimensional analysis or both.

## 2 Related Work

Argumentation persuasion assessment is the foundation of automatic debate systems, as they must be persuasive enough to argue effectively. Previous works have focused on the persuasiveness of arguments, including empirical studies (Thomas et al., 2017, 2019a), machine learning models (Persing and Ng, 2015; Zhang et al., 2016; Gleize et al., 2019) and works covering both (Al Khatib et al., 2020; Donadello et al., 2022). Some argument-based chatbots also take persuasiveness as a motivating factor (Rosenfeld and Kraus, 2016; Thomas et al., 2019b).

While these works mainly involve empirical laws or delicate models, within the bigger context of text evaluation, large language models (LLM) have become a new, powerful tool to tackle this task. Multiple exploitation methods have been proposed, such as conditional probability (Fu et al., 2023), score prompting (Wang et al., 2023a; Liu et al., 2023; Zhu et al., 2023) and pairwise comparison (Wang et al., 2023b). Several works, such as Bai et al. (2023), utilize multiple of them and introduce various methods to stabilize the results.

As for the challenges of judging debates with LLMs, a few pioneering works have researched some of them. For instance, Li et al. (2023) and Chan et al. (2023) focus on multi-dimensional assessment and propose different strategies to improve accuracy, such as peer review and group chat. Meanwhile, de Wynter and Yuan (2023) and Chen et al. (2023) have explored LLM's ability to handle argumentation tasks, measuring its capability of argumentation reasoning. These works are partially

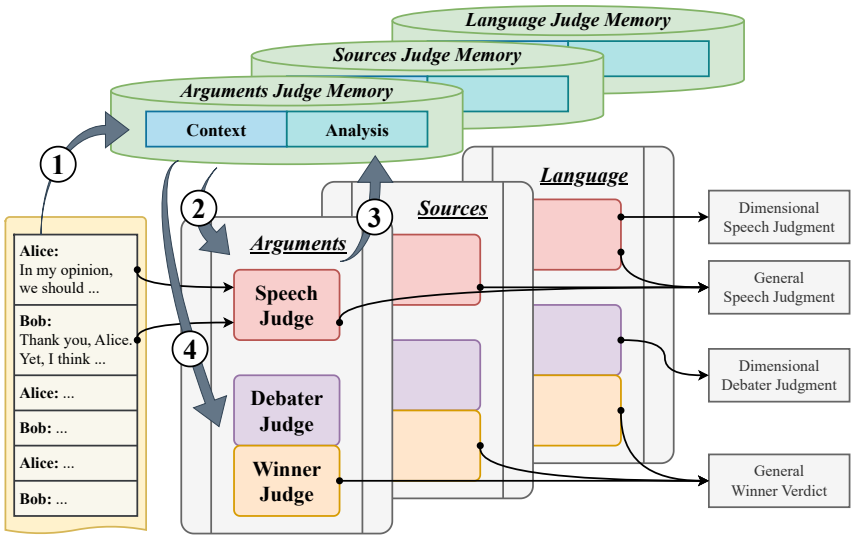


Figure 2: General structure of Debatrix. ①: add speech to context memory; ②: fetch relevant pieces of context and analysis; ③: add analysis and reflections to analysis memory; ④: fetch analysis for final judgment. The framework can generate speech judgments, debater judgments, and the winner verdict based on analysis from single or multiple dimensions.

in line with our work yet do not cover all 3 issues mentioned in Section 1. Finally, Li et al. (2019) proposed a dialog evaluation framework that works in a multi-turn manner, similar to our chronological design. However, this framework is designed for human annotators instead of LLM judges.

## 3 Debatrix

In this section, we provide a detailed overview of Debatrix, our fine-grained debate judging framework, including its general structure and workflow.

