MADS: Ensemble LLM-based Multi-Agent Debate System for Political Bias Detection

Yik Yu Ng
School of Computer Science
McGill University
Montreal, Canada
yik.ng@mail.mcgill.ca

Benjamin C. M. Fung
School of Information Studies
McGill University
Montreal, Canada
ben.fung@mcgill.ca

Elena Obukhova
Desautels Faculty of Management
McGill University
Montreal, Canada
elena.obukhova@mcgill.ca

Djedjiga Mouheb
Department of Computer Science and Software Engineering
Université Laval
Quebec City, Canada
djedjiga.mouheb@ift.ulaval.ca

Ching-Chun Huang
Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
chingchun@nycu.edu.tw

Shih-Chia Huang
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
schuang@mail.ntut.edu.tw

***Abstract*—Political bias detection in news media presents fundamental challenges: articles interweave facts, opinions, and selective framing, where subtle omissions can be as influential as explicit statements. Traditional supervised approaches require costly expert annotations that struggle to keep pace with evolving online content. We introduce MADS, an unsupervised framework leveraging multi-agent debate among large language models to detect political bias without training data. By orchestrating structured deliberation among three diverse LLMs, our approach enables models to challenge each other's interpretations through evidence-based argumentation, exposing biases that individual models miss while providing interpretable reasoning chains. Across evaluations on 3 datasets (over 7,000 articles) from major U.S. media outlets, MADS achieves 78.15% accuracy on expert-annotated data and performs comparably to supervised BERT baselines without any labeled examples. We also contribute a benchmark of 473,989 articles collected during the 2024 U.S. election period. Our results show that structured multi-agent debate provides robust, interpretable bias classification that no single model can consistently match across datasets and political leanings.**

***Index Terms*—Model ensemble, Multi-agent systems, Large language models, Bias detection, Unsupervised learning, Model Interpretability**

I. Introduction

Algorithmic content curation has created a paradox in modern news consumption: while readers access more diverse sources than ever before, personalization systems often narrow their exposure to politically aligned content, obscuring the partisan framing present across media outlets [1], [2]. Political bias in news media manifested through selective framing, strategic omissions, and partisan language choices, poses a fundamental challenge to informed democratic discourse [3]. Political bias detection in news articles presents unique computational challenges. News articles typically span thousands of tokens, interweaving factual reporting with editorial choices, source quotations with narrative framing, and explicit statements with strategic omissions. Unlike sentiment analysis or topic classification, bias detection requires understanding not only what is said but also how it is said and, critically, what is left unsaid [4]. This challenge is amplified by increasing political polarization across democratic nations [5], where even exposure to opposing viewpoints on social media can paradoxically increase partisan divisions [6]. The theoretical foundations of framing have been extensively documented [7], [8], yet automating the detection of such framing remains computationally challenging. Traditional supervised approaches to bias detection face fundamental limitations. Creating training datasets requires expensive expert annotation [9], and these models struggle to generalize beyond their training distribution. In addition, single-model approaches, whether supervised or unsupervised, remain vulnerable to systematic biases inherited from their training data. Recent studies have shown that large language models (LLMs) exhibit their own political biases that affect classification outcomes [10], [11], creating a paradox where potentially biased models are used to detect bias in text. We propose a novel solution, called *Multi-Agent Debate System* (*MADS*), through a conditional multi-agent debate framework that orchestrates deliberation among multiple LLMs to detect political bias without training data. MADS transforms bias detection from a single model classification task into a collaborative reasoning problem where diverse models challenge each other's interpretations through evidence-based argumentation. By requiring agents to ground their arguments in textual evidence, MADS exposes biases that individual models miss while providing transparent explanations. Our key insight is that not all bias detection decisions require the same level of scrutiny. When multiple models independently agree on a classification, simple aggregation suffices. However, when models disagree, often indicating subtle or ambiguous bias, structured debate can surface overlooked evidence and reconcile conflicting interpretations.

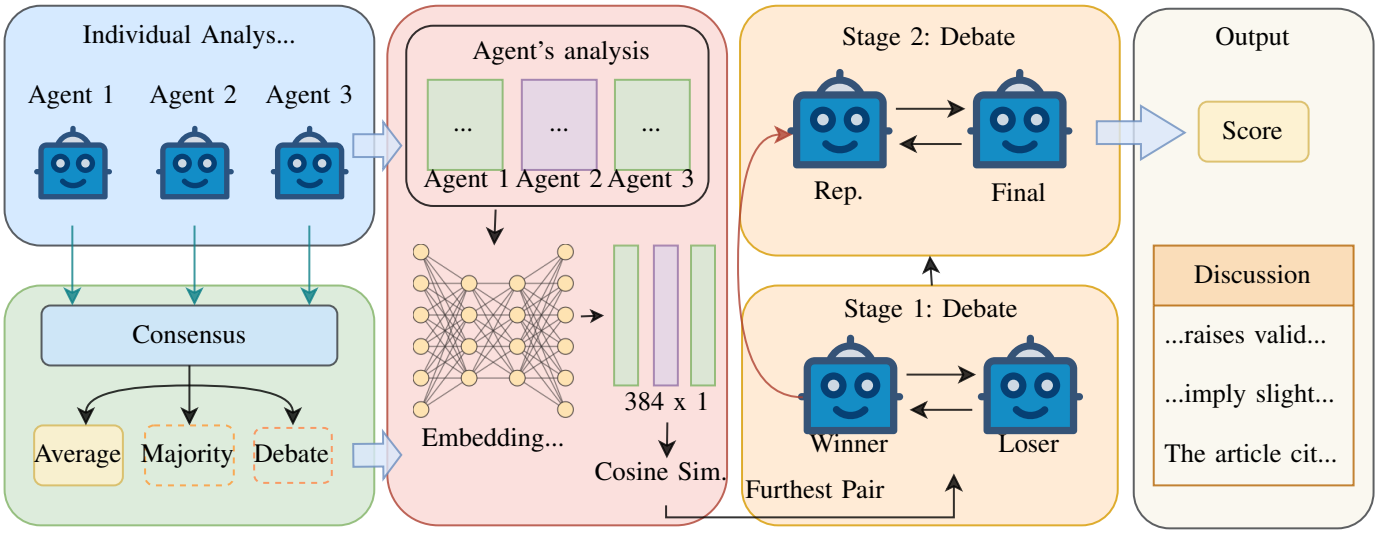

Fig. 1. Multi-agent debate pipeline for political bias detection.

