Debate-Feedback: A Multi-Agent Framework for Efficient Legal Judgment Prediction

Xi Chen `xich0108@bu.edu`
Mao Mao `maomao@bu.edu`
Shuo Li `lis23@m.fudan.edu.cn`
Haotian Shangguan `haosg19@bu.edu`

Abstract

The use of AI in legal analysis and prediction (LegalAI) has gained widespread attention, with past research focusing on retrieval-based methods and fine-tuning large models. However, these approaches often require large datasets and underutilize the capabilities of modern large language models (LLMs). In this paper, inspired by the debate phase of real courtroom trials, we propose a novel legal judgment prediction model based on the Debate-Feedback architecture, which integrates LLM multi-agent debate and reliability evaluation models. Unlike traditional methods, our model achieves significant improvements in efficiency by minimizing the need for large historical datasets, thus offering a lightweight yet robust solution. Comparative experiments show that it outperforms several general-purpose and domain-specific legal models, offering a dynamic reasoning process and a promising direction for future LegalAI research. Our code is released at `https://github.com/Xi7997/Debate_Feedback`.

1 Introduction

LegalAI leverages artificial intelligence technologies such as natural language processing, machine learning, and deep learning to address various legal tasks (Aletras et al., 2016; Katz et al., 2017; Zhong et al., 2020), including legal document analysis and consultation. A key area of LegalAI is Legal Judgment Prediction (LJP) (Zhong et al., 2018a; Ma et al., 2021; Cui et al., 2023), which focuses on predicting court judgments. LJP tasks typically use historical legal case data, including background information, case descriptions, statements from both parties, precedents, and court verdicts. Predictions range from binary outcomes (e.g., plaintiff vs. defendant wins) to multi-class tasks (e.g., sentence prediction). NLP technologies, combined with advanced models like LegalBERT (Chalkidis et al., 2019) and Lawformer (Xiao et al., 2021), have achieved strong results by learning from large datasets.

The debate model is a system that integrates large language modeling (LLM) with argumentative reasoning techniques to simulate the process of debate or contention (Irving et al., 2018; Nie et al., 2020), ultimately arriving at a decision or conclusion on a specific issue through the debate process. In a typical debate task, multiple LLM agents assume different roles and are deliberately guided to provide answers from various perspectives or positions. These generated arguments are then synthesized to assist the LLM in reaching a final conclusion (Zeng et al., 2022).

In this paper, we propose a Debate-Feedback model to explore an efficient and convenient method for predicting legal judgement. Fig[1] shows the general framework of the model in the task of predicting decision results. Specifically, Debate-Feedback can be divided into four steps. First, the collected historical legal cases L_i will be formatted into Case Background C_i, Plaintiff Claim P_i and Defendant Statement D_i. These information will be provided to the judge LLM for initial prediction. In the second step of the debate, multiple LLM agents will be guided to answer the prediction questions from different perspectives, and then exchange opinions and debate to generate their own comments E_i. In the verification phase, a pre-trained assistant model $\mathcal{E}$ will conduct a reliability analysis on each LLM's comments combined with case information. The results of the analysis will be provided to the judge LLM for reference together with each agent's comments. The judge LLM will give the prediction O_i for this round based on the above information $\mathcal{E} = E_i \oplus L_i$. More details are illustrated in the Methodology section. In summary, we introduce a Debate-Feedback model that enhances legal judgment prediction by incorporating a multi-agent debate process and reliability evaluation, providing a more efficient and

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 462–470
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

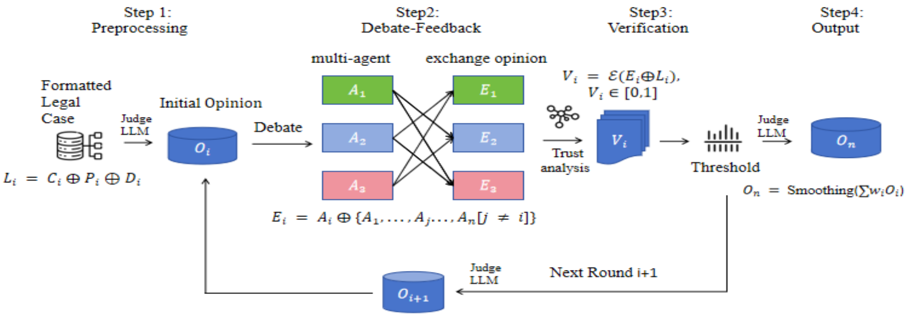

Figure 1: A brief introduction of Debate-Feedback Structure

accurate solution with reduced reliance on large datasets.

