Courtroom-LLM: An Innovative Courtroom-Analogous Framework for Text Classification

Anonymous ACL submission

Abstract

In this research, we introduce the Courtroom-LLM framework, a novel multi-LLM structure inspired by legal courtroom processes, aiming to enhance decision-making in ambiguous text classification scenarios. Our approach simulates a courtroom setting within LLMs, assigning roles similar to those of prosecutors, defense attorneys, and judges, to facilitate comprehensive analysis of complex textual cases. We demonstrate that this structured multi-LLM setup can significantly improve decision-making accuracy, particularly in ambiguous situations, by harnessing the synergistic effects of diverse LLM arguments. Our results from thorough evaluations on various NLP tasks show that the Courtroom-LLM framework surpasses both conventional single LLM classifiers and basic structured multi-LLM systems, underscoring the benefits of our legal proceedings-inspired model in enhancing NLP decision-making.

1 Introduction

In the field of natural language processing (NLP), the challenge of resolving ambiguous cases, where the correct answer is not clear-cut, remains a significant hurdle. This is where our research introduces a groundbreaking solution: the Courtroom-LLM framework. Inspired by the decision-making processes in legal courtrooms, this framework brings a novel approach to handling ambiguity in NLP, particularly in classification tasks.

This framework adopts a legal courtroom's procedural model, assigning roles analogous to judges, prosecutors, and defense attorneys within a Large Language Model (LLM) setup (Figure 1). Such a structure allows for a more rigorous and multi-faceted analysis of textual data, especially when dealing with ambiguous or unclear cases.

The methodology section of this paper outlines the operational phases of the Courtroom-LLM: Preliminary Hearing and Main Trial. This two-phase approach ensures a thorough evaluation, akin to the judicial process, where every potential interpretation is considered before reaching a verdict.

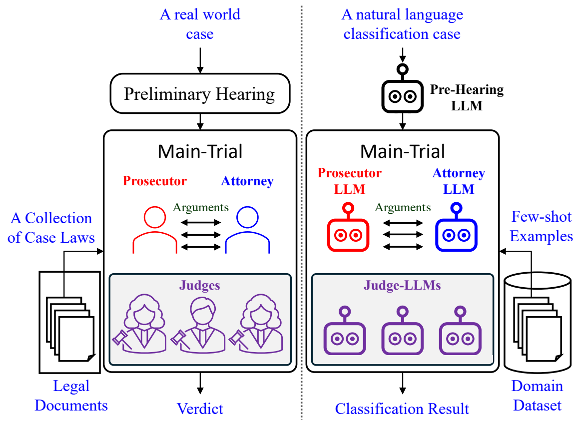


Figure 1: Comparison of the traditional courtroom system with our Courtroom-LLM framework.

Our extensive experimentation across diverse NLP classification tasks, spanning categories such as natural language understanding (RTE, BoolQ), natural language inference (QNLI, ANLI), and text classification (Emotion), has resulted in remarkably promising outcomes. Our experiments revealed that the Courtroom-LLM framework consistently outperforms both traditional single-LLM methods and simple parallel multi-LLM configurations, marking a striking average accuracy improvement of up to 13% over baseline experiments. In ambiguous cases where classification is tougher, the Courtroom framework showed a notable 35% performance boost.

In this paper, Section 2 discusses related work, setting the stage for our research. Section 3 details the Courtroom-LLM framework and its mechanics. Section 4 validates our approach through experiments in NLP classification. Section 5 concludes by discussing our findings and future research directions in NLP.

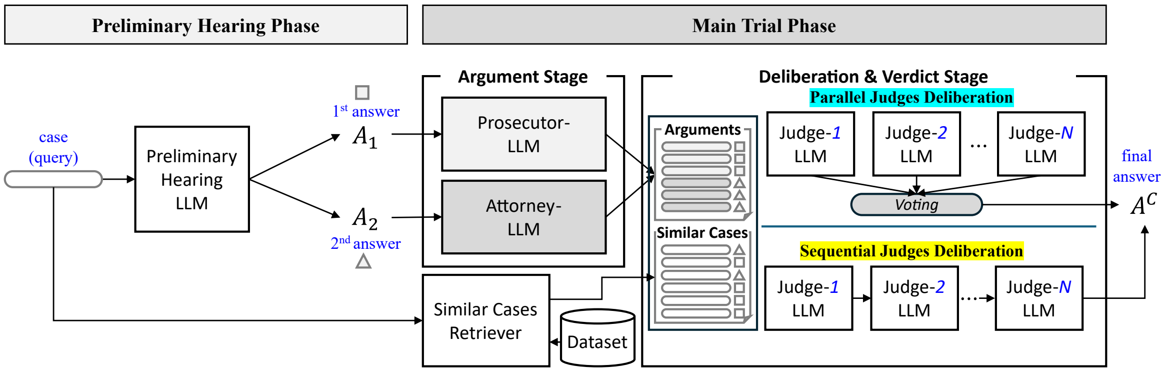


Figure 2: Overall architecture of Courtroom-LLM framework.

2 Related Works

Recent efforts to improve large language models include enhancing input prompts for precision, enriching queries with context, and considering changes to LLM structures for more accurate responses.

One of the most extensively studied research directions is *prompt engineering*, which has become crucial across various tasks. Innovations in this field involve adding sequential and systematic prompts that guide response generation and optimizing the order of prompts to improve results(Mao et al., 2023). Significant advancements include the use of chain-of-thought reasoning(Wei et al., 2022), providing *step-by-step* or *take-a-deep-breath* instructions(Shaikh et al., 2023; Yang et al., 2023), and abstracting initial queries to derive meaningful prompt blocks(Zheng et al., 2023).

To enhance LLMs' decision accuracy, recent approaches have included *additional information*, such as through retrieval functionalities. This supplementary information often comes from external search engines or internal databases(Lewis et al., 2020), employing algorithms like BM25(Yu et al., 2023) or measuring semantic textual similarities(Majumder et al., 2016).

Research on varying *LLM connection structures* includes methods like querying multiple LLMs(Li et al., 2024) and refining the answers through post-processing, simulating real-world debates among LLMs to converge on a consensus(Yao et al., 2023; Pi et al., 2022), assigning specific roles to LLMs to gather varied responses(Suzgun and Kalai, 2024), and inducing more refined tasks through LLM cooperation or competition(Lazaridou et al., 2016).

