Measuring and Mitigating Media Outlet Name Bias in Large Language Models

Seong-Jin Park[1] **and Kang-Min Kim**[1,2*]
[1]Department of Artificial Intelligence [2]Department of Data Science
The Catholic University of Korea, Bucheon, Republic of Korea
{sjpark,kangmin89}@catholic.ac.kr

Abstract

Large language models (LLMs) have achieved remarkable performance across diverse natural language processing tasks, but concerns persist regarding their potential political biases. While prior research has extensively explored political biases in LLMs' text generation and perception, limited attention has been devoted to biases associated with media outlet names. In this study, we systematically investigate the presence of media outlet name biases in LLMs and evaluate their impact on downstream tasks, such as political bias prediction and news summarization. Our findings demonstrate that LLMs consistently exhibit biases toward the known political leanings of media outlets, with variations across model families and scales. We propose a novel metric to quantify media outlet name biases in LLMs and leverage this metric to develop an automated prompt optimization framework. Our framework effectively mitigates media outlet name biases, offering a scalable approach to enhancing the fairness of LLMs in news-related applications.

GitHub Repository

1 Introduction

Extensive research has revealed the political biases inherent in large language models (LLMs) (Bang et al., 2024; Rozado, 2023; Lunardi et al., 2024; Fang et al., 2024). For instance, Liu et al. (2022) reported a liberal bias in the outputs of GPT-2, Santurkar et al. (2023) demonstrated that fine-tuning with human feedback tends to reinforce consistent left-leaning tendencies across various models, and Yang et al. (2024c) found that ChatGPT exhibits systematic left-leaning political bias in their responses. These findings suggest that LLMs may amplify specific political viewpoints, potentially shaping user perceptions (Messer, 2025), and may even reinforce political polarization in user interactions (Linegar et al., 2023). Nevertheless, an important dimension remains underexplored: *Do LLMs exhibit political biases toward the names of media outlets themselves?*

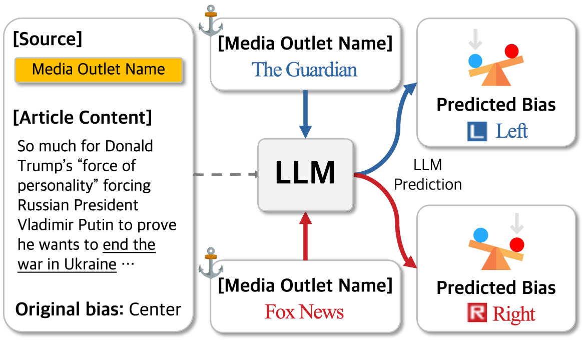


Figure 1: Problem definition. Media outlet names induce anchoring effects in LLM outputs. The example article obtained from CNN (Collinson, 2025).

Humans recognize differences in political perspectives across various media outlets and often develop biases toward them. The tendency of media outlets to exhibit political biases is rooted in the human inclination to prefer information that aligns with pre-existing beliefs (Taber and Lodge, 2006; Nickerson, 1998), which in turn incentivizes outlets to adopt and emphasize particular ideological stances (Baron, 2006; Bakshy et al., 2015; Garrett, 2009; Gentzkow and Shapiro, 2010). This phenomenon is widely recognized by the public, and platforms such as AllSides[1] systematically aggregate public and expert analyses to classify these biases (e.g., categorizing *Fox News* as right-leaning and *The Guardian* as left-leaning[2]).

In political psychology, it is well documented that people's interpretation of information is influenced by source cues where identical news content elicits different trust and bias perceptions depending on the outlet label (Iyengar and Hahn, 2009). This relates to the concept of media framing, where

[*]Corresponding author.

[1] https://www.allsides.com/media-bias

[2] Examples based on the AllSides Media Bias Chart

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29778–29797
November 4-9, 2025 ©2025 Association for Computational Linguistics

presentation and context (e.g., the source's identity) can alter an audience's judgment of the information's meaning and slant (Entman, 1993). Since LLMs are known to absorb the biases present in their training data (Bender et al., 2021), it is plausible that they may also internalize public biases associated with media outlet names.