This conditional approach in MADS maintains computational efficiency while ensuring thorough analysis of challenging cases. The contributions of this work are threefold:

- We introduce MADS, an unsupervised multi-agent debate framework for political bias detection that achieves strong performance without training data, performing on par with traditional supervised baselines including LSTM and BERT models while requiring no labeled examples for training.
- MADS provides interpretable bias detection through evidence-based argumentation, where each classification is accompanied by transparent reasoning chains that reveal the textual features, framing choices, and omissions identified as indicators of political bias.
- We construct a comprehensive evaluation benchmark including a custom dataset of 473,989 news articles collected from major media outlets during the six months preceding the 2024 U.S. presidential election (May–Nov).
- Our code and dataset are publicly available at: https://github.com/yyn-anson/MADS-Political-Bias-Detection

II. Related Work

A. Political Bias Detection

Political bias detection has evolved from lexicon-based approaches to sophisticated neural architectures. Early quantitative work compared media citations with congressional patterns [12] and analyzed partisan language through word frequencies [13]. A later study crowdsourced 10,000 article annotations for supervised classification [9]. The BASIL dataset introduced a fine-grained analysis that distinguishes lexical bias from informational bias [4]. Recent transformer-based approaches achieved significant gains, with BERT reaching 70% accuracy on AllSides data [14]. A critical gap remains in detecting omission bias: the selective exclusion of information. Prior work demonstrates that what outlets choose not to report can be more biased than their framing choices [15]. Recent work addresses this gap through multi-perspective approaches [16], [17]. Prompting-based methods show particular promise because LLM prompting can match traditional models without training data, although single-model approaches remain vulnerable to inherent biases [18]. Additional work on cross-lingual detection [19], [20], hybrid LLM-rule approaches [21], and the correlation between GPT-4 and human bias judgments [22] further informs our multi-model strategy.

B. Multi-Agent Systems and LLM Biases

Multi-agent frameworks demonstrate significant improvements in LLM reasoning capabilities. Debate among agents can improve factual accuracy through adversarial questioning [23], while divergent-thinking procedures can increase solution diversity [24]. Tree-of-thoughts reasoning explores multiple reasoning paths [25], and the reasoning bounds of multi-agent systems have been examined for complex tasks [26]. Architectural insights from generative agents [27] and conversation frameworks [28] inform the design of agent interactions. However, political contexts present unique challenges. LLM agents may converge toward inherent biases despite assigned political roles, including a consistent drift to the left regardless of their initial positions [29]. Neutral agents may also align consistently with Democratic positions, revealing systematic deliberation bias [30]. These findings highlight the need for mechanisms to prevent false consensus. Extensive research documents political biases in LLMs themselves. Political bias evaluations reveal consistent left-leaning responses [10], distinguish content bias from stylistic bias [31], and trace biases to training data composition [11]. LLM outputs also show significant alignment with WEIRD populations [32], while comprehensive surveys [33], [34] provide bias taxonomies. Mitigation strategies show mixed results: adversarial training [35] and self-reflection [36] achieve some bias reductions, and simple debiasing instructions can be effective [37],

though complete neutrality remains elusive [38]. Rather than attempting to debias individual models, our approach leverages multiple biased models in structured opposition. Ensemble methods can reduce individual biases when properly designed. Diversity maximization can improve robustness [39], but naive aggregation can amplify shared biases. Recent work identifies "bias reinforcement," in which debate strengthens rather than corrects existing biases [40]. Even diverse agents can converge on false consensus [11]. Our conditional debate framework addresses this by triggering deliberation only on disagreement and implementing confidence-based resolution to prevent artificial centralization.

III. Task Description

Problem Definition. We formulate political bias detection as an unsupervised multi-class classification task. Given a news article x containing n tokens, our goal is to predict its political bias $y \in \mathcal{Y} = \{\text{Left}, \text{Center}, \text{Right}\}$ without access to labeled training data. Unlike supervised approaches that learn from human-annotated examples, our framework must rely solely on the reasoning capabilities of pre-trained language models to identify bias patterns. The input is a raw news article text x of arbitrary length (typically 500–5,000 tokens), covering diverse topics and containing quotes, citations, and editorial commentary. The output consists of two components: (1) a categorical bias label $\hat{y} \in \mathcal{Y}$, and (2) a continuous bias score $\hat{s} \in [-3, 3]$ providing granular assessment, where negative values indicate left-leaning bias, positive values indicate right-leaning bias, and values near zero suggest balanced reporting.
Key Assumptions. Our approach assumes access to multiple LLMs with different training corpora and architectures (*model diversity*), that these models possess sufficient pre-trained knowledge about political discourse to perform zero-shot bias detection, and that articles are provided in their entirety without truncation. The current implementation focuses on English-language news from U.S. media outlets, with articles falling within the models' knowledge cutoff dates. Bias detection relies solely on textual content, excluding external knowledge bases, fact-checking, or article metadata such as publication source.

IV. Methodology

Building on the task formulation in Section 3, we deploy three LLM agents $\mathcal{A} = \{a_1, a_2, a_3\}$ to analyze article x, following the pipeline illustrated in Figure 1. Each agent a_i produces an *initial analysis* consisting of a lean score $s_i^{(0)} \in [-3, 3]$ and reasoning text $r_i^{(0)}$. We map scores to directions in $\mathcal{Y} = \{\text{Left}, \text{Center}, \text{Right}\}$ via

$$\text{dir}(s) = \begin{cases} \text{Left}, & s \le -\tau, \\ \text{Right}, & s \ge \tau, \\ \text{Center}, & \text{otherwise}, \end{cases} \qquad \tau = 1. \tag{1}$$

We denote $d_i^{(t)} = \text{dir}(s_i^{(t)})$ as the direction of agent i after round t (with $t = 0$ for the initial analysis). During the debate, agents can update $s_i^{(t)}$ and $r_i^{(t)}$. A *debate history* $\mathcal{H}$ is an ordered list of challenge/response JSON messages (see supplementary materials), shared among agents and appended each round. We clamp all scores to $[-3, 3]$ after each update.

A. Phase I: Independent Analysis

Each agent analyzes x with a standardized prompt and returns $(s_i^{(0)}, r_i^{(0)})$. We compute initial directions $d_i^{(0)} = \text{dir}(s_i^{(0)})$.

B. Phase II: Consensus Check and Routing

Let $\mathcal{D}^{(0)} = \{d_1^{(0)}, d_2^{(0)}, d_3^{(0)}\}$.