2 Related Work

Legal documents are characterized by lengthy texts and complex logic, which has led prior research to focus on two key approaches to address these challenges: training legal LLM and using retrieval augmentation.

2.1 Legal LLM

In-context Learning(ICL) is a learning paradigm widely applied in large language models (LLMs) by using a set of context examples to guide predictions during reasoning (Dong et al., 2024; Liu et al., 2021; Gutierrez-Pachas et al., 2022; Min et al., 2022). However, due to the often extensive length of legal texts, naive ICL methods are constrained by LLM input length limits. As a result, LegalAI solutions typically combine ICL with fine-tuning or pre-training of models to overcome these limitations. For instance, LegalBERT (Chalkidis et al., 2019) fine-tunes BERT on legal datasets, achieving strong results in legal text classification and provision retrieval. Similarly, Lawformer (Xiao et al., 2021) handles lengthy Chinese legal documents, while CaseLaw-BERT (Paul et al., 2023), fine-tuned on case law datasets, enhances legal case retrieval and judgment prediction. Despite their success, these approaches rely heavily on large, domain-specific datasets, which can limit their applicability across different legal systems and languages.

2.2 Retrieval Augmentation

Retrieving relevant legal precedents—court judgments or legal decisions from previous cases—is a mainstream approach to assist LLMs in making predictions, especially in overcoming the challenge of lengthy texts. By providing recommended samples, this method guides the LLM's reasoning process more effectively (Zhong et al., 2020; Huang et al., 2021). Ma et al. introduced a framework that deeply integrates legal precedents into judgment prediction (Wu et al., 2023), combining the reasoning capabilities of LLMs with domain-specific models to enable more accurate and context-aware predictions. Similarly, Caseformer (Su et al., 2024) employs a pre-training strategy that emphasizes distinctions between cases, enhancing case retrieval performance. Although retrieval augmentation improves the handling of long texts, it still relies on the availability of large datasets, and its reliance on specific legal systems and languages can limit broader applicability across different jurisdictions.

3 Methodology

In this section, we first systematically introduce our Feedback-Debate model, followed by an analysis of the limitations of the general debate architecture in specific legal scenarios, along with proposed solutions to address these shortcomings.

Overview Algorithm[1] presents the pseudo code for the debate-feedback framework in binary classification. The input is a preprocessed legal event text, labeled as S, and the main language model (LM) plays the role of the judge, predicting the probability of a legal judgment, $LM : S \rightarrow [0, 1]$. Two agents, t_{ne} and t_{po}, debate from opposing perspectives, providing inputs to refine the judgment. Each debate round involves these agents exchanging and debating their positions, with n defining the number of iterations.

The assistant model $\mathcal{E}$ evaluates the reliability of the agents' arguments and outputs a probability. If the reliability exceeds a threshold, the main LM adjusts its prediction by weighting the latest information, otherwise it defaults to the initial prediction. The final decision is smoothed over all

```
Algorithm 1: Debate-Feedback
Input: LM, $\mathcal{E} : \mathcal{S} \to [0,1]$; $n, T \in N$;
       $x \in S$ ; $t_{ne}, t_{po} : \mathcal{S} \to \mathcal{S}$;
Output: Final decision $y \in (0,1)$;
$O_0 \leftarrow LM(x)$;
for $i \leftarrow 1$ to $n$ do
    // Debate Step
    $a : a_{ne}, a_{po} \leftarrow t_{ne}(x), t_{po}(x)$;
    $e : e_{ne}, e_{po} \leftarrow t_{ne}(x \oplus a_{po}), t_{po}(x \oplus a_{ne})$;

    // Verification Step
    $v : v_{ne}, v_{po} \leftarrow \mathcal{E}(e_{ne}), \mathcal{E}(e_{po})$;
        $sum = LM(a, e, v)$;
    if Threshold(v) then
        $O_i =$
          $(1-T) * O_{i-1} + T * LM(x, sum)$;
    end
    else
        $O_i = LM(x)$;
    end
end
$y \leftarrow O_n$;
```