Political biases that LLMs hold toward media outlet names can give rise to anchoring effects. Anchoring refers to the cognitive phenomenon where individuals' estimations and decisions are heavily influenced by initially presented information (i.e., the anchor) (Tversky and Kahneman, 1974). Recent studies have shown that LLMs, like humans, are similarly susceptible to anchoring effects across various domains (e.g., finance and code generation) (Nguyen, 2024; Jones and Steinhardt, 2022; Lou and Sun, 2024). When a media outlet name appears within the context provided to an LLM, it may serve as an anchor, subtly influencing the model's generation process.

This issue becomes particularly critical given the increasing deployment of LLMs in news media generation and analysis (Ding et al., 2023; Brigham et al., 2024; Petridis et al., 2023; Gao et al., 2024), where impartiality and factual integrity are paramount. In particular, the strong document summarization capabilities of LLMs (Brown et al., 2020; Zhang et al., 2023; Laban et al., 2023) have accelerated the development of LLM-based news summarization systems (Zhang et al., 2024; Tam et al., 2023). Reflecting this trend, Bloomberg recently announced the integration of generative AI to assist in news content summarization within its widely used Bloomberg Terminal (Bloomberg, 2024). However, anchoring effects triggered by media outlet names in such applications could distort information delivery, potentially leading to significant socio-economic consequences (McCarthy and Dolfsma, 2014; Druckman and Parkin, 2005). Understanding and mitigating these risks is therefore essential for the responsible deployment of LLMs in news-related contexts.

In this paper, we conduct controlled experiments to measure media outlet name biases inherent in LLMs and propose a novel metric to quantify these biases. Specifically, we present news articles to LLMs while varying the attributed media outlet name and analyze whether the predicted political bias of the article shifts based on the source (Figure 1). To quantify this effect, we introduce a unified metric that captures two dimensions of political bias: magnitude and direction. Our approach confirms that media outlet names serve as anchors, systematically shaping the model's perception of an article's political stance.

In addition, we explore anchoring effects in summarization by prompting LLMs to generate summaries while varying the indicated media outlet and performing linguistic analyses on the outputs. Our findings reveal that media outlet names influence the summarization process, subtly altering the sentiment of the generated summaries. Building on these insights, we utilize our proposed metric as a reward function within an automated prompt optimization framework for LLMs, demonstrating its scalability and effectiveness in reducing media outlet name biases.

Our findings reveal several key insights: **(1)** LLMs consistently exhibit political bias to media outlet names, with variation across model families; **(2)** larger models and alignment-tuned (e.g., instruction tuning (IT) (Wei et al., 2022), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022)) models tend to show stronger biases; **(3)** LLMs also demonstrate similar levels of bias toward fictitious outlet names, reflecting sensitivity to political connotations; **(4)** articles with near-neutral inherent bias are more vulnerable to media-induced prediction shifts; **(5)** in summarization, media outlet names influence the sentiment of the article; **(6)** finally, media outlet name bias can be effectively mitigated through automated prompt optimization.

In summary, our contributions are as follows:

- We systematically evaluate media outlet name biases across diverse LLMs, providing key insights into the conditions and extent of biases.
- We propose a novel two-dimensional metric and framework to quantify media outlet name biases in LLMs, capturing both magnitude and direction.
- We demonstrate that our proposed metric serves as an effective signal for an automated prompt optimization framework, significantly mitigating media outlet name biases in article bias prediction tasks.

2 Measuring Media Outlet Name Bias in LLMs

To analyze media outlet name bias in LLMs, we conduct controlled experiments using real-world news articles. Each LLM is evaluated on two distinct tasks: political bias prediction and summarization, each performed under varying media outlet name conditions. In the news article political bias prediction task, we quantify changes in predicted bias as a function of the outlet's political leaning, using a two-dimensional metric framework. In the summarization task, we assess whether the attributed source name induces systematic changes in the semantics of the generated summaries.

2.1 Political Bias Prediction Shift

News Article Political Bias Prediction Let A denote the set of news articles and $a \in A$ a single article. Let $\mathcal{O}$ be the set of media outlet names. Given a prompt p and an attributed outlet name $o \in \mathcal{O}$, the LLM f_θ outputs a hard-label for the news article political bias prediction:

$$f_\theta(p, o, a) \in \{\text{left, center, right}\}, \tag{1}$$

which is then mapped to scalar values in $[-1,\ 0,\ 1]$ for subsequent computation.