Our study intersects the realms of prompt engineering, supplementary information provision, and exploration of LLM connection structures. By emulating a real-world courtroom system with LLMs, our research adopts an advanced approach to exploring connection structures and naturally incorporates prompt engineering by deriving materials for the final decision-making LLM from the arguments of prosecutors and attorneys. To our knowledge, this is the first attempt to implement a courtroom system through LLMs.

From an application perspective, this study is specifically focused on NLP classification tasks. Comparable approaches employed LLMs or unsupervised learning methods for classification tasks(Sun et al., 2023; Arora et al., 2022).

3 The Courtroom-LLM Framework

In modern society, the courtroom system crucially resolves ambiguous cases, with legal experts like prosecutors and defense attorneys advocating their positions. Ultimately, a judge assesses both sides' arguments to render a final verdict. This research utilizes LLMs to simulate this legal process, applying it to various natural language processing challenges, especially in classification tasks, as depicted in Figure 2.

The methodology employed in this research is structured into two main phases with an additional subdivision in the second phase: 1) The Preliminary Hearing phase, which entails an initial assessment of the presented case or input to determine its eligibility for further examination; and 2) The Main Trial phase, which is further divided into two key stages: a) The Argumentation Stage, where LLMs simulate the roles of prosecutors and defense attorneys to lay out their arguments and evidence, and b) The Deliberation and Verdict Stage, where the judge-LLM(s) meticulously analyze the presented arguments, weigh the evidence within the context of NLP tasks, and ultimately determine the most

Prompt for Preliminary Hearing

Choose the most appropriate label from the provided label list for the sentence.
Sentence:
sentence1: Security forces were on high alert after an election campaign in which more than 1,000 people, including seven election candidates, have been killed.
sentence2: Security forces were on high alert after a campaign marred by violence.
Label: entailment, non-entailment

Model Output (A_1)

entailment

Figure 3: Example of the preparatory prompt used in PH-LLM's initial decision-making process.

Prompt for Argument Stage (Prosecutor-LLM)

Please explain the context within these two sentence about the label *Entailment* ← This part will be *not-entailment* in case of an attorney-LLM.
The result should be one sentence.
Sentence:
sentence1: Security forces were on high alert after an election campaign in which more than 1,000 people, including seven election candidates, have been killed.
sentence2: Security forces were on high alert after a campaign marred by violence.

Model Output (Argument of prosecutor-LLM)

The context within these two sentences is that security forces were on high alert due to a campaign, either an election campaign or a campaign marred by violence.

Figure 4: Example prompts for generating arguments by the prosecutor- or attorney-LLM. The highlighted label part requesting explanation to the prosecutor and attorney are different respectively.

appropriate response or classification based on the evaluated data.

3.1 Preliminary Hearing Phase

In conventional legal contexts, a preliminary hearing determines if a case warrants a full trial by evaluating its gravity. However, in this study, this phase functions as a *preparatory* step for the Main Trial phase, adapting the process for NLP tasks.

Unlike real-world legal scenarios with clear plaintiffs and defendants, NLP cases lack such distinctions. It necessitates pre-determining the stances for prosecutor- and defense attorney-LLMs. To this end, the Preliminary Hearing LLM (PH-LLM) receives a case query and initially generates responses through a zero-shot approach, distinguishing between a first-best answer (A_1) and a second-best answer (A_2).

In *binary* classification, the responses supported by PH-LLM is designated as A_1, and the other option is labeled as A_2. For *multi-class* classification, the first and second highest-ranking responses to the query are assigned as A_1 and A_2, respectively.

Subsequently, A_1 is allocated to the prosecution, and A_2 to the defense. Based on the real-world notion that prosecutors usually support what is broadly accepted by law or common sense, we applied a similar rationale in our approach. We considered the PH-LLM's best answer, A_1, as the most universally accepted solution and thus assigned it to the Prosecutor-LLM's role in our model. The preparatory prompt for PH-LLM's initial decision-making is presented in Figure 3.

3.2 Main Trial – Argument Stage

In this stage, the prosecutor-LLM and defense attorney-LLM generate logical and evidential support for their respective positions on the given problem, backing answers A_1, and A_2.

Typically, these arguments, especially those from defense attorneys, are produced at considerable length, varying with the nature of the issue at hand. Differences in argument length between prosecution and defense can greatly affect the judge-LLM's capacity for making well-informed decisions. Given the known tendencies of current LLMs to exhibit biases towards longer textual inputs, there's an increased likelihood of a bias towards the defense's position(Mao et al., 2023). To prevent potential biases, this study imposes a length restriction on the arguments generated by both prosecutor-LLM and defense attorney-LLM, allowing only a specified length for their generated content. Figure 4 illustrates an example of a prompt designed for this purpose, guiding the creation of arguments within these limits.

3.3 Main Trial-Deliberation and Verdict Stage

At the Deliberation and Verdict Stage, judge-LLMs assess all arguments generated by the prosecutor-LLM and defense attorney-LLM to formulate a final decision. During this process, various precedential examples may be utilized, and a system of multiple judges can be implemented to enhance the deliberation depth.

Prompt for Deliberation and Verdict Stage

There are two opposite arguments about the two sentences. Referring to precedents, which one do you think is correct about this?

Sentence:
sentence1: Security forces were on high alert after an election campaign in which more than 1,000 people, including seven election candidates, have been killed.
sentence2: Security forces were on high alert after a campaign marred by violence.
Label: entailment, non-entailment

|attorney|
Both sentences describe a situation where security forces were on high alert due to a violent campaign, with the first sentence specifically mentioning the deaths of over 1,000 people and seven election candidates.
|prosecutor|
The label non-entailment refers to the fact that the second sentence does not necessarily imply the same level of violence as mentioned in the first sentence.

|precedent|
Case: Entailment
text: sentence1: As a result, peptic ulcer disease has been transformed from a chronic, frequently disabling condition to one that can be cured by a short regimen of antibiotics and other medicines.
sentence2: Antibiotics are used against peptic ulcer.
reason: The label of "entailment" is appropriate for this pair of sentences because sentence 1 implies or suggests that peptic ulcer disease can now be cured by a short regimen of antibiotics and other medicines. Sentence 2 directly states that antibiotics are used against peptic ulcer.
Case: Non-entailment
Sentence:
…

Model Output (Decision of a judge-LLM)

Based on the precedents, the correct label for the sentence pair is *entailment*.

