DR-CUP: A Dataset on Real-Time Commentary in U.S. Presidential Debates

Yu-Yu Chang[1]*, **Huan-Wen Ho**[1]*, **Chung-Chi Chen**[2]†, **Ming-Hung Wang**[1]†

[1]National Chung Cheng University, Taiwan, [2]AIST, Japan
Project Website: https://genchal.nlpfin.com/

Abstract

Presidential debates are critical platforms for political discourse, yet existing research lacks datasets tailored for analyzing real-time professional commentary. To address this, we introduce the **D**ataset on **R**eal-time **C**ommentary in **U**.S. **P**residential debates (DR-CUP), which aligns U.S. presidential debate transcripts (2016–2024) with professional commentary and annotations. DR-CUP supports research on commentary understanding, planning, and generation, offering insights into expert analysis and its role in contextualizing complex political discourse. In pilot studies, we evaluated state-of-the-art large language models (LLMs), revealing notable performance differences in understanding expert commentary and planning for generating professional commentary. DR-CUP is the first dataset to incorporate real-time cross-document alignment for debate data, providing a comprehensive resource for advancing research in political communication and computational social science.

Keywords: Corpus, Discourse Annotation, Representation and Processing, Natural Language Generation

Introduction

Presidential debates have emerged as critical platforms for political discourse, with their transcripts widely utilized in various research domains (Fein et al., 2007; Robertson et al., 2019; Haddadan et al., 2019). These debates provide rich insights for fields ranging from online crowd reaction analysis to argument mining and political psychology. Despite extensive research on presidential debate data, a significant gap persists: existing studies predominantly focus either on the debate's content (Jo et al., 2020b) or on accompanying social media discussions (Robertson et al., 2019). Consequently, a comprehensive dataset tailored for real-time professional analysis and commentary of presidential debates remains unavailable. To address this gap, we propose the **D**ataset on **R**eal-time **C**ommentary in **U**.S. **P**residential debates (DR-CUP), which provides researchers with a foundation for professional analysis of presidential debates.

Figure 1 shows an example of a debate transcript along with the real-time summary and commentary produced by professional journalists using the proposed DR-CUP. In the contemporary media landscape, live commentary (Zhang et al., 2016; Ishigaki et al., 2021; Marrese-Taylor et al., 2022) transforms passive viewership into an engaged and contextualized experience. Professional commentators provide instant analysis, enabling audiences to navigate the complexities of political discourse by unpacking nuanced arguments, identifying strategic communication techniques, and offering real-time fact-checking alongside historical and policy-related context. Live commentary thus serves several critical functions in modern political communication.

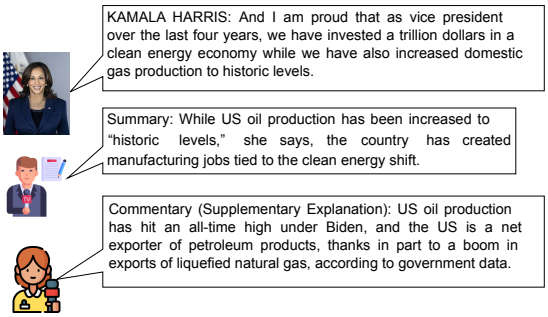


Figure 1: Example debate transcript together with the real-time summary and commentary produced by professional journalists in the proposed DR-CUP

First, it enhances accessibility by translating complex political language into insights that resonate with diverse audiences. By highlighting underlying messages and decoding rhetorical strategies, commentators make political discussions comprehensible to a broader public. Second, professional analysis deepens the understanding of debates by uncovering layers of meaning often missed by untrained observers. Third, in an era marked by information overload and media polarization, live commentary delivers balanced, analytically sound interpretations, guiding audiences through political exchanges with clarity and impartiality. The digital age has further amplified the importance of live commentary. Social media and streaming platforms provide instant global access to debates, and professional commentators play a vital role in

*Co-first author.
†The last two authors advised this project equally.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9850–9860
11–16 May 2026.

Dataset	Source	Period	Topic	# of Labels
Check-Worthy (Patwari et al., 2017)	D	2016	Fact-Checking	2
CLEF (Atanasova et al., 2018)	D	2016	Fact-Checking	2
Claim-Rank (Atanasova et al., 2019)	D	2016	Fact-Checking	2
CMU (Jo et al., 2020a)	R	2016	Proposition Type	4
M-Arg (Mestre et al., 2021)	D	2020	Argument Mining	3
DR-CUP (Proposed)	D & C	2016-2024	Commentary Aspect	11

Table 1: Dataset comparison. C, R, and D denote commentary, Reddit, and debate, respectively.

shaping public interpretation through real-time fact-checking, contextual analysis, and expert perspectives. This interpretive layer bridges the gap between political performances and meaningful public understanding. The proposed DR-CUP dataset offers a pivotal resource for advancing research on real-time expert commentary understanding, planning, and generation in political discourse.

In this study, we present the DR-CUP dataset, designed to analyze the dynamics of commentary and discussion in U.S. political debates. The dataset encompasses U.S. presidential election debates spanning from 2016 to 2024, providing aligned transcripts, professional commentary, and manual annotations with tailored label designs to facilitate an in-depth exploration of debate discourse. In our pilot explorations, we evaluated the applicability of large language models (LLMs) in understanding expert-level commentary, planning, and generating expert-like commentary. Our findings offer initial insights into the current capabilities and limitations of LLMs in supporting the understanding and generation of expert commentary. To the best of our knowledge, DR-CUP is the first dataset to include real-time cross-document alignment for presidential debate data and is among the most comprehensive datasets covering all debates for the 2024 U.S. presidential election.