- **Unanimous**: if all three directions coincide, return $\hat{s} = \frac{1}{3}\sum_i s_i^{(0)}$ and $\hat{d} = \text{dir}(\hat{s})$ (early exit).
- **Majority (2v1)**: if exactly two directions coincide, route to the majority debate procedure (Section IV-C1).
- **All Different (1v1v1)**: if $|\mathcal{D}^{(0)}| = 3$, route to the all-different debate procedure (Section IV-C2).

C. Phase III: Debate Resolution

When agents do not reach unanimous agreement, we conduct structured debate to resolve disagreements. The debate procedure differs depending on the disagreement type.

1) Majority Debate (2v1).: When two agents share a direction and one dissents, the higher-confidence agent from the majority pair is selected as the representative to debate the dissenter. Confidence is estimated using the entropy-based measure described in Section IV-C5. The representative and dissenter then engage in the standard debate procedure (Section IV-C3) for up to $R_{\max}$ rounds. If the agents converge on a shared direction, we check for panel unanimity across all three agents; if achieved, we return the mean of the three current scores. Otherwise, the confidence-based tiebreaker (Section IV-C5) determines the winner.

2) All-Different Debate (1v1v1).: When all three agents initially disagree, we conduct a two-stage debate. First, we identify the most divergent pair using embedding geometry: we embed each initial reasoning $r_i^{(0)}$ using `sentence-transformers/all-MiniLM-L6-v2` and compute ℓ_2-normalized embeddings $e_i = \phi(r_i^{(0)})/\|\phi(r_i^{(0)})\|_2$. Here, $i, j \in \{1, 2, 3\}$ index the three agents, and we select the pair with the lowest cosine similarity:

$$(p, q) = \arg\min_{i \neq j} e_i^\top e_j, \quad m = \{1, 2, 3\} \setminus \{p, q\}, \tag{2}$$

so that stage 1 exposes maximally distinct perspectives. In **Stage 1**, agents a_p and a_q debate for up to $R_{\max}$ rounds. If they converge, the higher-confidence agent becomes the representative; otherwise, the confidence-based tiebreaker selects one. In **Stage 2**, the stage-1 representative debates the remaining agent a_m, carrying forward the accumulated history $\mathcal{H}$. The same round mechanics, termination conditions, and confidence-based resolution apply. If panel unanimity is achieved at any point, we return the mean of all three current scores; otherwise, the final winner's score and direction are returned. The complete algorithmic procedures are formalized in the supplementary materials.

3) Round Mechanics.: At round $t = 1, 2, \ldots, R_{\max}$, agents alternate roles as challenger and target. The challenger reads the article, its own latest analysis $(s^{(t-1)}, r^{(t-1)})$, the target's latest analysis, and the history $\mathcal{H}$, and then emits a structured *challenge* JSON. The target replies with a *response* JSON. Both may update their scores $s^{(t)}$ and reasoning $r^{(t)}$ (clamped to $[-3, 3]$). After each exchange, we append to $\mathcal{H}$ and check termination conditions. Maintaining a shared history allows diverse opinions to propagate across agents, reducing inherited LLM bias, while ensuring format consistency across rounds.

4) Termination Conditions.: A debate between agents i and j terminates under three conditions. First, *pair consensus*: the debate ends immediately if $\mathrm{dir}(s_i^{(t)}) = \mathrm{dir}(s_j^{(t)})$. Second, *stagnation detection*: at each round, we embed the current argument text and compute its cosine similarity with the previous round's argument; if the similarity exceeds 0.9, we terminate early to prevent redundant exchanges and conserve tokens. Third, *round limit*: if $t = R_{\max}$ without consensus, the debate ends. When termination occurs without consensus (via stagnation or round limit), the confidence-based tiebreaker (Section IV-C5) selects the winner.

5) Confidence-Based Tiebreaking.: When a debate concludes without consensus, we select the winner based on prediction confidence rather than score magnitude. Each agent's prompt concludes with a categorical prediction statement (e.g., "The final prediction is [Left/Center/Right]"). We extract the token-level probability distribution $p_i = (p_i^L, p_i^C, p_i^R)$ over the three classes from the model's output logits at the prediction token position. The prediction entropy is then:

$$H(a_i) = -\sum_{c \in \mathcal{Y}} p_i^c \log p_i^c \tag{3}$$

The agent with lower entropy (higher confidence) is selected as the winner: $\text{winner} = \arg\min_i H(a_i)$. This mechanism avoids the bias toward extreme scores that magnitude-based selection would introduce, instead grounding the tiebreaker in the model's own distributional certainty.

D. Challenge–Response Protocol

We enforce a structured JSON exchange where the challenger must acknowledge valid points from the target, cite concrete evidence from x, and propose an adjusted lean score. The target responds with counterevidence and optionally updates its lean. The prompt templates are provided in the supplementary materials.

E. Output

The framework returns a direction label $\hat{d} \in \{\text{Left}, \text{Center}, \text{Right}\}$ and a calibrated lean score $\hat{s} \in [-3, 3]$. If panel unanimity is achieved at any time, $\hat{s}$ is the mean of the three current scores; otherwise, $\hat{s}$ equals the final winner's score.

F. Analysis of the Debate Mechanism

Among the 3,000 articles processed from the Ad Fontes dataset, approximately 81% achieved unanimous agreement across all three models, allowing for efficient early exit without debate. The remaining 19% triggered debate: 18% as majority (2v1) cases and 1% as all-different (1v1v1) cases, the latter involving articles with subtle bias, satirical content, or mixed political signals. In such cases, adversarial dialogue exposes bias invisible to single-agent analysis, ultimately converging on the correct classification. The quantitative impact of debate is analyzed in Section V-C2.

V. Experiment Setup & Result

A. Datasets

We evaluate our framework on three complementary datasets to ensure robustness across different annotation schemes and levels of supervision.

1) Ad Fontes Media.: The Ad Fontes Media dataset [41] provides article-level political bias ratings assigned by trained human analysts using a structured methodology that includes blind multi-analyst review and inter-rater reliability checks. We sample 3,000 articles with balanced class distribution (1,000 Left, 1,000 Center, 1,000 Right), as summarised in Table I. Unlike outlet-level datasets, each article in this dataset carries an individual bias rating independent of its source outlet, making it a higher-fidelity benchmark for article-level classification.