Figure 5: Example prompts for Judge-LLM in decision making using parallel and sequential judges deliberation.

Figure 5 illustrates an example of a prompt designed for judge-LLMs, providing collected precedents, and arguments of the attorney and prosecutor for a better judgment.

3.3.1 Case-Few-shot Prompting

In actual courtrooms, judges rely on statutes to guide their decisions, but such clear-cut rules are not available in our NLP context. To bridge this gap, we implemented the "Similar Cases Retriever" module. This tool selects data from the dataset that closely matches the input query, along with the corresponding labels, grouping them as similar cases. These cases are then presented to the judge-LLMs during the Main Trial phase, serving a role akin to legal precedents. This approach, resembling few-shot prompting, aids the judge-LLMs in their deliberation process. The process for assembling these similar cases is detailed in Algorithm 3.3.2, with illustrative examples provided in Table 1.

3.3.2 Multiple Judge-LLMs

This research framework is designed to be flexible, allowing for the inclusion of a single judge-LLM or the expansion to up to five judge-LLMs. For scenarios involving multiple judges, the framework employs two distinct decision-making approaches, named as "Parallel Judges Deliberation" and "Sequential Judges Deliberation".

Parallel Judges Deliberation: This method has all judge-LLMs simultaneously review the case information, with a final decision achieved through majority voting, reflecting a collaborative approach among the judges.

Sequential Judges Deliberation: Under this approach, judge-LLMs evaluate the case in a stepwise manner, with each judge considering the evaluations of their predecessors before adding their own. The last judge-LLM's decision incorporates the insights from all previous judges, culminating in the final verdict.

The prompt for the judge-LLMs' decision-making is illustrated in Figure 5.

Algorithm 1 Similar case retrieval process for few-shot examples.

procedure SELECT_PRECEDENTS(input)
  $E_{data} \leftarrow$ get_embeddings(Data)
  $E_{Input} \leftarrow$ get_embedding(Input)
  $Distances \leftarrow$ initialize empty list
  for E_x in E_{data} **do**
    $D \leftarrow$ cosine_similarity(E_x, E_{Input})
    $Distances$.append(D)
  end for
  $BestIndex \leftarrow$ argmax$(Distances)$
  $BestPrecedent \leftarrow$ Data$[BestIndex]$
  return $BestPrecedent$
end procedure

input text
question: What came into force after the new constitution was herald? sentence: As of that day, the new constitution heralding the Second Republic came into force.
randomly selected example
question: Who originally hosted Who Wants to Be a Millionaire for ABC? sentence: Hosted throughout its ABC tenure by Regis Philbin, the program became a major ratings success throughout its initial summer run, which led ABC to renew Millionaire as a regular series, returning on January 18, 2000.
selected example using similar cases retriever
question: When was the new constitution promulgated? sentence: As of that day, the new constitution heralding the Second Republic came into force.

Table 1: Selected few-shot case examples of QNLI dataset using random selection and similar cases retriever.

Subset	Data name	label	Original Size (Sampled rate)
Natural Language Understanding	RTE(Wang et al., 2019)	Entailment, Non-entailment	277 (100%)
	BoolQ(Clark et al., 2019)	yes, no	2,370 (21.09%)
Natural Language Inference	QNLI(Wang et al., 2019)	Entailment, Non-entailment	5,460 (9.15%)
	ANLI(Nie et al., 2020) R1	entailment, neutral, contradiction	1,000 (50.00%)
	ANLI R2	entailment, neutral, contradiction	1,000 (50.00%)
	ANLI R3	entailment, neutral, contradiction	1,200 (41.66%)
Classification	Emotion(Saravia et al., 2018)	sadness, joy, love, anger, fear, surprise	2,000 (25.00%)

Table 2: Dataset summary. For evaluation, we randomly selected 500 samples from each dataset, excluding RTE. For RTE, BoolQ, and QNLI, we utilized the validation sets, while for ANLI and emotion, we used the test sets.

4 Experiments

In this section, we delve into the series of experiments conducted to rigorously evaluate the impact of the Courtroom-LLM structure on enhancing performance across a spectrum of NLP classification tasks. The primary objective of these experiments is to quantify the performance gains attributed to the incorporation of the Courtroom-LLM framework and to identify specific conditions under which this innovative approach particularly excels.

To ensure a comprehensive analysis, we selected a diverse array of classification datasets widely recognized within the NLP community for their relevance and challenge. The datasets for our experiments, including their characteristics and categories, are summarized in Table 2.

To construct a similar case retriever, we utilized the embeddings of the 'en_core_web_sm' model from spacy(Honnibal et al., 2020). Additionally, for constructing PH-LLM, Prosecutor-LLM, Attorney-LLM, and Judge-LLM, we employed GPT-3.5-turbo(Ye et al., 2023) model from OpenAI. The temperature for the model was fixed at 0.5.

4.1 Performance Gains with Courtroom Structure

Table 3 presents how the Courtroom-LLM approach consistently excels in classification tasks across different datasets.

Our experiments show the sequential judge structure excels in most scenarios, consistently outperforming other methods. Limited by space, we presented only configurations with five multi-LLM judges, but other setups also demonstrate enhanced performance; see Appendix A for full results. Additionally, the impact of few-shot example quantity on performance was not distinctly evident.