### 3.1 Structure and Components

The overall structure of Debatrix is illustrated in Figure 2. Debatrix contains a collection of **chronological columns**, each being able to evaluate debates under a specific preference or dimension. Each chronological column processes speeches sequentially, generating speech judgments, including a score and a comment. When all speeches are analyzed, the column summarizes the past analyses, generating debater judgments (similar to speech judgments but for individual debaters) and the winner verdict (including the winner and a comment). Multiple columns can collaborate like a matrix, producing systematic judgments that cover multiple dimensions.

There are two groups of components in a chronological column: memory and judgment. They work together to analyze the debate thoroughly.

**Memory** provides long-term storage during the debate judging process. There are two types of memory: context memory records incoming speech context, and analysis memory stores intermediate analyses. Every incoming speech is added to the context memory before being analyzed by the speech judge. At any time, judges can fetch or query contents from both memories and add new analyses to the analysis memory.

**Judge** is the core component to analyze and judge the debate, including the speech judge, the debater judge, and the winner judge. The speech judge analyzes the stream of speeches, utilizing memories to understand them; the analysis is added to the memory and can be used to generate judgments. The debater and winner judges work after all speeches are processed. They use past analyses by the speech judges to generate debater judgments and the winner verdicts, respectively.

### 3.2 Chronological Column Workflow

LLMs have shown a strong capability in evaluating single passages and short conversations. To elaborate on this power for multi-turn long debates, we propose a dedicated workflow design for chronological columns, which can already be seen in Figure 2. Figure 3 illustrates a more detailed version of the workflow; the complete algorithm is listed in Appendix A.

The key point of our design is the speech analysis process. Speech analysis focuses on decomposing

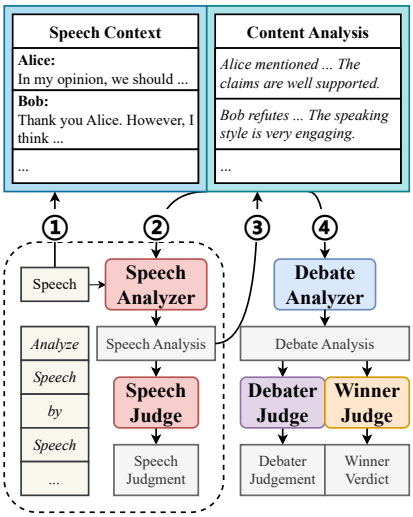


Figure 3: A more detailed version of the workflow of a single column. Blocks and numbers match the ones in Figure 2, except for the debate analyzer which is not shown in Figure 2.

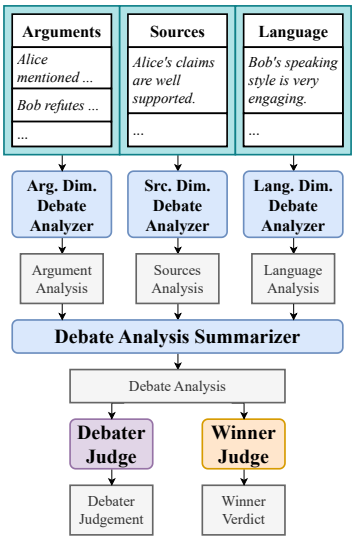


Figure 4: Combining dimensional debate analyses from multiple columns into one systematic analysis at the end of the debate. Note that each column possesses a memory containing dimensional content analysis, which allows a more nuanced understanding of the debate.

the content of the speech, such as how arguments interact with each other, what evidence is introduced to back arguments, and what language style the speech has shown. The column preference or dimension controls what is included in the analysis. The speech judge can use this information to generate the corresponding speech judgment and gain insight into the performance of the current speaking debater. Meanwhile, the analysis also acts as a summarized and digested version of the current speech, reducing the difficulty of subsequent speech judgment.

When the speech judge finishes analyzing the speech, the speech analysis is added to the analysis memory as part of the content analysis. At the end of the debate, the debater and winner judges must exploit the entire list of content analyses to judge a specific debater or compare between debaters. This is achieved by an extra debate analyzer, which converts all content analysis into a debater-directed debate analysis. Finally, the debater judges and the winner judge generate debater judgments and the winner's verdict based on the debate analysis.