Related Work

Datasets capturing various aspects of political discourse, particularly U.S. presidential debates, are crucial for advancing research in areas such as argument mining, fact-checking, and commentary analysis. Table 1 presents a comparative analysis of the proposed DR-CUP dataset and existing datasets. Prior work has utilized diverse datasets to explore these dimensions, often focusing on arguments, factual claims, and the structure of political discourse. For example, the Check-Worthy dataset (Patwari et al., 2017) and CLEF dataset (Atanasova et al., 2018), both derived from debate content, emphasize identifying claims warranting further scrutiny. These datasets feature binary labels and are tailored to fact-checking tasks. Similarly, Claim-Rank (Atanasova et al., 2019) provides annotated data for prioritizing fact-check-worthy claims, employing machine learning techniques to enhance predictive accuracy. The CMU dataset (Jo et al., 2020a), sourced from Reddit commentary on debates, broadens the scope by categorizing proposition types into four distinct labels. This dataset illustrates the integration of social media discourse with debate analysis, offering insights into public reactions and discourse dynamics. Meanwhile, M-Arg (Mestre et al., 2021), a multimodal dataset comprising textual and audio records of the 2020 U.S. presidential debates, focuses on argument mining through three categories: support, attack, and neutral. This dataset shows the increasing adoption of multimodal data to deepen understanding of argument structures.

Existing datasets typically focus on specific aspects of debates, such as factual claims or argument categories. In contrast, the DR-CUP dataset broadens the scope by incorporating both debate transcripts and commentaries from 2016 to 2024. It categorizes the commentaries into 11 labels, offering a more comprehensive view of the interaction between expert commentary and candidate statements, making it well-suited for deeper analysis. Our commentary labels were inspired by prior work in political discourse analysis and fact-checking, including the argument-based classification framework proposed by (Goffredo et al., 2023). Beyond fact-checking and argument mining, DR-CUP is also designed to support emerging tasks in grounded comment understanding and generation. For example, (Yang et al., 2019) proposed a model to generate reader comments on news articles, while (Liu et al., 2024) introduced SciNews, a dataset aimed at transforming scientific content into accessible news narratives. While DR-CUP itself is a dataset, it enables similar tasks under the unique constraint of real-time political debates. In our experiments, we demonstrate how DR-CUP can facilitate tasks such as commentary type classification and real-time comment generation, allowing future models to benchmark performance in politically salient, time-sensitive contexts.

Dataset

We introduce DR-CUP, a dataset that aligns U.S. presidential debate transcripts with real-time pro-

fessional commentary across ten debate events held between 2016 and 2024. These include presidential, vice presidential, and Republican primary debates, offering diverse political contexts for analysis. The dataset contains a total of 2,284 annotated commentary segments, each paired with corresponding transcript spans. This setup allows for fine-grained analysis of how experts interpret, summarize, and critique live political discourse. Our commentary data is sourced from Bloomberg expert analysis, while the transcripts are compiled from records provided by various media outlets.

Label Design

We propose seven main categories to annotate the dataset, and the *Commentator's Personal Opinion* (CPO) category is further expanded into five subcategories. We also provide example annotations on our GitHub repository, which include labeled commentary segments and their corresponding debate transcripts. The following is a detailed description of each label.

- **Key Summary (KS)**: This label indicates that the commentator is summarizing points raised by the debate moderators or contestants.
- **Supplementary Explanation (SE)**: This label is used when the commentator provides additional context or information sourced from experts, real-world events, or the current debate situation without expressing subjective opinions.
- **Commentator's Personal Opinion (CPO)**: This label captures instances where the commentator voices their viewpoint on a particular issue. It includes five subcategories:
  - **Performance of the Contestants (PC)**: Assesses contestants' discussion performance.
  - **Candidate Statements (CS)**: Analyzes specific claims made by the contenders.
  - **Analyzing or Conclusions (AC)**: Involves inferences or conclusions drawn by the commentator about a statement or occurrence.
  - **Market Performance (MP)**: Pertains to comments regarding the economic performance of a nation or stock market trends.
  - **Others**: Covers commentary on topics not addressed by the other sub-labels.
- **Fact-checking (FC)**: Verifies the accuracy of candidates' statements or external rumors.
- **Market Reactions (MR)**: Highlights commentary related to economic fluctuations or monetary market trends.
- **Public Opinion (PO)**: Represents descriptions of public sentiment on specific issues or polling trends.
- **Commentator's Question (CQ)**: Indicates that the commentator is posing a question about a particular issue.

Data Annotation

The dataset was compiled using debate scripts from the U.S. presidential election debates spanning 2016 to 2024 (as well as the Republican primary debates of 2023) and the corresponding professional commentary transcripts collected from Bloomberg. We extracted commentator comments from scripts and matched them with relevant debate transcript segments. Each comment was labeled with its corresponding segment; if no match was found, the segment was marked as "NO."

The annotators were non-U.S. graduate students with a background in computational linguistics. They engaged in in-depth discussions of Bloomberg commentaries and debate transcripts. The commentary label design was not created from scratch, but rather built upon both observation and prior research. We first identified common commentary functions from Bloomberg (e.g., summary, personal opinion), and refined the scheme by incorporating insights from argument-based classification and fact-checking research in political debates, especially (Goffredo et al., 2023). Two annotators performed the annotations, with agreement measured using Cohen's Kappa (Cohen, 1960) and Krippendorff's α (Krippendorff, 2011). Commentary type labels showed moderate agreement (Landis, 1977) with scores of 0.6540 and 0.6541.These results indicate a reasonable level of agreement between the two annotators, suggesting that the labeling scheme we designed is of acceptable quality. Nonetheless, discrepancies between annotators do occasionally occur. In such instances, the annotators engage in discussion to reach a consensus on the most appropriate label.