Comparatively, the sequential judge setup significantly surpasses the single-LLM-based classification across the board. Performance improvements

Task	LLM	Structure	# of decision maker	few-shot methods			
				0	1	2	3
Natural Language Understanding							
RTE	single		1	**0.6643**	0.5523	0.6282	0.5921
	multi	Parallel N-LLM	5	**0.6823**	0.5704	0.5957	0.6065
		Courtroom (Parallel Judge)	5	0.3501	0.7000	0.6600	**0.7100**
		Courtroom (Sequential Judge)	5	0.4259	0.7870	**0.7978**	**0.7978**
BoolQ	single		1	0.3720	**0.6060**	0.5580	0.5900
	multi	Parallel N-LLM	5	0.3740	**0.6140**	0.5580	0.5960
		Courtroom (Parallel Judge)	5	0.4980	0.4880	0.5020	**0.5640**
		Courtroom (Sequential Judge)	5	0.5540	0.5740	**0.5980**	0.5940
Natual Language Inference							
QNLI	single		1	0.6380	0.7540	0.6860	**0.7960**
	multi	Parallel N-LLM	5	**0.6740**	0.6280	0.6420	0.6200
		Courtroom (Parallel Judge)	5	0.4540	**0.7220**	0.6480	0.5920
		Courtroom (Sequential Judge)	5	0.5980	**0.8940**	0.8660	0.8500
ANLI R1	single		1	**0.5140**	0.4860	0.3920	0.3620
	multi	Parallel N-LLM	5	**0.5100**	0.4560	0.3760	0.3560
		Courtroom (Parallel Judge)	5	0.3900	0.4120	0.4000	**0.4340**
		Courtroom (Sequential Judge)	5	0.4420	0.5800	**0.5880**	0.5760
ANLI R2	single		1	**0.4480**	0.4380	0.3940	0.3580
	multi	Parallel N-LLM	5	**0.4720**	0.4320	0.3860	0.3640
		Courtroom (Parallel Judge)	5	0.3800	**0.4380**	0.4160	0.4400
		Courtroom (Sequential Judge)	5	0.4300	**0.5380**	0.5300	**0.5380**
ANLI R3	single		1	**0.3860**	**0.3860**	0.3460	0.3720
	multi	Parallel N-LLM	5	0.3800	0.3740	0.3480	**0.3940**
		Courtroom (Parallel Judge)	5	0.3360	**0.4760**	0.4420	0.4480
		Courtroom (Sequential Judge)	5	0.3460	0.5380	**0.5500**	0.5340
Text Classification							
Emotion	single		1	**0.4900**	**0.4900**	0.3300	0.2920
	multi	Parallel N-LLM	5	0.5020	**0.6180**	0.3540	0.2980
		Courtroom (Parallel Judge)	5	0.5540	**0.6820**	0.6800	0.6380
		Courtroom (Sequential Judge)	5	0.5120	**0.6620**	0.6480	0.6440

Table 3: Performance(Accuracy) of the Courtroom-LLM framework on various NLP classification tasks. **Bold** indicates the highest accuracy within each structure category and the same number of decision makers. A cyan background highlights the overall highest accuracy across the table. For the Courtroom-LLM structure, '# of decision makers' refers to the number of judges. Parallel N-LLMs use N independent LLMs for classification, finalized by majority voting.

ranged from a 35% increase in the RTE domain to a 34% enhancement in the QNLI task compared to baseline models.

4.2 Performance Analysis on Ambiguous Classification Cases

This research draws on the parallel between resolving legal disputes in courtrooms and addressing ambiguous classification tasks in NLP. We theorized that the Courtroom-LLM framework would effectively manage these complex classification challenges.

To validate this, we devised a simple method to distinguish between relatively difficult and straightforward classification tasks within our datasets: To differentiate between complex and simple classification tasks, we applied the following method:

1. For each input, classify using a single-LLM K times to obtain K predicted labels: $L = \{l_1, l_2, \ldots, l_K\}$.
2. Calculate the difference rate(D_{rate}) between the frequencies of the top two most common

	RTE		BoolQ		QNLI		ANLI R1		ANLI R2		ANLI R3		Emotion	
	Amb	**Nor**	**Amb**	**Nor**	**Amb**	**Nor**	**Amb**	**Nor**	**Amb**	**Nor**	**Amb**	**Nor**	**Amb**	**Nor**
Ratio	0.18	0.82	0.60	0.40	0.24	0.76	0.41	0.59	0.38	0.62	0.372	0.628	0.49	0.51
Single-LLM	0.41	0.60	0.51	0.64	0.50	0.87	0.35	0.56	0.28	0.54	0.37	0.38	0.40	0.82
Parallel N-LLM	0.41	0.60	0.54	0.64	0.50	0.87	0.32	0.56	0.26	0.54	0.36	0.38	0.45	0.82
Courtroom (Parallel Judges)	0.76 (+0.35)	0.82 (+0.22)	0.46 (-0.08)	0.46 (-0.18)	0.84 (+0.34)	0.69 (-0.19)	0.48 (+0.14)	0.36 (-0.20)	0.52 (+0.25)	0.39 (-0.15)	0.25 (-0.11)	0.25 (-0.13)	0.59 (+0.19)	0.76 (-0.06)
Courtroom (Sequential Judges)	0.71 (+0.31)	0.80 (+0.22)	0.40 (-0.11)	0.42 (-0.22)	0.79 (+0.29)	0.93 (+0.05)	0.40 (+0.05)	0.72 (+0.16)	0.42 (+0.14)	0.62 (+0.08)	0.29 (-0.09)	0.31 (-0.07)	0.56 (+0.17)	0.75 (-0.07)

Table 4: The proportion of ambiguous data(**Amb.**) and normal data(**Nor.**) separated by setting θ to 0.5, and the performance(accuracy) when using single-LLM, Parallel N-LLM and Courtroom-LLM (parallel and sequential judge methods) on the overall dataset. Blue indicates improved performance compared to single-LLM, while red indicates decreased performance compared to single-LLM. The performance of all data analyzed was evaluated using a few-shot size of 1, with 5 decision makers.

labels: $D_{rate} = (\text{freq}(l_{\text{top1}}) - \text{freq}(l_{\text{top2}}))$ / K.

3. Label cases as *'Ambiguous'* if $D_{rate} \leq \theta$; otherwise, they are *'Normal'* cases.

In this study, we set θ to 0.5 and K to 5.

Table 4 presents the proportion of each type within the datasets and the contribution of the Courtroom-LLM to performance for each data type. In most tasks, the table clearly shows a significant performance improvement in ambiguous cases when the Courtroom-LLM structure is applied, compared to the baseline. Even in normal cases, a slight enhancement in performance is noticeable.

This suggests that the Courtroom-LLM's argument-enhanced prompts excel in complex, ambiguous situations, indicating a synergistic effect from the structured debate within the LLMs. Moreover, the results reaffirm that the sequential judge method surpasses the parallel judge approach in these scenarios, highlighting its efficacy in detailed decision-making.