TrainingSet of Assistant model	
Training_X	{Case_background + Debater's opinion}
Training_Y	{Ground_truth **XOR** Debater's position}

Table 1: Dataset of assistant model.

rounds to produce a stable outcome. (Note that notation $\oplus$ does not mean xor, but rather combination in a non-additive sense.)

Reliability Analysis Through experiments, we observe that a simple debate model can sometimes lead to worse prediction results. This occurs because legal predictions differ from mathematical problems, as they often involve subjective tendencies. A straightforward example is when we guide multiple LLMs to debate from the perspectives of the plaintiff and defendant, it is challenging for them to reach a consensus. To address this issue, one of our solutions is to train an assistant model that learns from a large corpus of legal event annotations and assists in evaluating the reliability of different debate arguments, as shown in Table[1]. Specifically, the training set for the assistant model is generated from multiple runs of the unassisted Debate-Feedback model, which we refer to as Debate-Feedback (single) in the subsequent experimental section.

Smoothing Operation To mitigate the impact of a "failed" debate where the main LLM generates incorrect answers, we apply a smoothing operation. This involves saving the results of each prediction and assigning them a certain weight. Specifically, let $LM(x)$ represent the predicted result of the i-th debate and T be the weighting factor. The updated result is calculated as:

$$O_i \leftarrow (1-T) * O_{i-1} + T * LM(x) \quad (1)$$

where $T \in [0,1]$ represents the weight assigned to the latest prediction.

4 Experiment

4.1 Dateset and Baseline

Along with many influential LegalAI works, we also use CaseLaw as the main dataset. The **CaseLaw** dataset is a legal case dataset specifically used for natural language processing (NLP) and machine learning tasks in the legal field, especially in the fields of legal case retrieval and legal judgment prediction. This dataset contains a large number of court case texts that have been judged, usually including descriptions of legal facts, legal reasoning, and judgment results. In order to test the model's cross-language and cross-legal capabilities, we also used the Chinese dataset **CAIL18** (Xiao et al., 2018; Zhong et al., 2018b).

We compare Debate-Feedback with both general large language models and legal domain models. **GPT4o** and **GPT3.5-turbo** are representative general large language models at present (OpenAI et al., 2024), and they have been proven to have strong text analysis and logical reasoning capabilities. **LegalBert** (Chalkidis et al., 2019) and **Lawformer** (Xiao et al., 2021) are well-known legal domain model, they're able to capture the association between legal terms and cases well. In addition, **CNN** (Lecun et al., 1998) is also used as a classifier for feature extraction in the baseline evaluation, with **BERT** (Devlin et al., 2019) serving as the text embedding layer.

Considering that the debate-feedback framework can essentially be seen as a large language model reasoning framework, we also compare it with classic reasoning methods, including **Few-shot Learning**, **Chain of Thought(CoT)** (Wei et al., 2023) and **Reflexion** (Shinn et al., 2023). We use gpt-4o mini as the baseline model in this part and verified them on a smaller subset on a smaller subset of the datasets (12,000 samples from CaseLaw and 3,000 samples from CAIL18).

4.2 Regular LJP tasks

Trial Prediction The input for trial prediction includes a legal text, along with the opinions of the plaintiff and defendant. The predicted labels are Plaintiff wins, Defendant wins, Settlement, and Dismissed. Since Settlement and Dismissed are explicitly stated in the legal text, this can be reduced to a binary classification task with two labels: Plaintiff wins and Defendant wins. The CaseLaw dataset was used for this task, and Table[4] provides a sample.