### 3.3 Multi-Dimensional Collaboration

While we can configure a single column to produce general judgment directly, in Debatrix, combining multiple columns focusing on various dimensions is a better approach. One approach in our experiments is to summarize analyses from multiple columns into one systematic analysis. Figure 4 demonstrates this approach for debater assessment and winner judging, summarizing debate analyses under various dimensions; the same approach can also be applied to speech analyses when generating speech judgments.

It is worth noting that this is not the only way to collaborate columns. For instance, columns may interact with each other during the analyzing process, similar to the approach of Chan et al. (2023); we have examined whether such interactions are beneficial when summarizing debate analyses. However, they are not limited to this specific step — interactions may occur at any point, even when generating content analysis.

## 4 Debate Judging Benchmark

To assess LLMs and our proposed Debatrix framework with real competitive debates, we introduce a novel debate judging benchmark covering multi-dimensional and multi-debater scenarios. We include two debate sources in our debate judging benchmark: DebateArt for multi-dimensional 1v1 debates and debate competitions for high-quality, multi-debater debates.

### 4.1 DebateArt

One major part of our debate judging benchmark is based on DebateArt[1], an online debate platform that provides 1v1 debate arenas. The formation setting makes DebateArt different from many other

[1] https://debateart.org

| | # Speech | Speech Tok. | Debate Tok. |
|---|---|---|---|
| Min | 4.0 | 53.0 | 468.0 |
| Mean | 6.7 | 650.5 | 4,342.6 |
| Max | 10.0 | 2,368.0 | 12,337.0 |

Table 1: DebateArt debates content statistics, including the number of speeches in a debate, tokens in a speech, and tokens in a debate.

| Dimension | Pro | Tie | Con |
|---|---|---|---|
| General | 37 | 7 | 56 |
| Arguments | 33 | 11 | 56 |
| Sources | 14 | 67 | 19 |
| Langauge | 9 | 66 | 25 |

Table 2: DebateArt debates winner distribution, including general and dimensional ones.

debate forums that do not restrict the speaking order, as debates on this platform are closer to formal competitive ones.

In DebateArt, users vote to decide debate winners. Besides the common winner selection system, the platform also provides a categorical point assignment system, where voters must consider and vote on four metrics for comparative performance insights: arguments, sources, legibility, and conduct.[2] This voting system provides judging results under separate dimensions, which can be pro-winning, con-winning, or a tie. Moreover, voters must provide detailed explanations of their decisions, and their votes are supervised by experienced moderators, enhancing the quality of the votes. The weighted average of votes under each dimension decides the debate's winner. Details of how DebateArt debates are run are covered in Appendix B.1.

Our benchmark includes 100 valid debates with valid votes from DebateArt; we include our data collection process and filtering criteria in Appendix B.2. Table 1 lists the statistics of these debates. To align with oral debates that are not formatted, we merged two dimensions — legibility and conduct — into a single language dimension, representing the language style. We averaged their votes in these two dimensions for each vote and converted them into a single vote (pro/tie/con). Table 2 lists the winner distribution of all debates: while voters tend to give ties on sources and language, most debates have a specific winning side, largely because of the argument dimension.

[2] https://info.debateart.com/terms-of-service/voting-policy#casting-votes

| | Speech Tok. | Debate Tok. |
|---|---|---|
| Min | 1,478.0 | 13,571.0 |
| Mean | 1,892.5 | 15,139.9 |
| Max | 2,411.0 | 17,089.0 |

Table 3: Competition debates content statistics, including the number of tokens in a speech and a debate.

| Debater | OG | OO | CG | CO |
|---|---|---|---|---|
| # Wins | 8 | 16 | 8 | 6 |

Table 4: Competition debates winner distribution. In BP debates, the names of the teams are always fixed to OG, OO, CG, and CO; OG and CG form the pro side, and OO and CO form the con side. As some debates have two winners, the sum exceeds 22.