4.3 Performance Analysis Across Multiple Judge Utilization

In this analysis, we meticulously examine how the Parallel Judge Deliberation and Sequential Judge Deliberation systems influence decision-making trends and identify key factors that have contributed to performance enhancements observed with the Courtroom-LLM framework.

4.3.1 Judge Agreement

Figure 6 displays the agreement rates among judges in the parallel deliberation method, broken down by data type. It is evident that in almost all cases, the agreement rate surpasses 70%. Notably, the agreement rate tends to decrease for ambiguous data types, while it increases for normal cases. This pattern mirrors the intuitive real-world scenario where more challenging decisions typically result in lower agreement rates, underscoring the framework's ability to reflect the complexities of decision-making in nuanced situations.

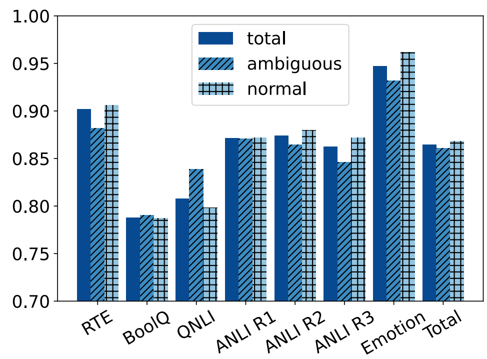
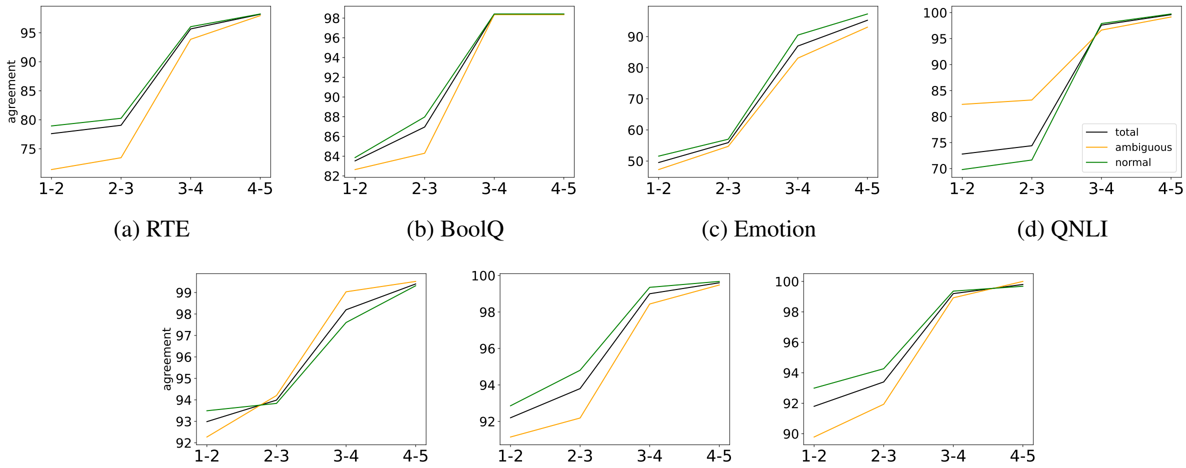


Figure 6: The average agreement rate for each dataset in the final response. It shows the average agreement rate for the overall dataset, as well as the average agreement rate for ambiguous data and normal data separately. In general, ambiguous data tends to exhibit lower agreement rates compared to normal data.

Figure 7 presents the agreement rates between consecutive judges in the sequential deliberation method, segmented by the type of data. It shows that, in almost all instances, agreement rates start lower but increase towards 100% as the deliberation sequence advances. A significant observation is that for ambiguous data, the initial agreement rates are usually lower than for normal data, indicating a heightened challenge in reaching consensus on more complex cases.

4.3.2 Comparison Initial and Final LLM Decision

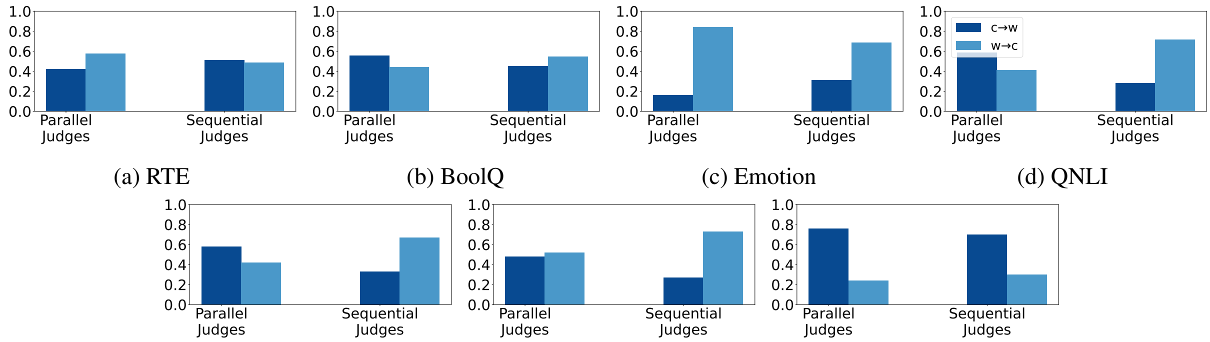
Figure 8 illustrates the performance changes in each deliberation method when the final decision diverges from the initial PH-LLM's judgment. In almost every instance, it is observed that incorrect de-



(a) RTE (b) BoolQ (c) Emotion (d) QNLI

(e) ANLI R1 (f) ANLI R2 (g) ANLI R3

Figure 7: Changes in agreement in the sequential judge deliberation method. The X-axis label '3-4' indicates the measurement of agreement between the third and fourth judges. The performance of all data analyzed was evaluated using a few-shot size of 1, with 5 decision makers.



(a) RTE (b) BoolQ (c) Emotion (d) QNLI

(e) ANLI R1 (f) ANLI R2 (g) ANLI R3

Figure 8: Rate of decision changes in Courtroom-LLM: comparing initial PH-LLM decisions with final outcomes. "w→c" represents transitions from incorrect to correct decisions, while "c→w" indicates shifts from correct to incorrect decisions in the final verdict. The performance of all data analyzed was evaluated using a few-shot size of 1, with 5 decision makers.

cisions (W: Wrong) initially made by PH-LLM are transformed into correct ones (C: Correct) through the deliberation process. This highlights the substantial impact that the arguments and discussions among prosecutor-LLM, defense attorney-LLM, and judge-LLM have on achieving accurate classification outcomes, demonstrating the effectiveness of the collaborative decision-making framework. For each task experimented in this paper, examples showcasing performance improvement over single-LLM through the actual courtroom-LLM framework can be found in Appendix B.