Article Prediction Article prediction is a multi-label classification task. The model receives a description of legal facts and the prediction content contains multiple labels of different relevant law articles. CAIL18 dataset is used in this task.

4.3 Evaluation Metrics

In this study, we evaluate the model performance using two key metrics: accuracy and F1-score.

Accuracy(Acc) is the proportion of correct predictions among all predictions. It is computed as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N}(y_i = y_{\text{true},i})}{N} \quad (2)$$

where N is the total number of predictions, y_i is the predicted label, $y_{\text{true},i}$ is the actual label, and $(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise.

F1-score(F1) is useful for imbalanced datasets as it balances precision and recall. In multi-class classification, F1-score is computed for each class and then averaged (macro F1-score). For a single class, F1-score is given by:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Where precision and recall are defined as:

$$\text{Precision} = \frac{\sum_{i=1}^{N} 1(y_i = c \wedge y_{\text{true},i} = c)}{\sum_{i=1}^{N} 1(y_i = c)} \quad (4)$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} 1(y_i = c \wedge y_{\text{true},i} = c)}{\sum_{i=1}^{N} 1(y_{\text{true},i} = c)} \quad (5)$$

For multi-class classification, the macro F1-score is calculated as the average F1-scores for all classes:

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} F1_c \quad (6)$$

where C is the number of classes.

Model	CaseLaw		CAIL18	
	Acc	F1	Acc	F1
CNN(with BERT)	0.58	0.54	0.39	0.11
Legal-BERT	0.63	0.61	0.22	0.03
Lawformer	0.53	0.31	0.38	0.12
GPT-3.5-turbo	0.49	0.27	0.26	0.04
GPT-4o	0.64	0.64	0.31	0.05
Debate-Feedback(single)	0.66	0.65	0.42	**0.16**
Debate-Feedback(assistant)	**0.67**	**0.66**	**0.45**	**0.16**

Table 2: Comparison of models on CaseLaw and CAIL18 datasets. All judge's and debaters' LMs in experiments are based on the GPT-4o model and T = 0.5.

4.4 Experimental Results

The experimental results demonstrate the effectiveness of the Debate-Feedback model, with the inclusion of an assistant model in the feedback loop enhancing prediction reliability and providing more robust results compared to the single Debate-Feedback model. These results validate the strength of our approach in improving the accuracy and consistency of legal judgment predictions. Our experimental results are shown in Table[2], Figure[2] and Figure[3].

CaseLaw Dataset Performance For the CaseLaw dataset, the Debate-Feedback model outperformed GPT-4o, GPT-3.5-turbo, Legal-BERT, CNN and Lawformer. The model with the assistant achieved an accuracy of 0.67 and an F1-score of 0.66, while the single Debate-Feedback model obtained slightly lower performance with an accuracy of 0.66 and an F1-score of 0.65. These results show that our method improves the performance of pre-train legal domain models, which only achieved an accuracy of 0.63 and an F1-score of 0.61. The assistant model's inclusion in the feedback loop improves the reliability of predictions, making it more robust compared to the single model.

CAIL18 Dataset Performance On the Chinese legal dataset CAIL18, the Debate-Feedback model achieved a remarkable accuracy of 0.45, significantly surpassing GPT-4o (accuracy 0.31) and GPT-3.5-turbo (accuracy 0.26). The model with an assistant component further improved the F1-score to 0.16, highlighting the ability of the assistant model to refine predictions and correct any inconsistencies in the debate phase. These results also suggest that the Debate-Feedback model is more versatile in handling cross-linguistic challenges compared to other models.

Model	CaseLaw		CAIL18	
	Acc	F1	Acc	F1
Few-shot	63.8%	64.1%	29.7%	5.03%
CoT (4-steps)	63.7%	64.0%	31.2%	6.17%
Reflexion	64.5%	65.0%	31.8%	8.12%
Debate-Feedback (single)	**66.2%**	**65.7%**	**41.9%**	16.1%
Debate-Feedback (assistant)	**67.1%**	**66.1%**	**44.8%**	**16.3%**

Table 3: Performance comparison of different reasoning methods on CaseLaw and CAIL18 datasets.