Annotators labeled each commentary based on its relationship to the debate segment. The process began with a binary distinction between objective and subjective content. Objective comments—those that restated debate content or introduced verifiable external facts—were categorized as *Key Summary*, *Fact-Checking*, *Supplementary Explanation*, *Observation on Public Opinion*, or *Observation on Market Reaction*. Subjective comments were further classified into five subtypes of *Commentator's Personal Opinion* (CPO), based on their tone or focus (e.g., evaluation of performance, analytical inference). Comments that posed questions were labeled as *Commentator's Question*. The annotators followed a shared internal guideline based on this decision procedure and conducted initial trial annotations to calibrate their understanding. Disagreements were resolved collaboratively.

Year	Event	KS	FC	SE	CQ	PO	MR	CPO					Total
								PC	CS	AC	MP	Others	
2016	First U.S. Presidential Debate	119	3	12	0	1	9	16	20	4	1	7	192
2016	Second U.S. Presidential Debate	166	24	12	5	0	4	19	10	6	0	14	260
2016	Third U.S. Presidential Debate	154	22	26	0	2	9	25	9	6	0	9	262
2020	First U.S. Presidential Debate	234	6	104	1	3	2	12	14	4	0	5	385
2020	U.S. Vice Presidential Debate	155	8	53	2	0	0	6	11	2	0	4	241
2020	U.S. Presidential Debate	221	5	70	0	1	1	8	12	1	0	3	322
2023	GOP Presidential Debate	94	3	24	0	0	0	8	5	1	0	3	138
2023	GOP Presidential Debate	50	3	3	0	0	0	9	10	0	0	2	77
2024	Biden-Trump Presidential Debate	76	13	11	0	1	9	11	8	4	2	1	136
2024	Harris-Trump Presidential Debate	128	12	68	0	4	6	8	22	16	1	6	271
Total		1,389	99	384	9	12	40	123	124	44	4	55	2,284

Table 2: This table presents the results of our statistical analysis of the labels from each debate.

Year	Event	Sen.		Char.	
		Trans.	Comm.	Trans.	Comm.
2016	1st Pres.	689	515	51,000	51,051
2016	2nd Pres.	511	261	39,176	24,646
2016	3rd Pres.	295	145	24,616	17,810
2020	1st Pres.	538	258	46,929	28,444
2020	VP Debate	711	580	47,497	37,816
2020	2nd Pres.	387	466	38,148	32,567
2023	GOP Pres.	868	801	54,395	53,965
2023	GOP Pres.	651	752	49,788	49,738
2024	Biden-Trump	617	715	49,143	44,758
2024	Harris-Trump	448	439	33,105	24,709
Total		5,715	4,932	433,797	365,504

Table 3: Total Char.(Characters) and Sen.(Sentences) counts across debates. Trans.(Transcript) counts exclude fragments labeled as 'No'. Comm. stands for Commentary.

Statistics

Table 2 summarizes the annotation statistics. *Key Summary* dominates with 60.84% of 2,283 commentaries, showing its role in distilling key points and rephrasing debates for clarity. We acknowledge the class imbalance, as over 60% of the comments are labeled as *Key Summary*. However, this reflects the label's intended function: to capture commentators' objective descriptions of debate events. Since a large proportion of expert comments focus on summarizing candidates' statements and overall debate dynamics, we found this distribution to be appropriate and thus did not modify the label design. *Supplementary Explanation* and *Commentator's Personal Opinion* follow at 16.82% and 15.33%, highlighting the challenge of automating commentary due to reliance on external knowledge. *Fact-checking*, a key focus in prior datasets, appears in this dataset, with notably higher instances in the 2016 and 2024 presidential debates than in 2020. Table 3 presents the scale of our dataset, detailing the sentence and word counts for both transcripts and commentaries.

This distinction between objective (summary-like) and subjective (opinion-based) commentary is central to our annotation framework. Notably, within the *Commentator's Personal Opinion* (CPO) category, most commentary focuses on evaluating the *Performance of the Contestants* (PC)—this subcategory is the most frequent among CPO comments. In contrast, observations on *Market Reactions* (MR) are extremely rare, making it one of the least represented labels in the entire dataset. This indicates that real-time expert commentary tends to emphasize subjective assessments of debate performance, while references to market reactions are much less common, reflecting the typical focus of professional commentators during live coverage.

Party	Authenticity	2016	2020	2024
Democratic	True	16	2	4
	False	3	2	3
Republican	True	8	3	0
	False	23	12	18

Table 4: Fine-grained labels on the *Fact-Check* category.

The fact-checking category is further analyzed by political party (Democratic/Republican) and authenticity (true/false). Table 4 shows that false information persists in 2024, even in major debates. The dataset enables research on topics like fact-checking, opinion mining, and summarization using DR-CUP subsets.

Experimental Setting

Task

We introduce three tasks based on the proposed dataset: commentary understanding, commentary planning, and commentary generation. Table 5 provides an overview. Commentary understanding categorizes a given commentary into one of 11 defined categories, requiring the LLM to perform this classification. This serves as a benchmark for model evaluation and enables automatic labeling.

With the advancement of LLMs, models can increasingly assist professionals; however, generating commentary remains challenging, as discussed in Section Statistics, with over 33% involving external knowledge. We propose the task of commentary planning, wherein models predict the type of commentary professionals would produce from a

Task	Input	Output
Understanding	Commentary	Label (11 classes)
Planning	Debate transcript	Label (11 classes)
Generation	Debate transcript	Commentary

Table 5: Overview of the three tasks defined on the DR-CUP dataset.

debate transcript, thereby facilitating real-time suggestions. In summary, commentary understanding classifies existing content, while commentary planning anticipates relevant aspects from a debate.