Measuring Prediction Shift For each article $a \in A$, we query the model once without a media outlet name (i.e., the baseline) and once with each media outlet $o \in \mathcal{O}$. We define the media outlet-induced prediction shift for article a and outlet o as:

$$d(o, a) = f_\theta(p, o, a) - f_\theta(p, \varnothing, a), \tag{2}$$

where $f_\theta(p, \varnothing, a)$ is the model's political bias prediction without outlet attribution. The *overall article political bias prediction shift for outlet o* is then computed as the average over all articles:

$$S(o) = \frac{1}{|A|} \sum_{a \in A} d(o, a). \tag{3}$$

Let $G = \{\text{left, center, right}\}$ denote the set of media outlet bias classes, and let $\mathcal{O}_g \subset \mathcal{O}$ be the subset of media outlets annotated with political bias class $g \in G$. The *average prediction shift for each bias class g* is computed as:

$$S(g) = \frac{1}{|\mathcal{O}_g|} \sum_{o \in \mathcal{O}_g} S(o). \tag{4}$$

Visualizing $S(g)$ across the political spectrum provides an intuitive view of model bias patterns. A perfectly unbiased model would produce a flat $S(g)$ curve, while a model highly sensitive to media outlet names would yield a curve with a steep slope. To quantify this behavior more systematically, we introduce the source-induced prediction shift (SIPS) metric, which captures both the magnitude and the directional alignment of bias effects induced by outlet attribution.

2.2 The SIPS Metric

Motivation Political bias induced by media outlet names can be decomposed into two components: (1) *Bias Magnitude* — the extent to which political bias predictions shift, regardless of direction, and (2) *Directional Consistency* — whether the direction of the shift aligns with the known political orientation of the outlet. These two aspects are partially independent and should be evaluated separately. To capture both dimensions in a unified manner, we define the SIPS metric.

Absolute Sensitivity We define absolute sensitivity (AS) to measure the overall magnitude of bias shifts for a given article a. Conceptually, AS computes the average magnitude of prediction shifts across all bias classes and scales the value to the $[0, 1]$ range. Formally, AS is defined as:

$$\text{AS}(a) = \frac{1}{Z} \sum_{g \in G} \frac{1}{|\mathcal{O}_g|} \sum_{o \in \mathcal{O}_g} |d(o, a)|\,, \tag{5}$$

where scaling factor $Z = \max(d(\cdot) \cdot |G|)$.

Agreement Coherence Agreement coherence (AC) measures the extent to which the direction of bias shifts aligns with the annotated political orientation g_o of each media outlet o. For media outlets labeled as left or right, we assess whether the sign of the prediction shift corresponds to the media's ideological polarity. For center-labeled outlets, we evaluate whether the magnitude of the shift remains small, reflecting the expected neutrality of such sources.

Formally, AC is defined as:

$$\text{AC}(a) = \frac{1}{|G|} \sum_{g \in G} \mathbf{1}_g\Big(\frac{1}{|\mathcal{O}_g|} \sum_{o \in \mathcal{O}_g} d(o, a)\Big), \tag{6}$$

and $\mathbf{1}_g(\cdot)$ is an indicator function defined as:

$$\mathbf{1}_g(\cdot) = \begin{cases} \mathbf{1}\,[\text{sign}(\cdot) = \text{sign}(g)] & \text{if } g \neq 0, \\ \mathbf{1}\,[|\cdot| \leq \delta] & \text{if } g = 0, \end{cases} \tag{7}$$

where δ is a small threshold (e.g., δ = 0.3[3]) that allows minimal shifts for center-labeled outlets.

[3] Set to 0.3 to ensure balanced spacing across classes in the prediction range of $[-1, 1]$.

SIPS We define the SIPS score as the root mean square of the averaged AS and AC across all articles $a \in A$. Let $\overline{\text{AS}} = \frac{1}{|A|}\sum_{a \in A} \text{AS}(a)$, which measures the average magnitude of bias shifts induced by media outlet attributions, and $\overline{\text{AC}} = \frac{1}{|A|}\sum_{a \in A} \text{AC}(a)$, which quantifies the average directional accuracy of bias predictions. The SIPS score is then:

$$\text{SIPS} = \sqrt{\frac{\overline{\text{AS}}^2 + \overline{\text{AC}}^2}{2}}, \qquad (8)$$

which captures both the magnitude and directional correctness of bias shifts across the article set.