Comparison with basic reasoning methods As shown in table[3], Debate-Feedback structure achieves significant advantages in comparison with several basic reasoning frameworks. The results show that Chain-of-Thought and Reflection perform only marginally better than Zeroshot, while our Debate-feedback framework consistently demonstrates superior performance, reinforcing the conclusions of our original experiments.

We believe there are two primary reasons why standard reasoning techniques like CoT and Reflection are less effective for this type of legal prediction problem:

Complexity of Legal Texts: The legal text itself is lengthy and logically complex, and simple prompts are difficult to be effective.

Nature of Legal Prediction: Legal prediction is always different from logical reasoning. It is not a step-by-step thinking toward the correct answer, but usually a discussion to unify or compromise the views of multiple parties. This is precisely why we designed the Debate-feedback framework, which is tailored to handle such tasks.

5 Conclusion

We propose a debate-feedback model based on LLMs for legal judgment prediction and demonstrated its feasibility through experiments. The inclusion of an assistant model and reliability analysis enhances prediction robustness. Future work could explore the application of debate models in other fields or further integrate them with LLMs.

6 Limitations

Our work currently has the following limitations:

(a) The experiments were limited to two datasets and two specific tasks, broader evaluations across additional datasets and tasks are necessary to fully validate the model's robustness and generalizability in different legal contexts.

(b) While the smoothing technique and assistant model (reliability analysis) were included in the framework, their individual contributions to the overall performance were not deeply investigated.

(c) This work does not integrate retrieval argument techniques, which presents a promising direction for future research to enhance the model's performance.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Ilias Chalkidis, Michael Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Legalbert: The muppets straight out of law school. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2898–2904.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Daniel A. Gutierrez-Pachas, Eduardo F. Costa, and Alessandro N. Vargas. 2022. Distribution of a markov chain in reverse-time with cluster observations in the extremes of a finite time window. *Preprint*, arXiv:2206.05607.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, and Matthias Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 79–88, New York, NY, USA. Association for Computing Machinery.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Daniel Katz, Michael Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Preprint*, arXiv:2107.13586.

Jiayuan Ma, Chao Liu, Furu Wei, and Deheng Huang. 2021. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4562–4571.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. *Preprint*, arXiv:2110.15943.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *Preprint*, arXiv:1910.14599.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law. *Preprint*, arXiv:2209.06049.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Weihang Su, Qingyao Ai, Yueyue Wu, Yixiao Ma, Haitao Li, Yiqun Liu, Zhijing Wu, and Min Zhang. 2024. Caseformer: Pre-training for legal case retrieval based on inter-case distinctions. *Preprint*, arXiv:2311.00333.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *Preprint*, arXiv:2105.03887.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *Preprint*, arXiv:1807.02478.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *Preprint*, arXiv:2204.00598.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018a. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018b. Overview of cail2018: Legal judgment prediction competition. *Preprint*, arXiv:1810.05851.

Hongyu Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

A Appendix

1. The choices about different numbers of rounds and debaters on the debate-feedback model (without assistant model).

As illustrated in Figures[2] and Figures[3], while the number of debaters and debate rounds may vary depending on the specific task, generally, using 2-4 debaters and conducting 2-3 rounds often yields favorable results. This configuration can serve as a useful reference for readers, helping to avoid unnecessary computational overhead.