### 4.2 Debate Competitions

To extend our benchmark with high-quality *formal* debates, we furthermore collected debates from world-class competitive debate competitions in recent years, including the World Universities Debating Championship (WUDC), the European Universities Debating Championship (EUDC), and the North American Debating Championship (NAUDC). All these competitions follow the British Parliamentary (BP) format (World Universities Debating Council, 2023), where four teams are divided into two sides of a motion but compete against all three other teams in the debate. Instead of predicting the winning side, our benchmark requires judging which of the four teams is the best. Please refer to Appendix B.3 for more details about BP debates.

We transcribed debate videos from knockout rounds of famous competitive debate competitions to obtain high-quality BP debates; for more data collection details, please refer to Appendix B.4. After filtering incomplete or damaged transcriptions, we obtained 22 debates with full transcriptions and final verdicts. Table 3 lists statistics of these competition debates; they are significantly longer than DebateArt debates.

It is worth noting that we were only able to collect the winning teams of these debates, which is not necessarily unique: in the finals, only one of the four teams wins the debate; yet, in the semi-finals and quarter-finals, two of them (they can even be mutual opponents) can win and proceed. Among all the collected debates, 6 are finals and have only one winner, and 16 have two winners. Table 4 demonstrates the distribution of winners in all these debates. To unify them, our benchmark treats predicting any winning teams as correct.

## 5 Experiments

Using our benchmark, we conduct experiments to evaluate the debate-judging performance of LLMs. We also compare our Debatrix framework with judging directly with LLMs.

### 5.1 Model & Framework Configuration

We utilize the latest GPT family as our target LLMs, including ChatGPT (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-0125-preview`); our experiments mainly focus on ChatGPT to test Debatrix under a limited context window ($16,385$ tokens). We set the temperature to 0 and repeated all experiments 3 times, measuring the average performance.

To control the dimension of the judge results, we provide judge preferences in the system prompt for every judge. We ask the judges to output comments only and then call the LLM again to generate their corresponding scores (integers from 1 to 10) or winner verdict to diminish mismatches.

We compare two variants of chronological workflow concerning relevant context fetching: half-analysis and full-analysis. In half-analysis, we fetch previous speeches; in full-analysis, we fetch content analysis of previous speeches instead of raw speeches. We denote Debatrix frameworks with these workflow variants as Debatrix-H and Debatrix-F, respectively.

### 5.2 Experiment Settings

We use DebateArt debates and competition debates to evaluate general winner prediction performance and DebateArt debates exclusively to evaluate dimensional ones.

**Metrics** We compare the winner verdict prediction with the true voting result of the debate; for scores, we compare the two debaters to generate an alternative score-based winner verdict and compare it with the true verdict. Specifically, for the sources and languages dimensions, we treat score differences within $\pm 3$ as ties, as score comparison is too strict (as can be seen in Table 2). We also prompt the model to allow tie verdicts. Due to the existence of ties, we measure the root mean square error (RMSE) between the model prediction and the votes. More specifically, we assign the values of pro, tie, and con votes/predictions to $0$, $0.5$, and $1$, respectively; then we match each vote to its corresponding model prediction and calculate the RMSE. Debates from debate competitions have four debaters and do not have ties. Therefore, we prompt the model only to give winning verdicts, measure the accuracy, and ignore the scores. As for Debatrix dimensions, we adopt the dimensions defined for DebateArt, and add a clash dimension which originally belongs to the arguments dimension, so that all dimensions are balanced.[3]

The baseline models we used are as follows:

- **Direct**: We choose ChatGPT without any frameworks as our baseline, which we denote as Direct. We prompt the model to analyze and judge a debate, and input the entire debate.
- Debatrix-H: While both Debatrix-H and

[3]The clash dimension is for competition debates exclusively; for DebateArt debates, we stick to their pre-defined dimensions.

| Method | Arg. | Src. | Lang. | Summary |
|---|---|---|---|---|
| Direct | 52.06 | **32.47** | 39.91 | 44.91 |
| Debatrix-H | 50.87 | 33.14 | **32.55** | 44.18 |
| Debatrix-F | **47.50** | 34.67 | 36.16 | **42.21** |