2) Baly et al.: We adopt the dataset from Baly et al. [14], which provides outlet-level political bias labels sourced from AllSides [42]. AllSides assigns ratings to news outlets using methodologies including blind surveys and independent editorial reviews. We randomly sample 3,000 articles with balanced class distribution; after excluding 36 articles that failed content extraction, 2,964 articles (994 Left, 983 Center, 987 Right) were used for evaluation, as shown in Table I. Because all articles from the same outlet share the same label regardless of individual article content, this dataset introduces label noise at the article level, contributing to greater classification difficulty compared to Ad Fontes.

3) Custom Dataset.: To assess performance on recent media content, we curated 1,300 news articles from the GDELT Project[1] (Table I). We initially collected 473,989 articles published between May and November 2024, covering the period before the 2024 U.S. presidential election. After political keyword filtering, we sampled 100 articles from each of 13 major U.S. media outlets: CNN, BBC, Fox News, Breitbart, The Guardian, New York Times, TIME, New York Post, MSNBC, The Nation, Washington Examiner, Newsweek, and Forbes. Outlet-level bias labels were assigned according to AllSides ratings [42]: 6 outlets are classified as Left (CNN, The Guardian, New York Times, MSNBC, The Nation, TIME), 3 as Center (BBC, Newsweek, Forbes), and 4 as Right (Fox News, Breitbart, New York Post, Washington Examiner), yielding 600 Left, 300 Center, and 400 Right articles respectively. We merge *Lean Left* into *Left* and *Lean Right* into *Right*, yielding three-way classification consistent with the other datasets.

[1] https://www.gdeltproject.org/

TABLE I
SUMMARY OF DATASETS USED IN EXPERIMENTS

Dataset	Left	Center	Right	Article Label	Outlet Label
Ad Fontes	1,000	1,000	1,000	Yes	No
Baly et al.[a]	1,000	1,000	1,000	Yes	No
Custom (GDELT)[b]	600	300	400	No	Yes

[a] Of 3,000 sampled articles, 36 were excluded due to content extraction failures, yielding 2,964 articles (994 Left, 983 Center, 987 Right) for evaluation.

[b] Class counts reflect outlet-level AllSides ratings: 6 Left-rated outlets (CNN, The Guardian, New York Times, MSNBC, The Nation, TIME), 3 Center-rated outlets (BBC, Newsweek, Forbes), and 4 Right-rated outlets (Fox News, Breitbart, New York Post, Washington Examiner), each contributing 100 articles.

B. Models

We deploy two tiers of LLM agents. MADS uses three large-scale models (14B–22B parameters): Qwen3-14B, GPT-OSS-20B, and Mistral-Small-22B. MADS Small uses three small-scale models (7B–8B parameters): Qwen2.5-7B, Llama 3.1-8B, and Mistral-7B. For embedding-based pair selection, we use all-MiniLM-L6-v2. Within each tier, models represent diverse training corpora and architectures to mitigate shared bias. All models are run locally on NVIDIA A6000 GPUs with standardized zero-shot prompting (see supplementary materials for prompt templates). Table II lists all model specifications.

C. Article-Level Performance

1) Performance on Ad Fontes Dataset.: Table V presents comprehensive performance results on 3,000 articles from the Ad Fontes dataset. Individual model accuracy ranges from 77.22% (GPT-OSS) to 79.50% (Qwen3-14B), revealing substantial variation that underscores the challenges in political bias detection and motivates our multi-agent approach. MADS achieves 78.15% accuracy and 78.29 Macro-F1. While Qwen3-14B achieves the highest individual accuracy on this dataset, its advantage does not hold across datasets or per-class metrics (see Sections V-C3 and V-C4). The majority of articles (80.9%) achieve unanimous consensus across all three agents, allowing for efficient early exit. The remaining 19.1% trigger debate: 17.8% as majority (2v1) cases and 1.3% as all-different (1v1v1) cases. We analyze the impact of debate on these subsets in Section V-C2. To contextualize our approach against existing baselines, Table III compares MADS with zero-shot LLM baselines reported by Haroon et al. [43] on Ad Fontes and supervised baselines from Baly et al. [14]. Both sets of baselines use only article content as input (excluding title and source metadata) to ensure fair comparison. While GPT-4o achieves 64% accuracy in their zero-shot evaluation, our framework achieves 78.15% accuracy with smaller open-source models (14B–22B parameters). We note that this comparison involves different prompting strategies, so the gap reflects both prompt design and the multi-agent framework rather than the debate mechanism alone. The isolated contribution of debate is quantified in Section V-C2. On the Baly dataset, despite requiring no training data, MADS achieves 52.40% accuracy, comparable to and marginally exceeding both LSTM (46.42%) and BERT (51.41%) models trained on labeled examples. The lower absolute accuracy on this dataset reflects the outlet-level labeling scheme discussed in Section V-A2.

2) Effect of Debate on Classification.: To isolate the contribution of the debate mechanism, Table IV reports accuracy before and after debate for the subset of articles where debate was triggered on the Ad Fontes dataset. For MADS, debate occurs in 573 of 3,000 articles. In majority (2v1) cases, debate produces a slight accuracy decrease of 1.05 percentage points, suggesting that the majority vote is already a strong signal that debate occasionally overturns. However, in all-different (1v1v1) cases, where all three agents initially disagree, debate yields a substantial gain of 12.28 percentage points on these articles with subtle or mixed political signals. For MADS Small, debate improves accuracy across all scenarios, with a 3.02 percentage point overall gain, indicating that smaller models benefit more consistently from deliberation. These results confirm that triggering debate selectively on disagreement cases preserves the strong consensus signal while improving the most ambiguous articles.

3) Per-Class Analysis.: Table V reveals that individual models exhibit inconsistent per-class strengths across datasets. For instance, Qwen3-14B leads Left F1 on Ad Fontes (80.66) but falls behind Mistral-22B on Baly Left (50.52 vs. 55.39), while Mistral-22B's advantage on Baly Left does not carry over to Ad Fontes. This cross-dataset instability confirms that no single model reliably dominates across all classes and distributions. MADS addresses this by combining complementary strengths through structured deliberation, achieving more consistent per-class performance across both datasets than any individual constituent model. On the Baly dataset, Center classification becomes particularly challenging, with MADS achieving 69.48% Center recall but only 44.79% precision, reflecting the difficulty of outlet-level label assignment where individual articles may not match their outlet's overall political leaning.