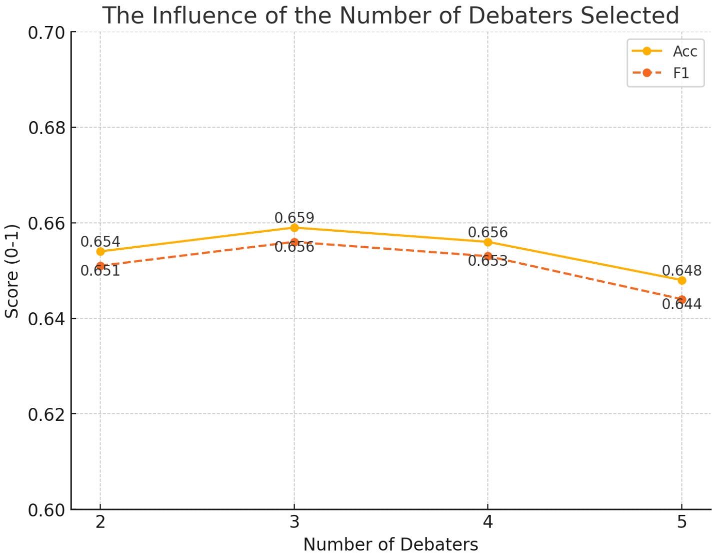

Figure 2: Influence of the number of debaters selected.

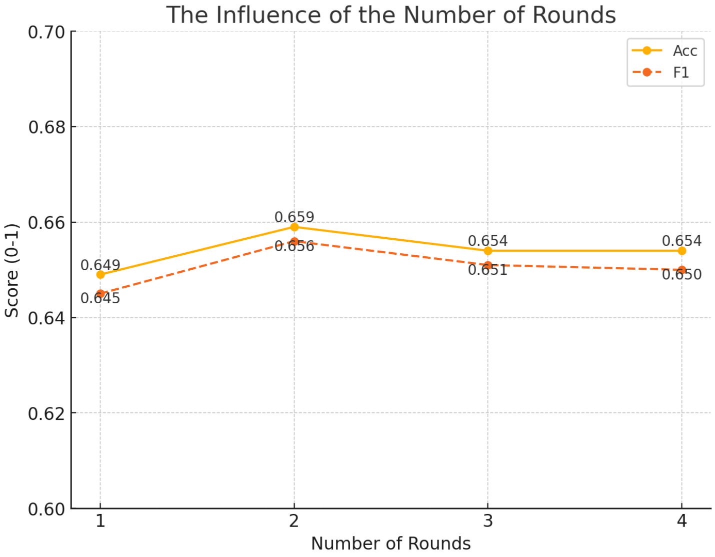

Figure 3: Influence of the number of rounds selected.

2. A sample of Debate-Feedback Structure with one round and three debaters in binary classification task, table[4].

<table>
<tr><td colspan="3">1. Initial Step</td></tr>
<tr><td colspan="3"># input:
Here is a legal case, your task is to predict the result of the judge, the labels can be one of the following:
[\'Plaintiff wins\',\'Defendant wins\'].
ONLY output the one of the labels without anything else.
Here is the case background:
The case, "The Adams Express Company v. Richard T. Haynes," took place in 1865, involving a transaction where Richard T. Haynes entrusted a trunk to the Adams Express Company for transport from Nashville, Tennessee, to Chicago, Illinois. Haynes claimed that the contents of the trunk were lost due to the negligence and fraudulent actions of the express company. The express company provided a receipt outlining the terms of service, which limited their liability ... "</td></tr>
<tr><td colspan="3"># output:
Plaintiff wins</td></tr>
<tr><td colspan="3">2. Debate Step</td></tr>
<tr><td>Debater 0</td><td>Debater 1</td><td>Debater 2</td></tr>
<tr><td># input:
Here is a legal case, your task is to predict the result of the judge and Debate withother experts, the labels can be one of the following:
[\'Plaintiff wins\',\'Defendant wins\'].
Please state your opinion and refer to factual examples whenever possible.
Your answer no more than 200 words.
Here is the case background:
...

output:
I predict that the judge will rule in favor of the plaintiff, Richard T. Haynes. Although the express company presented a receipt with terms limiting its liability, the crux of the case centers on the allegations of gross negligence and possible fraudulent actions ...</td><td># input:
Here is a legal case, your task is to predict the result of the judge and Debate withother experts, the labels can be one of the following:
[\'Plaintiff wins\',\'Defendant wins\'].
Please state your opinion and refer to factual examples whenever possible.
Your answer no more than 200 words.
Here is the case background:
...