Table 5: ChatGPT score comparison RMSE ($\times 100$) on DebateArt debates. Arg.: Arguments; Src. Sources; Lang.: Language. Summary predictions are generated by multi-column collaboration. Lower RMSE is better.

| Method | Arg. | Src. | Lang. | Summary |
|---|---|---|---|---|
| Direct | 52.23 | 41.37 | 47.31 | 45.01 |
| Debatrix-H | 50.37 | 38.43 | **33.41** | 44.03 |
| Debatrix-F | **47.67** | **37.67** | 44.22 | **41.75** |

Table 6: ChatGPT winner prediction RMSE ($\times 100$) on DebateArt debates. Notations follow Table 5. Lower RMSE is better.

| Method | Completion % | Accuracy |
|---|---|---|
| Direct | 36.36 | 16.67 |
| Debatrix-F | **100** | **51.52** |

Table 7: ChatGPT winner prediction accuracy on competition debates. Completion %: percentage of debates for which the LLM can actually *complete* a verdict within the context window.

Debatrix-F apply our framework, Debatrix-H is selected as a baseline to examine the effectiveness of speech analysis.

### 5.3 Main Results

Tables 5 and 6 demonstrate score comparison and winner prediction RMSEs on DebateArt debates using ChatGPT respectively. Debatrix supersedes the direct method both generally and on each dimension, except for the dimension of sources in score comparison, where the difference is relatively small compared to other dimensions. As for the two Debatrix variants, on the one hand, Debatrix-F consistently outperforms Debatrix-H on the arguments dimension, which contributes to its lead in general winner prediction; on the other hand, Debatrix-H performs better on sources and language dimensions, specifically the latter one.

Table 7 lists the winner prediction accuracy on competition debates. Due to the limited context window, bare ChatGPT failed to process over 60% of the debates, significantly hindering performance. For this reason, we do not conduct experiments with Debatrix-H since, with this variant, analyzing the last speech costs as many tokens as judging the entire debate. In contrast, Debatrix-F successfully solved this issue, aiding ChatGPT to judge all 22 debates and achieve a much higher accuracy.

## 6 Analysis

### 6.1 Half-Analysis or Full-Analysis?

From Tables 5 and 6, we can see that while Debatrix-F is consistently better than Debatrix-H when judging arguments, Debatrix-H is better on the language dimension. This indicates that either variant has its specific expertise, or in other words, different dimensions prefer different types of past context: for the arguments dimension, intense analysis is required for an accurate understanding of both side's arguments and counter-arguments, in which case the digested content analysis is beneficial; for the language dimension, however, the content analysis could underestimate or overestimate the language style of speeches, in which case a direct comparison between raw speeches may be better instead.

| Method | | DArt (S) ↓ | DArt (W) ↓ | Comp. ↑ |
|---|---|---|---|---|
| Direct | S | 49.99 | 51.16 | * |
| | C | 44.91 | 45.01 | * |
| Debatrix-H | S | 45.82 | 46.18 | * |
| | C | 44.18 | 44.03 | * |
| Debatrix-F | S | 48.61 | 48.71 | 34.85 |
| | C | 42.21 | 41.75 | 51.52 |

Table 8: Comparison between single-dimension and multi-dimensional collaborate approaches, using ChatGPT. Methods: S for Single, C for Collaborate; DArt (S): DebateArt debate score comparison RMSE; DArt (W): DebateArt debate winner prediction RMSE; Comp.: Competition debate winner prediction accuracy. We do not include Direct and Debatrix-H in competition debate results, as they cannot complete all judgments.

| Method | Single | Summary | Discussion |
|---|---|---|---|
| Accuracy | 34.85 | 42.42 | 51.52 |

Table 9: Winner prediction distribution on debate competition debates, using ChatGPT with Debatrix-F. Summary and Discussion are collaboration methods.