Furthermore, we utilize the predefined labels with LLMs to emulate human experts in generating commentaries based on debate transcripts. Specifically, for the generation task, the model is given a segment of the debate transcript as input and is required to generate a plausible expert-style commentary corresponding to that segment. Our objective is to examine the differences between LLM-generated commentaries and those authored by human experts.

Evaluation

We evaluate four high-performing LLMs: GPT-4o, Claude 3.5-Sonnet, Gemini 2.0-Flash, and DeepSeek R1-1776. Two prompting schemes, zero-shot and few-shot, are employed. In the zero-shot scheme, only the definition of each label is provided. In the few-shot scheme, two examples per label are included to enable in-context learning. We make all prompt templates and in-context examples used in our experiments publicly available in our GitHub repository. For the understanding and planning tasks, we adopt both micro-F1 and macro-F1 scores as evaluation metrics.

For the generation task, we employ BLEU, ROUGE, and BERTScore to measure the similarity between generated and professional commentaries. Additionally, we introduce an evaluation scheme in which LLMs, acting as proxies for the general public, rate the commentaries—providing a more objective complement to the subjective selection process. In the main evaluation, we focus on three high-performing API-based LLMs: GPT-4o, Claude 3.5-Sonnet, and Gemini 2.0-Flash. Although DeepSeek R1-1776 (70B) was considered, it was excluded from the main experiments due to its significantly smaller parameter size compared to the other three API models, in order to ensure consistency and comparability in our results.

We categorize all labels into two groups: **Key-Summary**, grouped under the **Summary** category, and the remaining ten labels under the **Commentary** category. Different evaluation criteria are defined for each. Our design is inspired by the concept of news value as defined in (Bednarek, 2010), adapted to assess the quality of commentary in the debate domain. This classification is grounded in the distinction between objective and subjective commentary. **Key-Summary** represents objective recounting of debate events, whereas the ten **Commentary** labels involve more subjective interpretations, evaluations, or reasoning. By incorporating this division, we can align each group with different communicative intents—factual clarity versus analytical depth—and formulate evaluation criteria accordingly. This approach not only builds on the principles of (Bednarek, 2010), but also facilitates more targeted assessment in downstream tasks such as summary quality scoring and commentary generation. Below are the definitions of the evaluation aspects.

- **Summary Scoring Metrics (1–5 scale):**
  - **Evidentiality (Evi.):** Measures the relevance between the summary and the debate transcript.
  - **Style:** Evaluates the writing style and presentation quality of the summary.
  - **Reliability (Re.):** Assesses consistency with the factual content of the transcript.
  - **Comprehensibility (Com.):** Evaluates whether the summary enables the general public to clearly understand the transcript content.
- **Commentary Scoring Metrics (1–5 scale):**
  - **Importance (Imp.):** Measures the relevance and societal impact of the commentary.
  - **Comprehensibility (Comp.):** Assesses clarity and the effective conveyance of the commentator's message.
  - **Expectedness (Exp.):** Evaluates alignment with or deviation from public expectations, i.e., the degree of surprise.
  - **Possibility (Po.):** Assesses the plausibility of the events or reasoning described.
  - **Reliability (Rel.):** Evaluates the trustworthiness of the commentary; higher scores indicate greater credibility.

Experimental Result

Understanding and Planning

Table 6 presents the experimental results of commentary understanding and commentary planning. Due to the significant imbalance in label distribution, the discussion emphasizes the macro-F1 scores. In addition, we also conducted the McNemar test. The baseline represents an ideal condition and is therefore excluded from the test. First, irrespective of the LLM used, the few-shot scheme improves performance in commentary understanding. Second, although GPT-4o achieves the highest performance in the commentary understanding task,

Task		Understanding		Planning	
Scheme	LLM	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Zero-shot	Majority	0.6084	0.0688	0.6084	0.0688
	GPT-4o	0.5910*	0.3803*	0.4638	0.0648*
	Claude 3.5 Sonnet	0.5684	0.3534	0.4911*	0.0606
	Gemini-2.0-flash	0.3784	0.3022	0.1755	0.0530
	DeepSeek R1-1776	0.3430	0.3010	0.0934	0.0469
Few-Shot	Majority	0.6084	0.0688	0.6084	0.0688
	GPT-4o	0.6071*	0.4631*	0.3462	0.0674
	Claude 3.5 Sonnet	0.5427	0.4037	0.4122*	0.0838*
	Gemini-2.0-flash	0.3462	0.3521	0.1450	0.0689
	DeepSeek R1-1776	0.4003	0.2824	0.1836	0.0737

Table 6: Experimental results of commentary understanding and commentary planning. The majority for the planning task always predicts the most frequent label, *Key Summary*. * denotes statistically significant differences between the best-performing and the second-best models according to the McNemar test ($p < 0.05$). The baseline represents an ideal condition and is therefore excluded from the test.

		Professionals			
		KS	SE	CPO	Others
GPT-4o	KS	42.51%	5.96%	4.67%	0.16%
	SE	1.93%	6.76%	1.61%	2.09%
	CPO	6.28%	3.54%	12.88%	0.64%
	Others	4.03%	0.97%	0.48%	5.48%

Table 7: GPT-4o Confusion matrix for the commentary understanding task. KS, SE, and CPO denote *Key-Summary*, *Supplementary Explanation*, and *Commentator's Personal Opinion* categories.

it attains only a macro-F1 score of 0.4631. This result shows that understanding commentary content remains a challenge even for these advanced LLMs. Third, across all tasks, the performance of Gemini is significantly inferior to that of other LLMs, as confirmed by the McNemar test, This observation offers an important insight for future studies that aim to utilize the proposed tasks and dataset, suggesting careful consideration of LLM selection.