Interpretation The SIPS score ranges from 0 to 1 and can be interpreted as follows:

- **SIPS** ≈ 1: Indicates strong prediction shifts that are either perfectly aligned with the known political orientations of the media outlets (AC↑), or reflect large shifts in magnitude regardless of direction (AS↑ & AC↓).
- **SIPS** ≈ 0: Suggests either little to no response to outlet names (AS↓), or highly inconsistent directional shifts (AC↓).

Because SIPS captures both the magnitude and the directional consistency of media outlet name-induced bias shifts, we recommend reporting AS, AC, and SIPS together for a comprehensive assessment of model behavior. Illustrative examples of AS, AC, and SIPS values under various prediction scenarios are provided in Appendix B.

2.3 Sentiment Shifts in Article Summarization

Summarization Method To examine whether media outlet names induce anchoring effects in news article summarization, we adopt a minimal-prompting strategy similar to that of Zhang et al. (2024). We instruct LLMs to generate summaries with minimal guidance. As in the political bias prediction task, we vary the attributed media outlet name across three political bias categories, while keeping the article content constant.

Analyzing Anchoring Effects To evaluate the impact of media outlet attribution on summaries, we follow the approach of Bang et al. (2024). We extract named entities using a named entity recognition (NER) model and analyze their sentiment to track changes in the proportion of positive, negative, and neutral expressions. Details on the NER and sentiment analysis methods are provided in Appendix C.1.

3 Experimental Setup

3.1 Representative Media Outlet Selection

We select representative media outlets based on the AllSides Media Bias Chart[4], which categorizes sources into three political bias classes: left, center, and right. To assess outlet popularity, we use SimilarWeb traffic data[5] from January 2025. Among the top 50 outlets by traffic, we identify the top three outlets per bias with available AllSides bias annotations. This procedure resulted in a final set of nine media outlets used in our experiments.

The selected media outlets are: **Left**: *Associated Press*, *The Guardian*, and *HuffPost*; **Center**: *BBC News*, *Forbes*, and *CNBC*; **Right**: *Fox News Digital*, *Daily Mail*, and *Breitbart News*.

3.2 LLM Selection

To analyze biases in widely used LLMs, we select the five most downloaded open-source LLM families on Hugging Face[6] and use their latest alignment-tuned versions. For proprietary LLM, we include the latest GPT-4.1 series. LLM selection process details are provided in Appendix C.2.

Selected models include Llama-3.3$_{\text{70B-Instruct}}$ (Grattafiori et al., 2024), Qwen-2.5$_{\text{72B-Instruct}}$ (Yang et al., 2024a), Phi-4$_{\text{14B}}$ (Abdin et al., 2024), Mistral-Small$_{\text{24B-Instruct}}$ (AI, 2024), Gemma-2$_{\text{27B-IT}}$ (Team et al., 2024), and GPT-4.1. To examine model size effects, we include scaled variants of Qwen-2.5$_{\text{7B-72B}}$, Llama-3$_{\text{8B, 70B}}$ and Llama-3.1$_{\text{8B, 70B}}$, and GPT-4.1$_{\text{mini, nano}}$. To assess impact of alignment tuning, we use base versions (i.e., pretrained only) of Qwen-2.5, Llama-3.1, and Mistral-Small. For reasoning-specialized models, we include QwQ$_{\text{32B}}$ (Team, 2025).

3.3 Political News Dataset Selection

We use two datasets: the AllSides dataset (Baly et al., 2020) and the Hyperpartisan News Detection dataset (Kiesel et al., 2019). From the AllSides dataset, we randomly sample 1,500 articles for each of the three bias classes (left, center, and right), resulting in a total of 4,500 articles. For summarization experiments and evaluations involving reasoning-specialized LLMs, we use a smaller subset of 450 articles (150 per class). From the Hyperpartisan dataset, we include all 1,273 articles