### 6.2 Single or Collaborate?

Besides running multiple columns in Debatrix, another method to obtain general winner predictions is to encode all dimensions into one *single* dimension and create a single column to perform all analyses. Table 8 compares these two approaches. While Debatrix still outperforms the Direct method using one single dimension, in all cases, collaborating multiple dimensions gives better results than merging them into a super dimension. This result illustrates the importance of multi-column design in Debatrix.

An unusual observation is that Debatrix-H supersedes Debatrix-F when using a single dimension. We argue that ChatGPT cannot properly summarize key information for multiple dimensions in one output, causing the content analysis to lose critical clues that may affect the final verdict. With multi-

| Model | Method | | DArt (S) ↓ | DArt (W) ↓ | Comp. ↑ |
|---|---|---|---|---|---|
| ChatGPT | Debatrix-F | C | 41.84 | 42.29 | 51.52 |
| GPT-4 | Direct | S | 44.14 | 46.39 | 34.85 |
| GPT-4 | Debatrix-F | S | 39.82 | 40.61 | 36.36 |
| GPT-4 | Debatrix-F | C | 36.07 | 37.58 | * |

Table 10: Comparison between ChatGPT with Debatrix-F and GPT-4. Notations follow Table 8. We did not conduct full experiments for GPT-4 with Debatrix-F due to the position bias issue.

| Model | Method | OG | OO | CG | CO |
|---|---|---|---|---|---|
| ChatGPT | Debatrix-F | 13 | 21 | 15 | 17 |
| GPT-4 | Direct | 0 | 0 | 10 | 56 |
| | Debatrix-F (S) | 0 | 5 | 9 | 52 |

Table 11: Winner prediction distribution on debate competition debates. OG, OO, CG, and CO are teams in BP debates, and their speaking order is OG, OO, OG, OO, CG, CO, CG, and finally CO.

dimensional collaboration, this issue is effectively solved, hence Debatrix-F performs better in this case.

### 6.3 Summary or Discussion?

In Section 3, we mentioned that column interaction could be an alternative way to combine multiple columns. Following Chan et al. (2023), we try to organize a discussion among columns based on their debate analyses to replace the one-step summary. Table 9 lists experiment results on competition debates. Although superseding the single-column setting, introducing discussion does not improve the performance compared to a direct summary; further investigation reveals that subsequent speakers tend to follow previous statements if some specific verdict is already included instead of combining their analyses. The result, while not completely denying the discussion method, suggests that such strategies could be better applied to rectify mistakes in analysis rather than giving a verdict.

### 6.4 Debatrix or GPT-4?

Table 10 compares ChatGPT with Debatrix-F and GPT-4 under various settings[4]. It can be seen that ChatGPT with Debatrix-F outperforms naive GPT-4 using single dimension; while introducing Debatrix helps GPT-4 reach the best performance on DebateArt debates, it does not gain much improvement on competition debates.

Table 11 summarizes the predicted winner in all complete runs for each model and method combination to further investigate this outcome. Surprisingly, while ChatGPT gives relatively balanced predictions, GPT-4 always predicts the closing side (CG and CO), in most cases CO; CO is the speaker of the last speech in the debate; even Debatrix can hardly change this. In all cases, they do not match the true winner distribution (Table 4), excluding the potential cause of imbalanced labels.

We conjecture that position bias (Ko et al., 2020; Wang et al., 2023b) could be a major factor that causes LLMs as powerful as GPT-4 to fail in judging BP debates. When arguments from all debaters are similarly strong, LLM may prefer the last speaker who can refute other debaters while not being refuted by others, thus seemingly more convincing.

## 7 Conclusion

In this paper, we propose a fine-grained debate judging framework based on LLM, Debatrix. We decompose the debate judging task into a speech-by-speech analysis to tackle multi-turn, long debates and elaborate multiple dimensions to generate systematic judgments. We introduce a novel debate judging benchmark to assess our framework and other automatic debate judging approaches, covering multi-dimensional and multi-debater scenarios. Under both settings, Debatrix significantly improves ChatGPT, aiding it in judging long debates that exceed the context window and outperforming bare GPT-4.