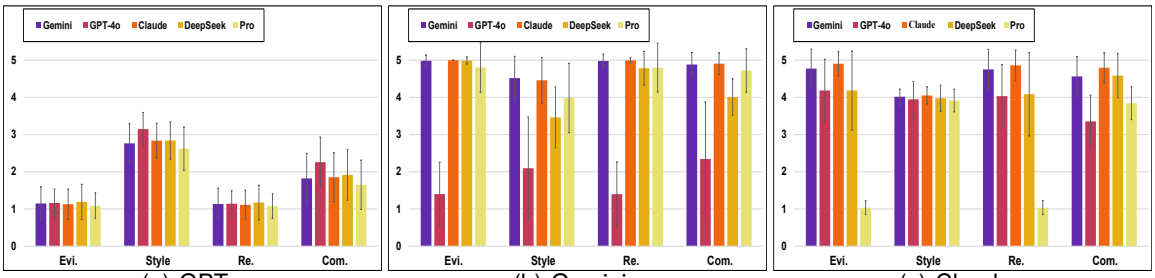
When comparing the performance gap between commentary understanding and planning tasks, it becomes evident that planning is more challenging than understanding. This is attributed to the plausibility of all options in the planning task. Professional commentators leverage not only information from the debate itself but also external sources that provide insights into the global political landscape, enabling more accurate and informed commentary. In contrast, the current approach relies solely on debate information and does not effectively incorporate external data.

Importantly, our experiments further show that for the planning task, the simple baseline method of always predicting the most frequent label *Key-Summary* achieves higher micro-F1 and a macro-F1 that is competitive with, but not uniformly superior to LLM-based models. This counterintuitive result suggests that, under the current data distribution and alignment strategy, LLMs struggle to extract sufficiently meaningful cues from transcript segments and are unable to outperform simple frequency-based heuristics. Several factors may contribute to this phenomenon. First, not all commentary categories have clear evidence in the transcript—categories such as fact-checking, market response, or high-level inference often rely on external information beyond the debate transcript. Second, our annotation protocol aligns commentaries to transcript spans and sometimes aggregates multiple segments, which can dilute the discriminative signals necessary for accurate prediction. In the few-shot setting, the context accessible to the model is limited to a narrow window of debate transcript, further increasing prediction uncertainty and randomness, especially for low-frequency or externally informed labels.

To enable LLMs to emulate expert-level performance more closely, it may be necessary to fine-tune the models with expert knowledge and up-to-date political context. This represents a promising avenue for future research.

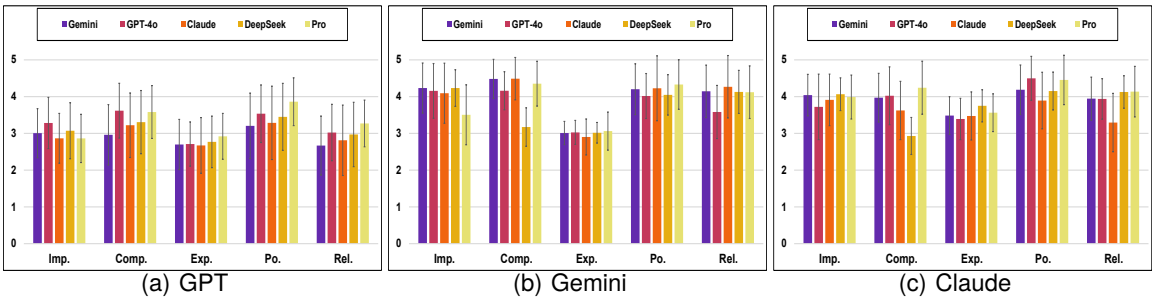
Table 7, Table 8, and Table 9 present the confusion matrices for each task using the best-performing model under few-shot settings. Table 7 shows the model struggles most with distinguishing between *Key-Summary* and *Supplementary Explanation* (extracted vs. paraphrased content). Because the understanding task provides only the commentary, without the corresponding transcript. To examine the effect of this limitation, we conducted an additional experiment on the 2024 Biden-Trump Presidential Debate where both the commentary and transcript were given as input, allowing us to assess the impact of transcript information on classification performance. As shown in Table 8, the result is that adding aligned transcripts improved *Key-Summary* recall from 77.64% to 80.88%, highlighting their value in providing contextual grounding for classification. Table 9 shows that the model often generates summaries during commentary planning, even without exposure to imbalanced training data. It also lacks the professional judgment to seek additional information when needed. Notably, the model frequently offers opinions where experts would only summarize, highlighting that deciding when to present opinions remains a key challenge.

To distinguish the performance of different models and determine whether there are statistically significant links between the outputs of distinct LLMs, we adopt the McNemar test (McNemar, 1947), a widely used technique in social science. Although, overall, most pairwise model comparisons reveal statistically significant differences—indicating that distinct LLMs indeed exhibit varying performance patterns in understanding and planning tasks—a



(a) GPT (b) Gemini (c) Claude

Figure 2: The evaluation results of different LLMs on summaries. Pro denotes content authored by professionals.



(a) GPT (b) Gemini (c) Claude

Figure 3: The evaluation results of different LLMs on commentaries.