5 Conclusion

Our research introduces the innovative Courtroom-LLM framework, a groundbreaking approach that draws inspiration from legal courtroom processes to address the inherent ambiguities in NLP classification tasks. By simulating a courtroom setting within large language models and assigning roles analogous to judges, prosecutors, and defense attorneys, we have developed a structured multi-LLM setup that significantly enhances decision-making accuracy, especially in complex and ambiguous scenarios.

Our evaluations confirm that Courtroom-LLM outperforms traditional LLM classifiers, offering a strong solution for intricate cases without labeled data. This success suggests that judicial-inspired models can significantly improve NLP decision-making. As a future direction, we plan to adapt our framework for broader NLP applications, including sequential labeling and generative tasks, to show its versatility and impact on advancing NLP technology.

Limitations

The Courtroom-LLM framework, despite its effectiveness in NLP classification, presents certain limitations:

1. **Scope of Application**: The current setup is designed for text-classification, derived from debates between prosecutor-LLM and defense attorney-LLM. Expanding this framework to accommodate generative NLP tasks and sequential labeling scenarios remains a challenge for future development.

2. **Resource and Time Intensity**: High classification accuracy comes at the cost of multiple LLM queries, requiring significant computational resources and time. Exploring the use of smaller LLMs within the framework could potentially address this issue.

3. **Handling of Neutral Labels**: The framework shows limitations in accurately classifying '*neutral*' labels in tasks like natural language inference, indicating a need for improved model sensitivity to nuanced classifications.

Future enhancements to the Courtroom-LLM framework should aim to address these limitations, broadening its applicability and efficiency in diverse NLP tasks.

References

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT 2019*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need.

Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4):647–665.

Junyu Mao, Stuart E. Middleton, and Mahesan Niranjan. 2023. Do prompt positions really matter?

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Qiang Fu, Yan Gao, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. *arXiv preprint arXiv:2201.11473*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

A Overall Accuracy

We experimented with the performance of few-shot 1 to 3 examples and judge-LLM configurations of 1, 3, and 5, using the data employed in the paper to validate the methodology. We conducted experiments for single-LLM, multiple-LLM without the Courtroom-LLM framework, and two versions of applying our framework (parallel and sequential judges). Table 5, Table 7, and Table 6 displays the performance of datasets for natural language understanding, natural language inference, and text classification task.

B Example of Judge-LLM

In this section, we present the formatted context input and corresponding outputs for the actual judge-LLM. We provide the input forms for the RTE dataset in natural language understanding task, the ANLI R1 dataset in natural language inference task, and the Emotion dataset in text classification task, along with the outputs of single-LLM, parallel N-LLM, and Courtroom-LLM(parallel judges), and Courtoom-LLM(sequential judges). The inputs for RTE, ANLI R1, Emotion datasets are shown in Table 8, Table 10, and Table 12. The outputs are shown in Table 9, Table 11, and Table 13. While there have been no alterations to the actual input data, redundant information overlapping with the actual datasets has been condensed in the respective tables.

Task	LLM	Structure	# of decision maker	few-shot methods			
				0	1	2	3
RTE	single		1	**0.6643**	0.5523	0.6282	0.5921
	multi	Parallel N-LLMs	3	0.6678	0.5342	0.6209	0.6064
			5	**0.6823**	0.5704	0.5957	0.6065
		Courtroom (Parallel Judges)	1	0.3610	0.5704	0.6029	0.5957
			3	0.3610	0.7292	**0.7690**	0.7076
			5	0.3501	0.7000	0.6600	0.7100
		Courtroom (Sequential Judges)	1	0.4223	0.5704	0.6029	0.5957
			3	0.3610	0.7906	0.7726	0.7906
			5	0.4259	0.7870	**0.7978**	**0.7978**
BoolQ	single		1	0.3720	**0.6060**	0.5580	0.5900
	multi	Parallel N-LLMs	3	0.3920	**0.6200**	0.5580	0.6060
			5	0.3740	0.6140	0.5640	0.5960
		Courtroom (Parallel Judges)	1	0.3460	0.4320	0.4300	0.4720
			3	0.3420	0.5020	0.5000	0.5180
			5	0.4980	0.4880	0.5020	**0.5640**
		Courtroom (Sequential Judges)	1	0.1840	0.3440	0.4240	0.4340
			3	0.1760	0.3300	0.4200	0.4180
			5	0.5540	0.5740	**0.5980**	0.5940

Table 5: Natural language understanding task accuracy comparison on RTE(Wang et al., 2019) and BoolQ(Clark et al., 2019) dataset: **Bold** indicates the highest accuracy within each structure category. A cyan background highlights the overall highest accuracy across the table. For the Courtroom-LLM structure, '# of decision makers' refers to the number of judges. Parallel N-LLMs use N independent LLMs for classification, finalized by majority voting.

Task	LLM	Structure	# of decision maker	few-shot methods			
				0	1	2	3
Emotion	single		1	**0.4900**	**0.4900**	0.3300	0.2920
	multi	Parallel N-LLMs	3	0.4960	0.5900	0.3340	0.3140
			5	0.5020	**0.6180**	0.3540	0.2980
		Courtroom (Parallel Judges)	1	0.4520	0.6700	0.6760	0.6680
			3	0.4500	0.6560	0.6500	**0.6860**
			5	0.5540	0.6820	0.6800	0.6380
		Courtroom (Sequential Judges)	1	0.4440	0.6700	**0.6760**	0.6680
			3	0.4480	0.6740	0.6440	0.6500
			5	0.5120	0.6620	0.6480	0.6440

Table 6: Classification task accuracy on Emotion(Saravia et al., 2018) datasets: **Bold** indicates the highest accuracy within each structure category. A cyan background highlights the overall highest accuracy across the table. For the Courtroom-LLM structure, '# of decision makers' refers to the number of judges. Parallel N-LLMs use N independent LLMs for classification, finalized by majority voting.