[4]Due to the high cost of GPT-4, we only conduct experiments on 30 out of 100 DebateArt debates and all competition debates.

## Limitations

Despite the above results, this paper has a few limitations for which we appreciate future studies. First, the workflow implementation in Debatrix, while already showing competence, can be further studied and polished. Second, the position bias issue on GPT-4 for BP debates remains even when Debatrix is applied, calling for a more powerful tool to rectify this phenomenon.

## References


Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

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

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations?

Adrian de Wynter and Tommy Yuan. 2023. I wish to have an argument: Argumentative reasoning in large language models.

Ivan Donadello, Anthony Hunter, Stefano Teso, and Mauro Dragoni. 2022. Machine learning for utility prediction in argument-based computational persuasion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5592–5599.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons.

Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Trans. Interact. Intell. Syst.*, 6(4).

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger,

Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Rosemary J. Thomas, Judith Masthoff, and Nir Oren. 2019a. Can i influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. *Frontiers in Artificial Intelligence*, 2.

Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. 2017. Adapting healthy eating messages to personality. In *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, pages 119–132, Cham. Springer International Publishing.

Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. 2019b. Is argumessage effective? a critical evaluation of the persuasive message generation system. In *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems: 14th International Conference, PERSUASIVE 2019, Limassol, Cyprus, April 9–11, 2019, Proceedings*, volume 11433, pages 87–99. Springer.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

World Universities Debating Council. 2023. *World Universities Debating Championships Debating & Judging Manual*.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges.

# A Debatrix Column Workflow Algorithm

Algorithm 1 demonstrates the complete workflow of Debatrix chronological columns. Here we do not include multi-column collaboration.

# B Benchmark Data Details

## B.1 DebateArt Debate Procedure[5]

Users instigate debates in DebateArt. The instigator needs to provide a debate topic and set debate configurations such as character limit, time limit (12 hours to 2 weeks), and the number of rounds (up to 5); they can also include a description with pertinent details like definitions, expanded resolution, special rules, and scope limitations.

The instigator may elect to be pro or con, leaving the other position to the contender; the contender can be any community user willing to accept the challenge or another user requested directly by the instigator. No matter which case, once the contender enters, the debate starts, and both sides publish their arguments. If any side fails to propose an argument within the time limit, they will automatically forfeit the round; in our work, debates with forfeited turns are treated as incomplete.

When all arguments have been published, the community or the appointed judges select the debate's winner by voting. Voters need to follow the specified voting system and give fair verdicts. The debate is finished when the winner has been selected according to the votes.

## B.2 DebateArt Debate Collection

To collect debates with valid content and votes, we first crawled the list of finished (result announced) debates on DebateArt and then filtered out debates that had no valid votes or were interrupted (not using all preset rounds). Next, we crawled the debate details of the remaining debates, including topic (motion), debaters, description (info slide), arguments (speeches), and votes. The raw arguments are in HTML format; hence, we use

[5]This section mainly refers to https://info.debateart.com/help/debates.

**Algorithm 1:** Debatrix Chronological Column Workflow

**input :** judge preference $P$, debate motion $M$, a list of debaters $\{D_1, \ldots, D_m\}$, an info slide $I$, a list of speaker-speech tuples $\{(d_1, s_1), \ldots, (d_n, s_n)\}$

**output :** a list of speech comments $\{C_{s_1}, \ldots, C_{s_n}\}$, a list of debater comments $\{C_{D_1}, \ldots, C_{D_m}\}$, a winner verdict $V$

sc_mem $\leftarrow \emptyset$; // speech context memory

ca_mem $\leftarrow \emptyset$; // content analysis memory

ci $\leftarrow (P, M, \{D_1, \ldots, D_m\}, I)$; // common inputs

**for** $i \leftarrow 1$ **to** $n$ **do**

  rel_sc $\leftarrow$ `QuerySpeechContext(`sc_mem, $s_i$`);`
  // query relevant contents to new speech

  ca_mem $\leftarrow$ ca_mem $\cup \{(d_i, s_i)\}$;

  ca $\leftarrow$ `AnalyzeContent(`ci, $d_i$, $s_i$, ca_mem, rel_sc`);` // analyze speech

  ca_mem $\leftarrow$ ca_mem $\cup$ ca;