		Professionals			
		KS	SE	CPO	Others
GPT w/o	KS	77.65%	34.58%	23.77%	0.81%
	SE	3.53%	39.25%	8.20%	34.68%
	CPO	11.47%	20.56%	65.57%	4.95%
	Others	7.35%	5.61%	2.46%	59.57%
GPT with	KS	80.88%*	16.82%	15.57%	0.81%
	SE	3.53%	34.58%	4.92%	7.26%
	CPO	9.71%	39.25%	76.23%*	29.95%
	Others	5.88%	9.35%	3.28%	61.99%*

Table 8: Confusion matrices (Recall) of GPT-4o for the commentary understanding task. Comparing the impact of aligned transcripts on the 2024 Biden-Trump Presidential Debate.

few comparisons did not reach statistical significance ($p > 0.05$). The comparison between Gemini and DeepSeek was not statistically significant, suggesting no substantial difference in their performance on understanding tasks. This may imply similar comprehension strategies or error distributions between the two models.

On the other hand, Planning Task contains more instances of non-significant results. In planning tasks, models such as Gemini, Claude, and DeepSeek may exhibit more consistent approaches, resulting in less pronounced differences compared to those observed in understanding tasks. These non-significant comparisons serve as a reminder that, while general performance differences exist among models, some LLMs may converge in their behavior on specific task types. Therefore, when aiming to generate diverse commentary, it may be beneficial to select LLMs that demonstrate distinct behavioral patterns, thereby enhancing the richness and variety of the output.

Commentary Generation

By using traditional evaluation metrics, GPT-4o, Gemini, Claude, and DeepSeek got BLEU scores of 0.0024, 0.0058, 0.0075, and 0.0073 respectively; ROUGE scores of 0.0832, 0.1053, 0.1134, and 0.1050; and BERTScores of 0.8324, 0.8448, 0.8472, and 0.8442. Both BLEU and ROUGE scores are relatively low, while BERTScore performs notably well. While n-gram-based metrics such as ROUGE and BLEU, as well as semantic similarity metrics like BERTScore, are suitable for quantifying the surface and semantic similarity between generated and reference commentaries, they do not fully capture the diversity and multifaceted nature of expert commentary. To capture the diversity and multifaceted nature of expert reviews as thoroughly as possible, we referred to the news value concept proposed by (Bednarek, 2010) and defined different evaluation dimensions for both the **Key-Summary** and the **Commentary**. Our results indicate that LLM-generated commentaries can differ in wording and content details from human experts, yet still appear reasonable and fluent. Manual validation (Tables 9) further shows that, in several dimensions, the LLMs' outputs are comparable to human commentaries. Therefore,

		Professionals			
		KS	SE	CPO	Others
Claude	KS	39.29%	11.76%	15.30%	7.09%
	SE	0.48%	0.00%	0.00%	0.00%
	CPO	9.34%	3.86%	3.06%	0.64%
	Others	5.64%	1.61%	1.29%	0.32%

Table 9: Claude3.5-Sonnet Confusion matrix for the commentary planning task. KS, SE, and CPO denote *Key-Summary*, *Supplementary Explanation*, and *Commentator's Personal Opinion* categories.

instead of relying solely on automatic metrics, we also adopt a news value framework to analyze the generated commentaries across multiple dimensions. This approach enables a more comprehensive assessment of commentary quality, capturing both textual similarity and human-judged value.

Figure 2 and 3 presents the evaluation results using GPT-4o, Gemini, and Claude as evaluator models. First, all models exhibit a common phenomenon: when evaluating content generated by themselves, they tend to assign relatively higher scores compared to those given to outputs from other models. Notably, Gemini tends to assign markedly lower scores when evaluating outputs from GPT-4o, and Claude gives particularly low scores to summaries written by experts.

Second, in terms of commentary evaluations, all models consistently assign higher scores to expert-written commentaries across various dimensions. This supports the validity of our evaluation framework in capturing the qualities of expert-level commentary. Furthermore, although commercial APIs generally achieve slightly higher scores than DeepSeek, an interesting observation is that DeepSeek, having utilized a considerable amount of data distilled from GPT, performs relatively well under GPT's evaluations, but fares worse under evaluations from Gemini and Claude. At this point, it remains difficult to definitively determine which model performs best overall. However, we emphasize that our proposed evaluation criteria and framework are capable of effectively identifying the professional features characteristic of expert-written commentary.

To further our investigation, we engaged two undergraduate students to perform the same evaluation using the same criteria. Each annotator scored 50 summaries and 50 commentaries per LLM and expert, resulting in a total of 500 evaluations. Table 10 and 11 shows the average results from the three LLMs and two human raters. First, human evaluators consistently gave higher scores to expert-written summaries, indicating that these summaries are better aligned with the debate transcripts and more effectively capture the key themes and issues than those produced by LLMs. In contrast, GPT-generated summaries received the lowest ratings, suggesting challenges in summarizing debate content effectively. Ranking-wise, Gemini's evaluations showed the highest alignment with human judgments.

LLM	Gemini	GPT-4o	Claude	DeepSeek	Pro
Evi.	3.62(1.91)	2.49(1.71)	3.67(1.92)	3.45(1.88)	2.30(1.91)
Style	3.69(1.78)	3.24(2.02)	3.74(1.81)	3.52(1.79)	3.55(1.74)
Re.	3.61(1.90)	2.49(1.68)	3.65(1.92)	3.37(1.86)	2.32(1.92)
Com.	3.79(1.77)	2.96(1.61)	3.88(1.80)	3.62(1.70)	3.46(1.67)
Annotator	Gemini	GPT-4o	Claude	DeepSeek	Pro
Evi.	3.62(1.53)	1.15(0.43)	3.61(1.59)	3.58(1.60)	4.27(0.99)
Style	3.67(1.00)	2.40(0.98)	3.84(0.89)	3.63(1.08)	3.45(1.11)
Re.	3.18(1.51)	1.09(0.34)	3.20(1.54)	3.27(1.52)	3.78(1.13)
Com.	3.36(1.24)	1.62(1.00)	3.55(1.23)	3.74(1.26)	4.37(0.85)