[4] We use Version 10.1, the latest available version as of January 2025

[5] https://www.similarweb.com/

[6] https://huggingface.co/

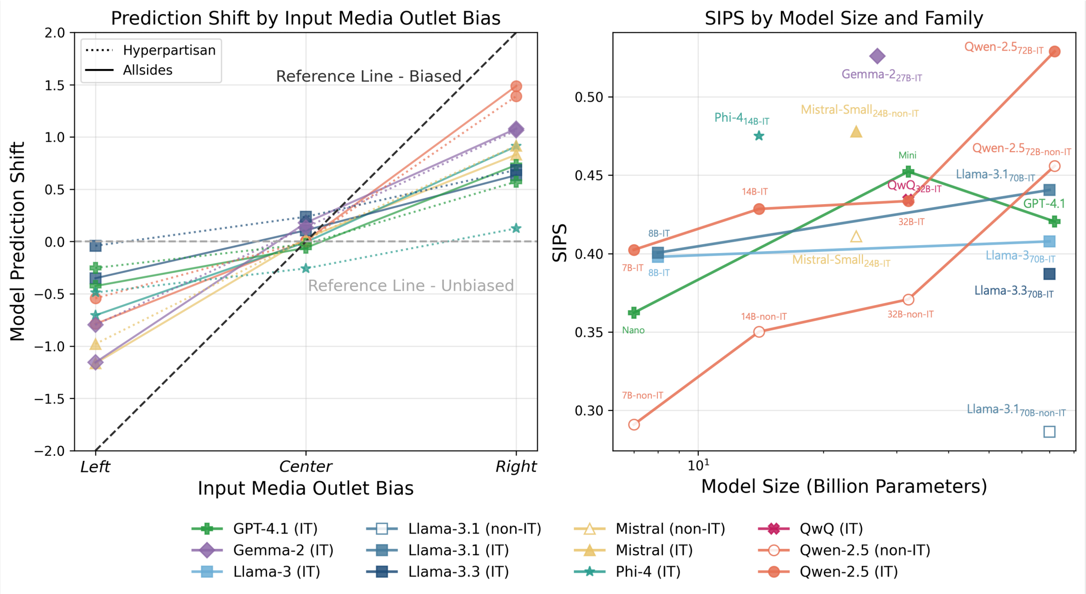


Figure 2: SIPS scores of major models and response variation according to the bias class of the input media outlet name. *(Left)* Visualization of the degree of response shift (i.e., $S(g)$ curve). *(Right)* SIPS scores of representative model families and series, sorted by model size. IT refers to alignment-tuned variants.

with available bias annotations. All datasets we used were free to use for research.

3.4 Implementation Details

Prompts for LLMs For article bias prediction, we design prompts following the structure proposed by Maab et al. (2024), including role assignment, task description, bias target explanation, and output guidelines. For summarization tasks, we directly adopt the prompt template utilized in Zhang et al. (2024). The actual prompt templates are provided in Appendix C.3.

LLM Generation Configuration For article bias prediction, we configure open-source LLMs using the multiple-choice method from Robinson and Wingate (2023), while QwQ$_{32B}$ used its recommended sampling settings. For summarization, we follow the setup of Wu et al. (2021), setting the decoding temperature to 0.3 to encourage factual consistency and focused generation. Detailed generation settings are provided in Appendix C.4.

4 Results and Analysis

4.1 News Article Political Bias Prediction

LLM-wise Analysis All six LLMs evaluated exhibit significant media outlet name biases in a directionally coherent manner across all datasets (Figure 2 *Left*). Qwen-2.5$_{72B\text{-}Instruct}$, Gemma-2$_{27B\text{-}IT}$, Mistral-Small$_{24B\text{-}Instruct}$, and Phi-4$_{14B}$ show clear prediction shifts, with Mistral and Gemma more sensitive to left-leaning sources and Qwen-2.5 to right-leaning ones. In contrast, GPT-4.1 exhibits modest bias magnitude but clear direction.