  $C_{s_i}$ $\leftarrow$ `JudgeSpeech(`ci, $d_i$, ca`);` // judge speech

**end**

all_ca $\leftarrow$ `FetchContentAnalysis(`ca_mem`);`
// fetch all content analyses

da $\leftarrow$ `AnalyzeDebate(`ci, all_ca`);` // analyze debate

**for** $i \leftarrow 1$ **to** $m$ **do**

  $C_{D_i}$ $\leftarrow$ `JudgeDebater(`ci, $D_i$, da`);` // judge debater

**end**

$V$ $\leftarrow$ `DecideWinner(`ci, da`);` // decide winner of debate

**return** $\{C_{s_1}, \ldots, C_{s_n}\}, \{C_{D_1}, \ldots, C_{D_m}\}, V$

`markdownify`[6] to convert them into Markdown documents.

We further filtered out debates that do not fit in our benchmark, including debates that do not use the categorical point assignment system, are not formal, and contain very short speeches, which may indicate a forfeit. Although many are close to being professionals, we also excluded some very long debates to ensure that inputs do not exceed the LLM's context window during the experiment.

### B.3 BP Debates

The British Parliamentary (BP) format is a widely accepted competitive debate style, followed by famous competitions like WUDC, EUDC, and NAUDC (World Universities Debating Council, 2023). Each BP debate contains four teams, with a total of eight debaters. There are two teams on each side of the debate: on one side are Opening Government (OG) and Closing Government (CG); on the other side are Opening Opposition (OO) and Closing Opposition (CO). They follow the order specified below to give speeches:

- First speaker (the "Prime Minister") from OG;
- First speaker (the "Leader of Opposition") from OO;
- Second speaker (the "Deputy Prime Minister") from OG;
- Second speaker (the "Deputy Leader of Opposition") from OO;
- First speaker (the "Government Member") from CG;
- First speaker (the "Opposition Member") from CO;
- Second speaker (the "Government Whip") from CG;
- Second speaker (the "Opposition Whip") from CO.

Each speech lasts for 7 minutes, with limited tolerance for timeouts. In general, OG should define the motion, propose arguments, and refute arguments from OO; OO should rebut OG's case and propose constructive arguments for their side; CG and CO should provide further supplementary analysis in favor of their side, respectively.

The Points of Information (POI) is a special feature in BP debates. A POI is a formalized interjection from any debater on the opposite side to the current speaker. The current speaker has the right to decide whether the POI is accepted or rejected; once accepted, the debater offering the POI can make an argument or ask a question within 15 seconds, and the current speaker should respond properly before continuing their speech. In our benchmark, we mark POI conversations as quoting blocks and prompt LLMs to pay attention to them, as engaging in POIs may contribute to the debaters' overall performance.

### B.4 Competition Debate Collection

Many debate competitions, including WUDC, EUDC, and NAUDC, only provide essential information about debates, such as motions, info slides, teams, and winners. They do not have official transcriptions; unofficial ones are often incomplete. Fortunately, in recent years, many of these competitions have provided official video recordings of debates not long before the finals.

We selected debates starting from the quarter finals from WUDC (2020-2023), EUDC (2019-2022) and NAUDC (2021 and 2023), and downloaded their video recordings. Next, we extracted audio files from the recordings and used Whisper (Radford et al., 2023) to recognize the speeches. We manually checked and formatted the output results into valid transcriptions.

Due to missing, damaged, or incomplete recordings, not every debate was available for transcription; we only kept debates whose transcription was complete. Finally, we merged their transcriptions with debate information to produce the final data.

[6] https://github.com/matthewwithanm/python-markdownify