4) Model Comparison and Ensemble Justification.: The best individual model varies by dataset: Qwen3-14B leads on Ad Fontes (79.50%) but drops to 52.33% on Baly, while Mistral-22B leads on Baly (53.54%) but scores 77.84% on Ad Fontes. While MADS does not achieve the highest accuracy on either dataset individually (78.15% vs. Qwen3-14B's 79.50% on Ad Fontes; 52.40% vs. Mistral-22B's 53.54% on Baly), it delivers consistent performance across both without requiring advance knowledge of which model will perform best, making it a reliable choice for real-world deployment where such oracle information is unavailable. Comparing the two ensemble tiers, MADS substantially outperforms MADS Small on Ad Fontes (78.15% vs. 75.66%), confirming that larger models benefit the ensemble. On Baly, the gap narrows (52.40% vs. 51.87%), suggesting that outlet-level label noise reduces the advantage of model scale.

TABLE II
MODEL SPECIFICATIONS FOR ALL AGENTS USED IN MADS AND MADS SMALL.

Tier	Model	HuggingFace ID	Context
Large	Qwen3-14B	`Qwen/Qwen3-14B`	32k
	GPT-OSS-20B	`openai/gpt-oss-20b`	128k
	Mistral-Small-22B	`mistralai/Mistral-Small-Instruct-2409`	32k
Small	Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	32k
	Llama 3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	128k
	Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	32k
Embedding	all-MiniLM-L6-v2	`sentence-transformers/all-MiniLM-L6-v2`	–

TABLE III
COMPARISON WITH EXTERNAL BASELINES. **LEFT:** ZERO-SHOT LLM BASELINES ON AD FONTES (FROM HAROON ET AL. [43], ARTICLE CONTENT ONLY). **RIGHT:** SUPERVISED BASELINES ON THE BALY DATASET (FROM BALY ET AL. [14], ARTICLE CONTENT ONLY).

Method	Acc (%)
GPT-4o [43]	64
Llama2-13B [43]	49
Mistral-7B [43]	59
MADS	**78.15**

Method	Acc (%)	Macro F1
LSTM [14]	46.42	0.4544
BERT [14]	51.41	0.4826
MADS	**52.40**	**0.5220**

TABLE IV
EFFECT OF DEBATE ON ACCURACY FOR THE AD FONTES DATASET. WE REPORT THE INITIAL ACCURACY (AVERAGE OF INDIVIDUAL AGENTS) AND FINAL ACCURACY (AFTER DEBATE) FOR THE SUBSET OF CASES WHERE DEBATE OCCURRED.

Model Scale	Debate Scenario	N	Initial Acc (%)	Final Acc (%)	Gain (%)
MADS	All Debate Cases	573	49.39	49.21	-0.18
	Majority (2v1)	535	50.40	49.35	-1.05
	All Different (1v1v1)	38	35.09	47.37	+12.28
MADS Small	All Debate Cases	863	50.17	53.19	+3.02
	Majority (2v1)	782	51.32	54.48	+3.16
	All Different (1v1v1)	81	39.09	40.74	+1.65

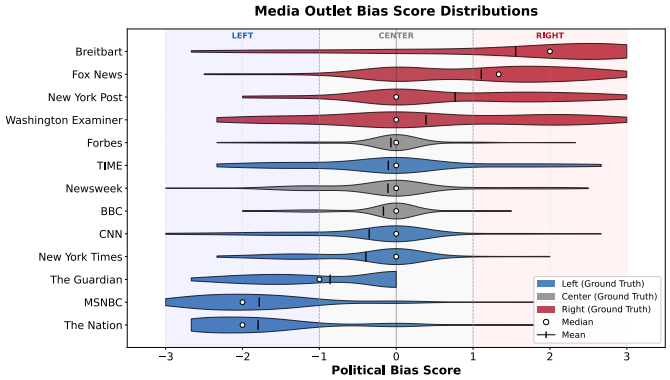

Fig. 2. Distribution of bias scores for articles across 13 major US media outlets. Colors indicate AllSides ground truth (blue: Left, gray: Center, red: Right). Background shading demarcates bias regions: Left, Center, Right.

D. Outlet-Level Performance

For real-world applications, characterizing the overall bias of media outlets is often more valuable than analyzing individual articles. Our custom GDELT dataset enables this by aggregating article-level predictions across 100 articles per outlet for 13 major U.S. sources using mean score aggregation. Figure 2 visualizes the resulting distributions. The outlet ordering by mean predicted bias score largely corresponds with conventional understanding of media bias. Outlets with strong partisan leanings show tighter distributions with clear directional skew, while centrist outlets exhibit wider distributions spanning multiple bias regions. The outlet-level classification achieves 53.85% accuracy with a Macro-F1 of 55.57%. Notably, our method achieves perfect precision (100%) for Left and Right outlets, meaning no left-leaning outlet is ever classified as right-leaning or vice versa. All misclassifications occur at the Left-Center and Center-Right boundaries (Fig. 3), reflecting the inherently ambiguous distinction between centrist and mildly partisan sources. The absence of severe cross-spectrum errors indicates reliable high-level characterization suitable for media monitoring and reader awareness applications.

VI. CHALLENGES AND LIMITATIONS

Our unsupervised framework faces several inherent challenges. First, bias detection without labeled data relies entirely on LLMs' pre-trained knowledge, which may embed their own biases. The task is complicated by articles interweaving multiple events and perspectives across thousands of tokens,

TABLE V
PER-CLASS PERFORMANCE (PRECISION, RECALL, AND F1-SCORE), ACCURACY, AND MACRO-F1 ACROSS AD FONTES AND BALY DATASETS