Model	AllSides			Hyperpartisan		
	SIPS	AS	AC	SIPS	AS	AC
Qwen-2.5$_{72B\text{-}Instruct}$	**0.529**	0.439	**0.605**	0.465	0.376	**0.540**
Mistral-Small$_{24B\text{-}Instruct}$	0.478	0.426	0.525	**0.466**	**0.396**	0.527
Phi-4$_{14B}$	0.475	<u>0.468</u>	0.482	0.362	0.339	0.383
Llama-3.3$_{70B\text{-}Instruct}$	0.387	0.358	0.414	0.370	0.337	0.400
Gemma-2$_{27B\text{-}IT}$	<u>0.510</u>	**0.479**	<u>0.540</u>	<u>0.466</u>	<u>0.385</u>	<u>0.535</u>
GPT-4.1	0.421	0.266	0.532	0.356	0.189	0.467

Table 1: Calculated SIPS, AS, and AC scores. Highest scores are in **bold**; second-highest are <u>underlined</u>.

We then apply the SIPS metric for further analysis. As shown in Table 1, Qwen-2.5$_{72B\text{-}Instruct}$ records the highest SIPS on the AllSides dataset, indicating strong bias, while Llama-3.3$_{70B\text{-}Instruct}$ shows the lowest, suggesting milder bias. Most models achieve high AC scores, reflecting strong alignment with human-annotated polarity directions. On the Hyperpartisan dataset, AC remains stable while AS slightly decreases, consistent with the flatter dotted lines in Figure 2 (*Left*).

Further analysis reveals that SIPS increases with model size and alignment tuning (Figure 2, *Right*). SIPS scores scale nearly linearly with parameter count in Qwen-2.5 and Llama models, though GPT-4.1$_{mini}$ unexpectedly shows the highest SIPS within its family. IT models consistently exhibit higher SIPS than their base versions, supporting

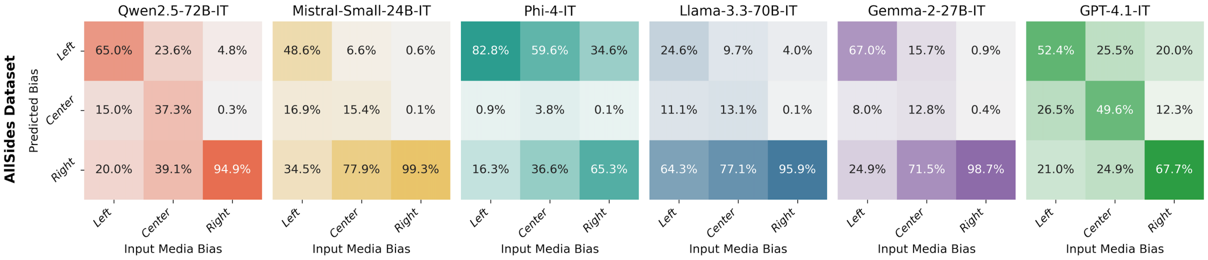


Figure 3: Predicted bias heatmap by input media bias across models. Predicted bias is calculated by averaging model outputs, with values within (–0.3, 0.3) classified as center. IT refers to alignment-tuned variants.

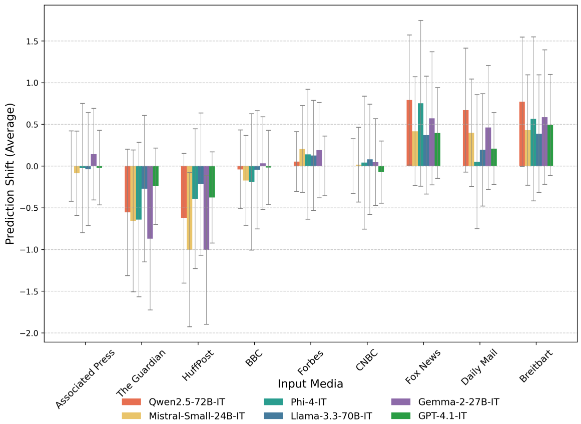


Figure 4: Average prediction shift by media outlet name, sorted from left- to right-leaning. Gray lines indicate standard deviations.

Model	ΔG_{left}	ΔG_{right}	ΔF_{left}	ΔF_{right}
Qwen-2.5$_{\text{72B-Instruct}}$	-0.041	0.356	-0.280	0.445
Mistral-Small$_{\text{24B-Instruct}}$	-0.238	0.297	-0.334	0.267
Phi-4$_{\text{14B}}$	-0.210	-0.018	-0.388	0.121
Llama-3.3$_{\text{70B-Instruct}}$	-0.045	0.199	-0.033	0.192
Gemma-2$_{\text{27B-IT}}$	-0.043	0.352	-0.261	0.365

Table 2: Average prediction shift for fictitious media outlet names. ΔG_{left} and ΔG_{right} indicate average prediction shifts for generated left- and right-biased media names, while ΔF_{left} and ΔF_{right} refer to those for formulated names.

prior findings that alignment amplifies bias (Itzhak et al., 2024). In addition, the reasoning-specialized QwQ$_{\text{32B}}$ shows no significant difference.