Task	LLM	Structure	# of decision maker	few-shot methods			
				0	**1**	**2**	**3**
QNLI	single		1	0.6780	0.7540	0.6860	**0.7960**
	multi	Parallel N-LLMs	3	**0.7080**	0.7940	0.7080	0.7920
			5	0.7020	0.6280	0.6420	0.6200
		Courtroom (Parallel Judges)	1	0.4580	**0.8800**	0.8600	0.8740
			3	0.4860	0.5740	0.4880	0.4880
			5	0.4540	0.7220	0.6480	0.5920
		Courtroom (Sequential Judges)	1	0.6160	0.8800	0.8600	0.8740
			3	0.6030	**0.8940**	0.8520	0.8480
			5	0.5980	**0.8940**	0.8660	0.8500
ANLI R1	single		1	**0.5140**	0.4860	0.3920	0.3620
	multi	Parallel N-LLMs	3	**0.5220**	0.4560	0.3780	0.3620
			5	0.5100	0.4560	0.3760	0.3560
		Courtroom (Parallel Judges)	1	0.3640	0.5600	0.5620	**0.5800**
			3	0.3740	0.4180	0.4220	0.4260
			5	0.3900	0.4120	0.4000	0.4340
		Courtroom (Sequential Judges)	1	0.4620	0.5600	0.5620	0.5800
			3	0.4440	0.5760	0.5660	0.5600
			5	0.4420	0.5800	**0.5880**	0.5760
ANLI R2	single		1	**0.4480**	0.4380	0.3940	0.3580
	multi	Parallel N-LLMs	3	0.4540	0.4340	0.3860	0.3640
			5	**0.4720**	0.4320	0.3860	0.3640
		Courtroom (Parallel Judges)	1	0.3580	**0.4380**	0.4320	0.4360
			3	0.3620	0.4360	0.3960	0.4100
			5	0.3800	**0.4380**	0.4160	0.4400
		Courtroom (Sequential Judges)	1	0.4320	0.5420	**0.5500**	0.5260
			3	0.4280	0.5380	0.5320	0.5300
			5	0.4300	0.5380	0.5300	0.5380
ANLI R3	single		1	**0.3860**	**0.3860**	0.3460	0.3720
	multi	Parallel N-LLMs	3	0.3980	0.3700	0.3500	0.3660
			5	0.3800	0.3740	0.3480	**0.3940**
		Courtroom (Parallel Judges)	1	0.3740	**0.5600**	0.5380	0.5420
			3	0.3900	0.4460	0.4340	0.4780
			5	0.3360	0.4760	0.4420	0.4480
		Courtroom (Sequential Judges)	1	0.3760	0.5600	0.5380	0.5420
			3	0.3940	0.5520	**0.5640**	0.5260
			5	0.3460	0.5380	0.5500	0.5340

Table 7: Natural language inference task accuracy comparison on QNLI(Wang et al., 2019) and ANLI(Nie et al., 2020): **Bold** indicates the highest accuracy within each structure category. A cyan background highlights the overall highest accuracy across the table. For the Courtroom-LLM structure, '# of decision makers' refers to the number of judges. Parallel N-LLMs use N independent LLMs for classification, finalized by majority voting.

Context
Sentence: sentence1: Eric Harris and Dylan Klebold, seniors at the suburban Denver school, ... sentence2: 13 persons were killed by two students in 1999. Label: entailment,non-entailment
Arguments
|prosecutor| $\leftarrow A_1$ The label entailment is that the event described in sentence 2 is the same as the massacre described in sentence 1 where Eric Harris and Dylan Klebold killed a teacher and 12 students, representing the violent destruction of the perception of schools as safe havens. |attorney| $\leftarrow A_2$ The label ""non-entailment"" refers to the fact that sentence 2 does not fully capture the magnitude and impact of the event described in sentence 1, which involved the killing of a teacher, the injuring of numerous individuals, and the shattering of the perception of schools as safe places.
Precedents
Case: entailment text: sentence1: Rotorua has banned criminals with five or more dishonesty convictions ... sentence2: The Central Business District (CBD) is part of Rotorua. reason: The label of ’entailment’ is appropriate for this sentence pair because sentence 2 directly follows from and is implied by sentence 1. In sentence 1, it is mentioned that criminals with five or more dishonesty convictions are banned from entering the Central Business District (CBD) of Rotorua. Sentence 2 simply states that the Central Business District (CBD) is part of Rotorua, which is a logical consequence of the information provided in sentence 1. Therefore, sentence 2 can be inferred from sentence 1, indicating an entailment relationship between the two sentences. Case: non-entailment text: sentence1: The court in Angers handed down sentences ranging from four months suspended to 28 years for, ... sentence2: Franck V. comes from Angers. reason: The reason for labeling the sentence as ‘non-entailment’ is that sentence 2 does not necessarily follow or logically derive from sentence 1. While sentence 1 provides information about Franck V.’s involvement in a sex ring in Angers, sentence 2 simply states that Franck V. comes from Angers. There is no direct or implied connection between Franck V.’s origin and his involvement in the sex ring mentioned in sentence 1. Therefore, sentence 2 does not necessarily follow from sentence 1, leading to the label of ’non-entailment’.

Table 8: Natural language understanding task input example of judge-LLM on RTE dataset. The precedent size is 1, and A_1 label is ‘entailment’ and A_2 is ‘non-entailment’. All the data included in the natural language understanding task share the same input format.

Answer Label
entailment
Single-LLM Output
non-entailment
Parallel *N*-LLM Output
non-entailment
Parallel *N*-LLM Output list
[non-entailment , non-entailment , non-entailment , non-entailment , non-entailment]
D_{rate}
1.0
Courtroom(Parallel Judges)
[non-entailment , non-entailment , non-entailment, non-entailment, non-entailment]
Courtroom(Sequential Judges) Output
entailment
Courtroom(Sequential Judges) Output list
[non-entailment, non-entailment, entailment, non-entailment, entailment]
Courtroom(Sequential Judges) Responses
|Judge 0| The correct label for this sentence pair is "non-entailment." |Judge 1| The correct label for this sentence pair is "non-entailment." |Judge 2| |Judge 2| The correct label for this sentence pair is "entailment." |Judge 3| The correct label for this sentence pair is "non-entailment." |Judge 4| The correct label for this sentence pair is "entailment."