	Method	Left			Center			Right			Acc (%)	Macro-F1
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1		
Ad Fontes	Llama 3.1 (8B)	63.30	91.60	74.87	78.30	51.60	62.21	88.02	78.68	83.09	73.96	73.39
	Qwen2.5 (7B)	72.27	82.10	76.87	69.91	68.30	69.09	83.41	73.97	78.41	74.79	74.79
	Mistral (7B)	69.35	86.20	76.86	74.30	63.90	68.71	85.60	76.78	80.95	75.63	75.51
	MADS Small	**71.89**	**82.10**	**76.66**	**70.86**	**71.50**	**71.18**	**86.44**	**73.37**	**79.37**	**75.66**	**75.74**
	Qwen3 (14B)	78.43	83.03	80.66	73.06	74.63	73.84	87.70	81.08	84.26	79.50	79.59
	GPT-OSS (20B)	71.33	84.74	77.46	80.23	62.97	70.56	81.17	84.46	82.78	77.22	76.93
	Mistral (22B)	68.42	90.48	77.92	79.29	63.31	70.40	89.28	80.63	84.73	77.84	77.68
	MADS	**76.61**	**81.20**	**78.84**	**71.31**	**75.99**	**73.57**	**88.09**	**77.48**	**82.44**	**78.15**	**78.29**
Baly	Llama 3.1 (8B)	43.34	75.03	54.94	48.43	25.47	33.38	67.74	48.51	56.54	49.98	48.29
	Qwen2.5 (7B)	49.27	57.60	53.11	45.81	49.39	47.53	64.06	47.85	54.79	51.69	51.81
	Mistral (7B)	48.45	61.64	54.26	47.85	46.53	47.18	64.75	48.29	55.32	52.27	52.25
	MADS Small	**49.91**	**56.54**	**53.01**	**45.68**	**54.80**	**49.82**	**66.50**	**44.11**	**53.04**	**51.87**	**51.96**
	Qwen3 (14B)	51.80	49.30	50.52	44.20	66.73	53.18	75.84	41.03	53.25	52.33	52.32
	GPT-OSS (20B)	47.50	61.07	53.43	51.61	34.18	41.13	60.48	63.42	61.92	52.94	52.16
	Mistral (22B)	48.88	63.88	55.39	48.01	51.58	49.73	73.07	45.09	55.76	53.54	53.63
	MADS	**51.90**	**46.78**	**49.21**	**44.79**	**69.48**	**54.47**	**74.59**	**41.03**	**52.94**	**52.40**	**52.20**

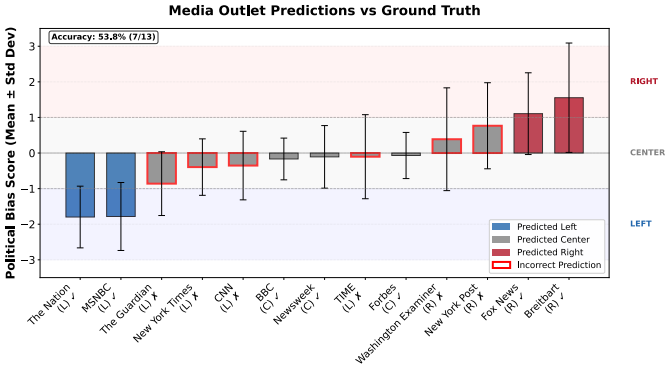

Fig. 3. Media outlet bias predictions ordered by mean score. Ground truth is shown in parentheses below outlet names, with checkmarks (✓) for correct and crosses (×) for incorrect predictions. Background shading indicates bias regions.

where legitimate arguments may exist for multiple bias interpretations. Second, model-intrinsic uncertainties introduce limitations: stochastic generation produces score variability, instruction-following capabilities vary across models, and hallucination risks persist during debates. Additionally, debate does not guarantee truth convergence; persuasive but incorrect arguments can prevail. The stagnation detection mechanism mitigates unproductive exchanges but cannot distinguish genuine deadlock from cases where additional rounds might yield convergence, and the confidence-based tiebreaker relies on token-level probabilities that may not always correlate with classification accuracy.

Third, the current evaluation is limited to English-language news from U.S. media outlets and uses models with strong English-language capabilities. Consequently, the reported results do not establish that MADS generalizes to other languages or political contexts. The availability and performance of suitable LLMs vary across languages, and political bias can be expressed through language-specific and culturally specific framing. Extending MADS will require multilingual datasets and evaluations with language-specific or multilingual LLMs.

VII. CONCLUSION

This work demonstrates that political bias detection can be transformed from a supervised learning problem requiring expensive annotations to an unsupervised reasoning task through structured multi-agent debate. MADS orchestrates deliberation among diverse LLMs, achieving 78.15% accuracy on Ad Fontes and performing better than supervised baselines on the Baly dataset, all without any training data. Our analysis reveals three critical findings. First, ensemble approaches provide robustness that single-model deployment cannot guarantee: the "best" model varies unpredictably across datasets, making MADS a reliable choice when the optimal individual model is unknown. Second, structured adversarial dialogue exposes subtle framing biases invisible to individual models, with the debate mechanism yielding a 12.28 percentage point accuracy improvement on the most ambiguous articles where all agents initially disagree. Third, the framework provides interpretable classifications through documented reasoning chains, essential for media literacy applications where users must understand and verify bias judgments. Looking ahead, promising directions include alternative debate formats, cross-lingual extension for global media monitoring, and integration with retrieval systems for real-time context. More broadly, the framework's success suggests that other subjective NLP tasks, such as stance detection and framing analysis may similarly benefit from multi-agent reasoning approaches.

ACKNOWLEDGMENT

This research is supported by National Cybersecurity Consortium (NCC) R&D Spearhead Grants (2024-1301), NSERC Discovery Grants (RGPIN-2024-04087), NSERC DND Supplements (DGDND-2024-04087), and Canada Research Chairs Program (CRC-2019-00041).

During the preparation of this work, the author(s) used ChatGPT/ClaudeAI to aid in refining the language and presentation of the work. After using the tool, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

REFERENCES

[1] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, p. e2023301118, 2021.

[2] M. L. Khan, I. Adjerid, and E. Karahanna, "Social media algorithmic curation and dissemination of political information," *Journal of Management Information Systems*, vol. 39, no. 3, pp. 782–815, 2022.

[3] W.-F. Chen, K. Al Khatib, H. Wachsmuth, and B. Stein, "Analyzing political bias and unfairness in news articles at different levels of granularity," in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, 2020, pp. 149–154.

[4] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang, "In plain sight: Media bias through the lens of factual reporting," 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6343–6349.

[5] L. Boxell, M. Gentzkow, and J. M. Shapiro, "Cross-country trends in affective polarization," *The Review of Economics and Statistics*, pp. 1–60, 2022.

[6] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[7] R. M. Entman, "Framing: Toward clarification of a fractured paradigm," *Journal of Communication*, vol. 43, no. 4, pp. 51–58, 1993.

[8] D. Chong and J. N. Druckman, "Framing theory," *Annual Review of Political Science*, vol. 10, pp. 103–126, 2007.

[9] C. Budak, S. Goel, and J. M. Rao, "Fair and balanced? quantifying media bias through crowdsourced content analysis," *Public Opinion Quarterly*, vol. 80, no. S1, pp. 250–271, 2016.

[10] D. Rozado, "The political biases of ChatGPT," *Social Sciences*, vol. 12, no. 3, p. 148, 2023.

[11] S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov, "From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 11 737–11 762. [Online]. Available: https://aclanthology.org/2023.acl-long.656/

[12] T. Groseclose and J. Milyo, "A measure of media bias," *The Quarterly Journal of Economics*, vol. 120, no. 4, pp. 1191–1237, 2005.