Media Outlet-level Analysis LLM predictions tend to reflect the political orientation of the input media outlet (Figure 3). Qwen-2.5$_{\text{72B-Instruct}}$, Mistral-Small$_{\text{24B-Instruct}}$, Llama-3.3$_{\text{70B-Instruct}}$, and Gemma-2$_{\text{27B-IT}}$ show strong sensitivity to right-leaning sources. GPT-4.1 is sensitive across all bias classes, consistent with its high AC despite a low SIPS. At the outlet level (Figure 4), the *Associated Press* has notably little effect on model predictions. This may be due to its recent reclassification from neutral in 2022[7], which is likely underrepresented in LLM training data. These trends are consistent across both datasets (Appendix D.1).

To examine whether LLMs react to the political connotations of media names, we introduce fictitious left- and right-biased outlets. We use two methods: a formulated approach combining parts of real outlets (e.g., *Millennial Times*) and a generated approach prompting GPT-4.1 to create politically biased fictional names. We manually verified all names to ensure they do not exist. Across both methods, all models exhibited clear bias (Table 2), suggesting that LLMs react to the implied ideological cues in media names. A full list of fictitious media names is provided in Appendix D.2.

Article-level Analysis To identify the most influential article-level factor, we analyze affected cases in Figure 5. A key insight is that the article's inherent bias plays a central role. In the AllSides dataset, center-labeled articles show a higher proportion of affected cases than others, though this signal may be diluted due to outlet-level labeling. In contrast, the Hyperpartisan dataset with article-level annotations reveals a much higher affected rate for non-hyperpartisan articles. This aligns with the intuition that articles lacking strong internal bias make model predictions more uncertain and thus more susceptible to external cues like outlet names.

Importantly, this effect is not merely due to neutral articles being predicted as center-biased. For instance, in the Hyperpartisan dataset, models such as Llama-3.3$_{\text{70B-Instruct}}$, Mistral-Small$_{\text{24B-Instruct}}$, and Gemma-2$_{\text{27B-IT}}$ showed minimal differences (i.e., less than 0.1) in average predictions between hyperpartisan and non-hyperpartisan articles when no media outlet name was given, yet the affected