Table 10: Average results of LLM and human annotators for summaries

LLM	Gemini	GPT-4o	Claude	DeepSeek	Pro
Imp.	3.78(1.86)	3.86(2.00)	3.65(1.84)	3.79(1.87)	3.46(1.83)
Comp.	3.81(1.89)	4.01(2.17)	3.80(2.02)	3.22(1.98)	4.04(2.19)
Exp.	3.03(1.60)	3.10(1.64)	3.01(1.59)	3.17(1.66)	3.11(1.74)
Po.	3.86(1.99)	4.01(2.13)	3.77(2.02)	3.90(2.06)	4.23(2.32)
Rel.	3.57(1.75)	3.62(1.89)	3.41(1.83)	3.73(1.84)	3.81(1.99)
Annotator	Gemini	GPT-4o	Claude	DeepSeek	Pro
Imp.	3.83(0.84)	3.70(0.97)	3.90(0.85)	3.83(0.94)	3.70(1.02)
Comp.	4.23(0.67)	4.51(0.62)	4.22(0.50)	4.23(0.83)	4.51(0.63)
Exp.	1.94(0.85)	2.18(0.92)	2.04(0.87)	1.94(0.80)	2.18(0.86)
Po.	3.83(0.58)	3.78(0.55)	3.90(0.45)	3.83(0.66)	3.78(0.45)
Rel.	3.89(0.58)	3.90(0.55)	3.98(0.46)	3.89(0.63)	3.90(0.56)

Table 11: Average results of LLM and human annotators for commentaries

Conclusion and Future Direction

This paper introduced DR-CUP, a dataset that aligns U.S. presidential debate transcripts with real-time professional commentary from 2016 to 2024. This resource provides a foundation for studying commentary understanding, planning, and generation, and fills a critical gap in political discourse research by incorporating expert-level analysis across a wide range of debate contexts. Through extensive pilot experiments, we benchmarked state-of-the-art LLMs on multiple tasks and revealed the challenges these models face, especially in generating insightful and context-rich commentary.

The DR-CUP dataset opens new directions in NLP and political communication. By aligning professional commentary with debate transcripts, it enables research on AI-generated political commentary that integrates external knowledge like real-time news or policy documents for richer, context-aware analysis. Beyond technical advances, DR-CUP supports interdisciplinary studies on how expert commentary shapes public opinion and audience perception during high-stakes debates. Currently, it draws mainly from Bloomberg, which ensures quality but limits ideological diversity. Future versions could include more sources to broaden coverage and improve model generalization.

Ethics Statements and Limitations

The first limitation is that about 33% of the commentary relies on external knowledge, such as expert information, real-world events, or the current debate context. Since the model is limited to using only the debate transcript, it struggles to generate expert-level commentary and lacks the ability to query supplementary materials, which further restricts its capacity to provide deeper insights.

The second limitation of our paper is that the model tends to generate summaries rather than more subjective supplementary explanations and opinions. This is because the majority (60.84%) of the commentary in the dataset is labeled as Key Summary, while *Supplementary Explanation* and "*Commentator's Personal Opinion*" are relatively less frequent. We retained all 11 categories during annotation, despite the fact that some categories are heavily imbalanced, in order to faithfully reflect the composition of professional commentary. In the future, we will consider adjusting the labeling scheme in response to data imbalance or augmenting commentaries from other media sources to mitigate the effects of skewed distributions on model performance.

Acknowledgements

This work has been co-funded by AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain." This work was also supported by the National Science and Technology Council, Taiwan, under the Grant NSTC 114-2221-E-194-051, NSTC 114-2628-E-194-005-MY3, and NSTC 114-2634-F-110-001-MBK.

References

Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, 11(3):1–27.

Monika Bednarek. 2010. Evaluation in the news: a methodological framework for analysing evaluative language in journalism. *Australian Journal of Communication*, 37(2):15–50.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Steven Fein, George R Goethals, and Matthew B Kugler. 2007. Social influence on political judgments: The case of presidential debates. *Political Psychology*, 28(2):165–192.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *ACL*, pages 4684–4690.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020a. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1008–1018, Marseille, France. European Language Resources Association.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2020b. Extracting implicitly asserted propositions in argumentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24–38.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. 2022. Open-domain video commentary generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7326–7339. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262.

Craig T Robertson, William H Dutton, Robert Ackland, and Tai-Quan Peng. 2019. The democratic role of social media in political debates: The use of twitter in the first televised us presidential debate of 2016. *Journal of Information Technology & Politics*, 16(2):105–118.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air ....

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5077–5089, Hong Kong, China. Association for Computational Linguistics.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1361–1371. Association for Computational Linguistics.

A. Appendices

A.1. Annotation and Evaluation Workflow

To ensure the reproducibility of our dataset and address the detailed setup of our human annotation and evaluation, we provide further specifics regarding our protocols. The data annotation was conducted by two non-U.S. graduate students with backgrounds in computational linguistics, while the evaluation of the generation quality was performed by two different non-U.S. undergraduate students. To ensure ethical research practices, all participants received reasonable compensation that met local standard hourly wage rates.

Annotation. During the initial phase, the two annotators conducted independent annotations on the same complete debate event. When label disagreements occurred, they engaged in collaborative discussions until a consensus was reached to determine the final labels. Using this agreed-upon standard as a baseline calibration, they then divided the workload to process the remaining debate data. For each segment, the judgment process was strictly based on the predefined label definitions and textual content, ensuring careful alignment with the debate transcripts.