[13] M. Gentzkow and J. M. Shapiro, "What drives media slant? evidence from U.S. daily newspapers," *Econometrica*, vol. 78, no. 1, pp. 35–71, 2010.

[14] R. Baly, G. Da San Martino, J. Glass, and P. Nakov, "We can detect your bias: Predicting the political ideology of news articles," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 4982–4991. [Online]. Available: https://aclanthology.org/2020.emnlp-main.404/

[15] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: An interdisciplinary literature review," *International Journal on Digital Libraries*, vol. 20, pp. 391–415, 2019.

[16] Y. Liu, S. Cromp, H. Yu, I. Lee, Y. Luo, S. Ezzini, F. Pastore, S. Wang, and L. Moonen, "Neus: Neutral multi-news summarization for mitigating framing bias," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3131–3148.

[17] T. Spinde, C. Kreuter, W. Gaissmaier, F. Hamborg, B. Gipp, and H. Giese, "Do you think it's biased? how to ask for the perception of media bias," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021, pp. 61–69.

[18] I. Maab, E. Marrese-Taylor, S. Padó, and Y. Matsuo, "Media bias detection across families of language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 4083–4098.

[19] D. Azizov, Z. M. Mujahid, H. AlQuabeh, P. Nakov, and S. Liang, "SAFARI: Cross-lingual bias and factuality detection in news media and news articles," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 12 217–12 231.

[20] F. Trhlík and P. Stenetorp, "Quantifying generative media bias with a corpus of real-world and generated news articles," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 4420–4445.

[21] S. Raza, M. Rahman, and S. Ghuge, "Dataset annotation and model building for identifying biases in news narratives," *Text2Story Workshop at ECIR*, pp. 5–15, 2024.

[22] R. Hernandes and G. Corsi, "Llms left, right, and center: Assessing gpt's capabilities to label political bias from web domains," 2024. [Online]. Available: https://arxiv.org/abs/2407.14344

[23] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023.

[24] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi, "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint arXiv:2305.19118*, 2023.

[25] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Proceedings of NeurIPS 2023*, 2024.

[26] Z. Wang, H. Yin, and Y. Song, "Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?" in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024, pp. 6106–6131.

[27] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[28] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, and J. Liu, "Autogen: Enabling next-gen llm applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[29] A. Taubenfeld, Y. Dover, R. Reichart, and A. Goldstein, "Systematic biases in LLM simulations of debates," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 251–267. [Online]. Available: https://aclanthology.org/2024.emnlp-main.16/

[30] A. Bandaru, F. Bindley, T. Bluth, N. Chavda, B. Chen, and E. Law, "Revealing political bias in llms through structured multi-agent debate," 2025. [Online]. Available: https://arxiv.org/abs/2506.11825

[31] Y. Bang, D. Chen, N. Lee, and P. Fung, "Measuring political bias in large language models: What is said and how it is said," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 11 142–11 159.

[32] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 29 971–30 004.

[33] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 2, pp. 1–79, 2024.

[34] R. Lin *et al.*, "Investigating bias in LLM-based bias detection: Disparities between LLMs and human perception," *arXiv preprint arXiv:2403.14896*, 2024.
[35] R. Liu, C. Jia, J. Wei, G. Xu, and S. Vosoughi, "Mitigating political bias in language models through reinforced calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 857–14 866.
[36] A. Borah and R. Mihalcea, "Towards implicit bias detection and mitigation in multi-agent LLM interactions," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 9306–9326.
[37] D. Ganguli *et al.*, "The capacity for moral self-correction in large language models," *arXiv preprint arXiv:2302.07459*, 2023.
[38] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.
[39] S. F. Tekin, F. Ilhan, T. Huang, S. Hu, and L. Liu, "LLM-TOPLA: Efficient LLM ensemble by maximising diversity," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 11 951–11 966.
[40] J. Oh, M. Jeong, J. Ko, and S.-Y. Yun, "Understanding bias reinforcement in llm agents debate," 2025. [Online]. Available: https://arxiv.org/abs/2503.16814
[41] V. Otero, J. Von Wald, E. Fox-Ramirez, and B. Heldebrandt, "Methodology white paper," https://adfontesmedia.com/methodology-white-paper/, 2023, accessed: 2024.
[42] AllSides, "Allsides media bias ratings," https://www.allsides.com/media-bias/ratings, accessed: 2024.
[43] M. Haroon, M. Wojcieszak, and A. Chhabra, ""whose side are you on?" estimating ideology of political and news content using large language models and few-shot demonstration selection," 2025. [Online]. Available: https://arxiv.org/abs/2503.20797

SUPPLEMENTARY MATERIAL

A. Algorithmic Procedures

This section presents the formal algorithms for our multi-agent debate framework. Algorithm 1 describes the overall inference pipeline, while Algorithm 2 details the pairwise debate procedure used in all debate scenarios.

Algorithm 1 Overall Inference Pipeline

Require: Article x, agents $\{a_1, a_2, a_3\}$, round cap $R_{\max}$

Phase I: For each i, query a_i to obtain $(s_i^{(0)}, r_i^{(0)})$; set $d_i^{(0)} = \mathrm{dir}(s_i^{(0)})$.
if $d_1^{(0)} = d_2^{(0)} = d_3^{(0)}$ (*unanimous*) **then**
  return $\hat{s} = \frac{1}{3}\sum_i s_i^{(0)}$ {Early exit}
end if
if exactly two directions coincide (*majority 2v1*) **then**
  Select higher-confidence agent from majority pair as *rep*; let *dis* be the dissenter.
  $(\text{winner}, \mathcal{H}) \leftarrow \textsc{Debate}(\text{rep}, \text{dis}, R_{\max}, \mathcal{H} = [])$
  if all three agents share the same direction **then**
    return $\hat{s} = \frac{1}{3}\sum_i s_i$ {Panel unanimity}
  else
    return $(s_{\text{winner}}, d_{\text{winner}})$
  end if
end if
{*All different (1v1v1)*}
Compute normalized embeddings e_i of $r_i^{(0)}$. Set $(p, q) = \arg\min_{i \neq j} e_i^\top e_j$; m is the remaining index.
$(\text{rep}, \mathcal{H}) \leftarrow \textsc{Debate}(a_p, a_q, R_{\max}, \mathcal{H} = [])$ {Stage 1}
if $d_1 = d_2 = d_3$ **then**
  return $\hat{s} = \frac{1}{3}\sum_i s_i$ {Panel unanimity}
end if
$(\text{winner}, \mathcal{H}) \leftarrow \textsc{Debate}(\text{rep}, a_m, R_{\max}, \mathcal{H})$ {Stage 2}
if $d_1 = d_2 = d_3$ **then**
  return $\hat{s} = \frac{1}{3}\sum_i s_i$ {Panel unanimity}
else
  return $(s_{\text{winner}}, d_{\text{winner}})$
end if

B. Complete Prompt Templates

This supplement provides the complete prompt templates used in our multi-agent ensemble system for political bias detection. The system employs three distinct types of prompts:

1) **Individual Analysis** prompts for initial bias assessment,
2) **Discussion Challenge** prompts for generating counter-arguments, and
3) **Discussion Response** prompts for responding to challenges.