Table 9: Natural language understanding task output example of judge-LLM input on RTE dataset. The precedent size is 1, and A_1 label is 'entailment' and A_2 is 'non-entailment'.

Context
premise: Helena Sukov (]) (born 23 February 1965) is a former professional tennis player from the Czech Republic. hypothesis: Helena Sukova enjoyed attending professional football games. Label: entailment,neutral,contradiction
Arguments
|prosecutor| $\leftarrow A_1$ The hypothesis that Helena Sukova enjoyed attending professional football games cannot be determined from the given premise. |attorney| $\leftarrow A_2$ The hypothesis that Helena Sukova enjoyed attending professional football games cannot be supported or refuted based on the given information about her career as a professional tennis player.
Precedents
Case: entailment text: premise: Judy Tegart Dalton is a retired professional tennis player... hypothesis: Judy and player Margaret Court played on the same doubles team in tournaments. reason: The reason for labeling the sentence as 'entailment' is that the hypothesis is supported by the information provided in the premise. The premise states that Judy Tegart Dalton won nine Grand Slam doubles titles, and five of those titles were with Margaret Court. Therefore, it can be inferred that Judy and Margaret Court played on the same doubles team in tournaments. Case: neutral text: premise: Alexandra Lendon Bastedo was a British actress, best known for her role as secret agent Sharron Macready ... hypothesis: Alexandra Lendon Bastedo despised her role as Sharron Macready. reason: The label of the sentence is 'neutral' because it simply states a fact about Alexandra Lendon Bastedo, without expressing any positive or negative sentiment. The sentence mentions her role as Sharron Macready in the 1968 British espionage/science fiction adventure series "The Champions" and states that she is best known for it. It does not provide any opinion or judgment about her feelings towards the role. Case: contradiction text: premise: This is a list of cities in Azerbaijan. Azerbaijan is a country in the South Caucasus region ... hypothesis: Azerbaijan has over 100 cities and 4,620 villages. reason: The label of the sentence is 'contradiction' because the hypothesis states that Azerbaijan has over 100 cities and 4,620 villages, which contradicts the information presented in the premise. The premise clearly states that Azerbaijan has a total of 77 cities, including 12 Federal-level cities, and provides the specific numbers for smaller "rayon"-class cities, urban-type settlements, and villages. Therefore, the hypothesis contradicts the information given in the premise.

Table 10: Natural language inference task input example of judge-LLM on ANLI R1 dataset. The precedent size is 1, and A_1 label is 'neutral' and A_2 is 'contradiction'. All the data included in the natural language inference task share the same input format.

Answer Label
Neutral
Single-LLM Output
Contradiction
Parallel *N*-LLM Output
Contradiction
Parallel *N*-LLM Output list
[Contradiction, Contradiction, Contradiction, Contradiction, Contradiction]
D_{rate}
1.0
Courtroom(Parallel Judges)
[Neutral, Contradiction, Neutral, Neutral, Contradiction]
Courtroom(Sequential Judges) Output
Neutral
Courtroom(Sequential Judges) Output list
[Contradiction, Neutral, Neutral, Neutral, Neutral]
Courtroom(Sequential Judges) Responses
|Judge 0| Based on the precedents, the correct label for the sentence "The hypothesis that Helena Sukova enjoyed attending professional football games cannot be determined from the given premise" is contradiction. |Judge 1| Based on the precedents, the correct label for the sentence "The hypothesis that Helena Sukova enjoyed attending professional football games cannot be determined from the given premise" is neutral. |Judge 2| Based on the precedents, the correct label for the sentence is neutral. |Judge 3| Based on the precedents, the correct label for the sentence is neutral. |Judge 4| Based on the precedents, the correct label for the sentence " is neutral.

Table 11: Natural language inference task output example of judge-LLM on ANLI R1 dataset. The precedent size is 1, and A_1 label is 'neutral' and A_2 is 'contradiction'.

Context
Sentence: i jest i feel grumpy tired and pre menstrual which i probably am but then again its only been a week and im about as fit as a walrus on vacation for the summer Label: sadness,joy,love,anger,fear,surprise
Arguments
|prosecutor| $\leftarrow A_1$ The speaker is feeling grumpy, tired, and possibly premenstrual, which may contribute to their sadness, but they also acknowledge that it has only been a week and they are not in the best physical shape. |attorney| $\leftarrow A_2$ The speaker is expressing their current state of feeling grumpy, tired, and premenstrual, possibly attributing it to their hormonal cycle, but also acknowledging that it has only been a week since their last period and they are not physically fit.
Precedents
Case: sadness text: im pretty sure and its been about a week and a half so although im feeling kind of betrayed and disillusioned by men at the moment everythings okay reason: The label of "sadness" is assigned to this sentence because the person expresses feeling betrayed and disillusioned by men. These negative emotions indicate a sense of sadness or disappointment. Despite stating that everything is okay, the initial emotions conveyed suggest a underlying feeling of sadness. Case: anger text: im not condoning terrorist action but you feel so furious and powerless reason: The label of 'anger' is appropriate for this sentence because the speaker expresses a strong emotion of fury. The use of the word "furious" indicates a high level of anger and frustration. Additionally, the feeling of powerlessness further emphasizes the intensity of the speaker's anger.

Table 12: Text classification task input example of judge-LLM on Emotion dataset. The precedent size is 1, and A_1 label is 'sadness' and A_2 is 'anger'.

Answer Label
anger
Single-LLM Output
suprise
Parallel *N*-LLM Output
anger
Parallel *N*-LLM Output list
[suprise, anger, anger, anger, anger]
D_{rate}
0.04
Courtroom(Parallel Judges)
[sadness, anger, sadness, sadness, sadness]
Courtroom(Sequential Judges) Output
anger
Courtroom(Sequential Judges) Output list
[anger, sadness, anger, anger, anger]
Courtroom(Sequential Judges) Responses
|Judge 0| The correct label for the given sentence is "anger." |Judge 1| The correct label for the given sentence is "sadness." |Judge 2| The correct label for the given sentence is "anger." |Judge 3| The correct label for the given sentence is "anger." |Judge 4| The correct label for the given sentence is "anger."

Table 13: Text classification task output example of judge-LLM on Emotion dataset. The precedent size is 1, and A_1 label is 'sadness' and A_2 is 'anger'.