Evaluation. For the generation quality evaluation, the 50 evaluated instances (summaries and commentaries) were randomly sampled exclusively from the four debate events held between 2023 and 2024. We deliberately chose not to provide prior training on U.S. politics to the evaluators. Because the target audience for professional expert commentary is the general public, having undergraduate students without special political training perform the evaluation more authentically simulates the perception and comprehension levels of general readers when consuming these commentaries.

Inter-evaluator agreement. Table 12 presents the results of the inter-evaluator comparison. We primarily compared the feedback provided by the two annotators after they reviewed the expert commentary. The results indicate a divergence in their preferences, suggesting that professional commentary elicits varying responses depending on the individual reader.

Comm.	Cohens Kappa	Sum.	Cohens Kappa
Imp.	0.0941	Evi.	0.1037
Comp.	0.1140	Style	0.0799
Exp.	0.2046	Re.	0.1736
Po.	0.0007	Com.	-0.0320
Rel.	-0.0447		

Table 12: The kappa results of comparison between two human evaluators.

	Compared model			
Reference	GPT	Gemini	Claude	DeepSeek
		Zero-shot		
GPT	–	**1.9e-50**	**1.5e-2**	**3.85e-40**
Gemini	**1.9e-50**	–	**5.9e-2**	**1.0e-3**
Claude	**1.5e-2**	**5.9e-2**	–	**2.39e-6**
DeepSeek	**3.85e-40**	**1.0e-3**	**2.39e-6**	–
		Few-shot		
GPT	–	**2.31e-57**	**4.09e-30**	**4.93e-32**
Gemini	**2.31e-57**	–	**3.53e-19**	**1.80e-12**
Claude	**4.09e-30**	**3.53e-19**	–	2.68e-1
DeepSeek	**4.93e-32**	**1.80e-12**	2.68e-1	–

Table 13: McNemar test p-values for commentary understanding. Values < 0.05 indicate rejection of the null hypothesis. The results highlighted in bold are statistically significant.

A.2. Comparison Between Different Models

To distinguish the performance of different models and determine whether there are statistically significant links between the outputs of distinct LLMs., we adopt the McNemar test (McNemar, 1947), a widely used technique in social science. Tables 13 and 14 present the test results for the *Understanding* and *Planning* tasks, respectively. Each table compares different LLMs, with the p-values representing the outcomes of the McNemar tests.

Although, overall, most pairwise model comparisons reveal statistically significant differences—indicating that distinct LLMs indeed exhibit varying performance patterns in understanding and planning tasks—a few comparisons did not reach statistical significance ($p > 0.05$). In Table 13 (Understanding Task), the comparison between Gemini and DeepSeek was not statistically significant, suggesting no substantial difference in their performance on understanding tasks. This may imply similar comprehension strategies or error distributions between the two models.

On the other hand, Table 14 (Planning Task) contains more instances of non-significant results. In planning tasks, models such as Gemini, Claude, and DeepSeek may exhibit more consistent approaches, resulting in less pronounced differences compared to those observed in understanding tasks. These non-significant comparisons serve as a reminder that, while general performance differences exist among models, some LLMs may converge in their behavior on specific task types. Therefore, when aiming to generate diverse commentary, it may be beneficial to select LLMs that demonstrate distinct behavioral patterns, thereby enhancing the richness and variety of the output.

Reference	GPT	Gemini	Claude	DeepSeek
		Zero-shot		
GPT	–	**3.52e-44**	**1.71e-41**	**3.85e-40**
Gemini	**3.52e-44**	–	**1.5e-2**	1.55e-1
Claude	**1.71e-41**	**1.5e-2**	–	2.72e-1
DeepSeek	**3.85e-40**	1.55e-1	2.72e-1	–
		Few-shot		
GPT	–	**7.99e-22**	**1.49e-24**	**1.96e-13**
Gemini	**7.99e-22**	–	5.9e-2	**1.0e-3**
Claude	**1.49e-24**	5.9e-2	–	**2.39e-6**
DeepSeek	**1.96e-13**	**1.0e-3**	**2.39e-6**	–

Table 14: McNemar test p-values for commentary planning.

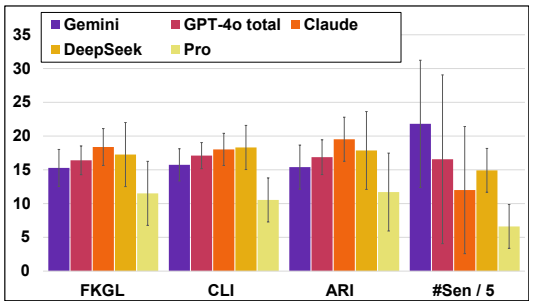


Figure 4: The results of text readability evaluation by different indicators.

A.3. Readability

We examined the differences between texts generated by LLMs and those written by human experts, focusing on textual complexity. We utilized several readability metrics, including the Flesch-Kincaid Grade Level (Kincaid et al., 1975), Coleman-Liau Index (Coleman and Liau, 1975), Automated Readability Index (Smith and Senter, 1967), as well as the number of sentences (# Sen) as an indicator of text length. As illustrated in Figure 4, expert-written texts consistently scored markedly lower across all four metrics compared to those generated by LLMs. This suggests that expert-authored content tends to be more concise and easier to read. In contrast, LLM-generated texts are generally more verbose and expansive in scope, which can diminish their overall readability.

A.4. Prompt Format

The comprehensive prompt templates for all tasks, along with a selection of our experimental results, are hosted at `https://lurl.cc/sCFVjk`.