B.1 Individual Analysis Prompt Template: Each model uses the following prompt to perform initial political bias assessment. The prompt follows a two-step approach: article understanding followed by bias analysis.

Listing 1. Individual Analysis Prompt Template

```
System Instruction:
You are an expert media analyst specializing in
    political bias detection.
Use this scoring scale:
- -3: Strong support for Democrats or criticism of
    Republicans
- -2: Moderate Democratic support
- -1: Slight Democratic lean
-  0: Neutral or balanced
-  1: Slight Republican lean
-  2: Moderate Republican support
-  3: Strong support for Republicans or criticism of
     Democrats
Note: All Democratic articles should score at most
    -1,
whereas all Republican articles should score at
    least +1.
User Message:
First, read the article and write a paragraph
demonstrating your understanding of its content.
####################
{ARTICLE_CONTENT}
####################
Step 1: Article Understanding
Summarize the article's main topic, events, and
    subjects.
Step 2: Political Bias Analysis
Analyze political leaning by considering:
- Political figures, parties, or policies mentioned
- Whether portrayed favorably, neutrally, or
    critically
- Democratic/Republican values promoted or
    criticized
- Multiple perspectives vs. one-sided framing
- Language tone, framing choices, and sources
- Overall implicit/explicit political direction
IMPORTANT: Write a distinct reasoning paragraph
separate from the article understanding.
Your output must be valid JSON:
{
  "article_understanding": "[Summary paragraph]",
```

Algorithm 2 Debate$(a_i, a_j, R_{\max}, \mathcal{H})$

Set $t \leftarrow 0$; let current (s_i, s_j) and (d_i, d_j) be initial values.
while $t < R_{\max}$ **do**
  $t \leftarrow t + 1$; assign challenger $\leftarrow a_i$ if t odd else a_j; target $\leftarrow$ the other.
  Query challenger with *generate_discussion_challenge*; append to $\mathcal{H}$.
  Query target with *generate_discussion_response*; update scores and reasoning; clamp to $[-3, 3]$.
  Recompute (d_i, d_j).
  if $d_i = d_j$ **then**
    break {Pair consensus}
  end if
  Embed current arguments; compute cosine similarity with previous round.
  if cosine similarity > 0.9 **then**
    break {Stagnation detected}
  end if
end while
if $d_i = d_j$ **then**
  winner $\leftarrow$ agent (same direction) with lower entropy H
else
  winner $\leftarrow$ agent (different direction) with lower entropy H {Confidence tiebreak}
end if
return (winner, $\mathcal{H}$)

```
  "reason": "[Bias reasoning and justification]",
  "lean": [Integer from -3 to 3]
}
Return only the JSON object.
```

B.2 Discussion Challenge Prompt Template: When models disagree, the challenge prompt enables one model to question another's analysis by acknowledging their perspective while presenting counterevidence.

Listing 2. Discussion Challenge Prompt Template

```
System Instruction:
You are an expert media analyst engaging in
    collaborative bias analysis.
Always respond with valid JSON only.
Scoring scale: -3 (Strong Democrat) ... +3 (Strong
    Republican)
User Message:
ARTICLE CONTENT:
{ARTICLE_CONTENT}
YOUR ANALYSIS:
Score: {OWN_SCORE} ({OWN_DIRECTION})
Reasoning: {OWN_REASONING}
TARGET ANALYSIS:
Score: {TARGET_SCORE} ({TARGET_DIRECTION})
Reasoning: {TARGET_REASONING}
{CONVERSATION_HISTORY}
Consider the target's interpretation.
Generate a structured challenge engaging with their
    perspective.
Output ONLY this JSON format:
{
  "understanding": "[Acknowledgment of target's
      valid points]",
  "challenge": "[Evidence-based counterargument
      citing passages]",
  "adjusted_lean": <score from -3 to 3 after
      reconsideration>
}
```

B.3 Discussion Response Prompt Template: The response prompt allows the challenged model to defend its position or acknowledge valid points and adjust its bias assessment.

Listing 3. Discussion Response Prompt Template

```
System Instruction:
You are an expert media analyst engaging in
    collaborative bias analysis.
Always respond with valid JSON only.
Scoring scale: -3 (Strong Democrat) ... +3 (Strong
    Republican)
User Message:
ARTICLE CONTENT:
{ARTICLE_CONTENT}
YOUR INITIAL ANALYSIS:
Score: {OWN_SCORE} ({OWN_DIRECTION})
Reasoning: {OWN_REASONING}
CHALLENGER'S ANALYSIS:
Score: {CHALLENGER_SCORE} ({CHALLENGER_DIRECTION})
Reasoning: {CHALLENGER_REASONING}
{CONVERSATION_HISTORY}
CHALLENGER'S LATEST ARGUMENT:
{CHALLENGE_TEXT}
Respond to the challenge. You may:
1. Defend your original score
2. Acknowledge valid points and adjust your score
Provide your response as JSON:
{
  "response_text": "[Your detailed response]",
  "score_changed": true/false,
  "new_score": [Updated or original score],
  "change_reasoning": "[Justification for change or
      no-change]"
}
```

B.4 Prompt Design Rationale: The prompt design incorporates several principles for effective multi-agent collaboration:

- **Structured Output:** JSON formatting ensures consistent parsing.
- **Perspective Acknowledgment:** Challenges require acknowledging the target's perspective before countering.
- **Evidence-Based Reasoning:** Prompts emphasize citing textual evidence.
- **Score Adjustment:** Models may revise scores based on peer analysis.
- **Conversation History:** Preserves context across multiple discussion rounds.
- **Clear Scoring Guidelines:** Consistent scale (-3 to $+3$) across all prompts.