output
In the case of "The Adams Express Company v. Richard T. Haynes," it is likely that the plaintiff, Haynes, would prevail due to the allegations of gross negligence and fraud. Despite the express company\'s attempt to enforce the terms of limited liability, these clauses typically ...</td><td># input:
Here is a legal case, your task is to predict the result of the judge and Debate withother experts, the labels can be one of the following:
[\'Plaintiff wins\',\'Defendant wins\'].
Please state your opinion and refer to factual examples whenever possible.
Your answer no more than 200 words.
Here is the case background:
...

output
In the case of "The Adams Express Company v. Richard T. Haynes," I predict that the judge will rule in favor of the plaintiff, Richard T. Haynes. The key issue revolves around whether the express company\'s actions constituted gross negligence or fraud, which would negate their ...</td></tr>
<tr><td># input
Now I will give you the opinion of another expert, if their opinions differ from yours, you can support their views or offer a counter-argument.
This round of conversation will be provided to the judge for reference, so remember to express your stance clearly.
Your answer no more than 200 words.
Here are their opinions:
...

output
I largely agree with the perspective of the first expert regarding the implications of gross negligence and fraud in "The Adams Express Company v. Richard T. Haynes." The pivotal issue is whether Haynes can convincingly demonstrate
...</td><td># input
Now I will give you the opinion of another expert, if their opinions differ from yours, you can support their views or offer a counter-argument.
This round of conversation will be provided to the judge for reference, so remember to express your stance clearly.
Your answer no more than 200 words.
Here are their opinions:
...

output
I strongly agree with the assessment of the first expert regarding the likelihood of the plaintiff, Richard T. Haynes, prevailing in "The Adams Express Company v. Richard T. Haynes." The focus on the express company\'s
...</td><td># input
Now I will give you the opinion of another expert, if their opinions differ from yours, you can support their views or offer a counter-argument.
This round of conversation will be provided to the judge for reference, so remember to express your stance clearly.
Your answer no more than 200 words.
Here are their opinions:
...

output
I find that I am in strong agreement with the opinion of the second expert regarding the case of "The Adams Express Company v. Richard T. Haynes." The emphasis on the express company\'s potential failure to demonstrate
...</td></tr>
<tr><td colspan="3">3. Verification Step</td></tr>
<tr><td>{input: the summary of Debater 0's statement to the assistant model}
{output: reliability: 0.75}</td><td>{input: the summary of Debater 1's statement to the assistant model}
{output: reliability: 0.8}</td><td>{input: the summary of Debater 2's statement to the assistant model}
{output: reliability: 0.45}</td></tr>
<tr><td colspan="3">4. Output step</td></tr>
<tr><td colspan="3"># input:
{Follow the conversation from Step 1}
Some experts discussed the case in this round and here is the summary, please provide your final opinion.
Same as above, ONLY output the one of the labels without anything else.
Here is the debate summary:
[\'In the case of ...]
{insert the reliability from Step 3}

output
Plaintiff wins</td></tr>
</table>

Table 4: An Example of Debate-Feedback Structure

3. Performance of the smoothing mechanism.

Debate-Feedback Mechanism	**Prediction Correction**	**Prediction Degradation**	**Accuracy Rate**
Without Smoothing	102	115	62.8%
With Smoothing	93	11	65.7%

Table 5: Performance of smoothing mechanism.

In our initial experiments, we unexpectedly discovered that a simple smoothing operation was particularly useful in improving prediction accuracy. Specifically, we tested the Prediction Correction Rate and Prediction Degradation Rate with and without smoothing on a binary CaseLaw dataset containing 3000 samples, as shown in table[5].

- **Prediction Correction**: When the initial prediction of the model is wrong, and it is corrected by the debate-feedback framework.
- **Prediction Degradation**: When the initial prediction of the model is correct, but becomes incorrect due to the framework.

We found that the Prediction Degradation Rate was particularly high without smoothing, while the Prediction Correction Rate was about the same. This means the smoothing mechanism helps models avoid relying too heavily on the influence of a certain debater.