[7] https://www.allsides.com/news-source/associated-press-media-bias

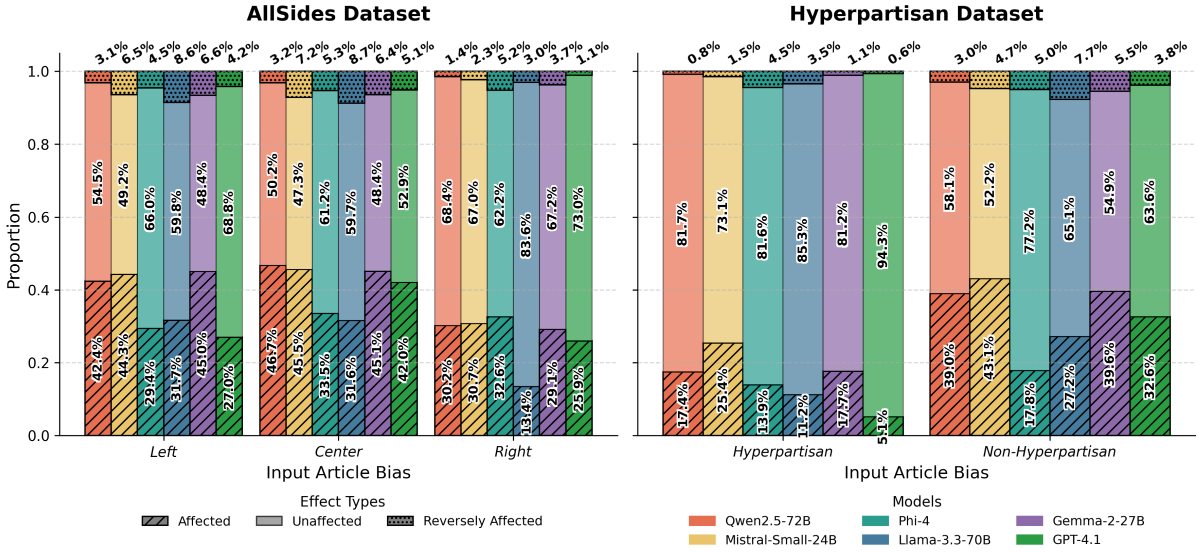


Figure 5: Proportions of prediction outcomes by article bias. Affected indicates cases where the prediction shifted in the same direction as the media outlet name; unaffected refers to cases with no shift; and reversely affected indicates shifts in the opposite direction of the media outlet name. All models are alignment-tuned variants.

rates differed substantially. This indicates the effect stems from structural sensitivity.

Mechanistic Understanding of Bias To provide a mechanistic understanding of how LLMs exhibit bias toward media outlet names, we conduct a saliency analysis (Simonyan et al., 2014) to quantify the attention models place on outlet names during article bias prediction. Using gradient-based saliency scores, we analyze two models (Mistral-Small$_{\text{24B-Instruct}}$ and Phi-4$_{\text{14B}}$) on a subset of the AllSides dataset consisting of 450 articles.

Both models demonstrated that media outlet name tokens received significantly higher saliency scores than content tokens, with outlet names exceeding 3× the mean token saliency (detailed results in Appendix D.3). Mistral-Small$_{\text{24B-Instruct}}$ showed outlet token saliency means of 2.61 compared to content token means of 0.85, while Phi-4$_{\text{14B}}$ exhibited even stronger patterns with outlet token means of 2.84 versus content means of 0.83. These findings confirm that LLMs' directional bias, which aligns with public perception of media outlet political slant, is mechanistically rooted in the models' disproportionate attention to outlet names rather than article content.

Model	Δ|Pos. ER|	Δ|Neg. ER|	Δ|Neu. ER|
Qwen-2.5$_{\text{72B-Instruct}}$	0.0546	0.1163	0.1248
Mistral-Small$_{\text{24B-Instruct}}$	0.0845	0.1587	0.1821
Phi-4$_{\text{14B}}$	0.0536	0.1177	0.1349
Llama-3.3$_{\text{70B-Instruct}}$	0.0619	0.1409	0.1644
Gemma-2$_{\text{27B-IT}}$	0.0569	0.1283	0.1352

Table 3: Changes in the average proportion of entities by sentiment when media outlet names are included in the summarization task. Δ|Pos. ER| indicates the average change in positive entities, Δ|Neg. ER| for negative entities, and Δ|Neu. ER| for neutral entities.

4.2 News Article Summarization

Entity-level Analysis Media outlet names affect not only article political bias prediction but also summarization. As shown in Table 3, the sentiment of named entities in generated summaries varies depending on the attributed outlet. For Mistral-Small$_{\text{24B-Instruct}}$, the model with the greatest variance, the proportion of positive entities changes by an average of 8.45% compared to summaries generated without outlet attribution. Negative and neutral entities vary even more, by 15.87% and 18.21% respectively, indicating larger fluctuations in non-positive sentiments. No statistically significant trend is observed with respect to summary length. How such variations influence human perception of sentiment or political stance remains an open question for future research. Detailed results by media bias class and summary length are provided in Appendix E.1.

Content-level Analysis We conduct additional evaluation using a pretrained political bias classifier[8] that predicts article-level bias as 0 (left),

[8] https://huggingface.co/matous-volf/political-leaning-politics

1 (center), or 2 (right). The results confirm our earlier findings: summaries generated with left- and right-leaning media outlet names shift political stance compared to those with center-leaning outlet names. Although the effect varies by model, the Gemma-2$_{\text{27B-IT}}$, which shows the second highest SIPS score, exhibits aligned directional bias in this task. However, not all models show proportional shifts, aligning with our stated limitations. Detailed quantitative results and qualitative examples are provided in Appendix E.2.

Human Evaluation To explore how humans perceive changes in political bias levels across LLM-generated news summaries conditioned on different media outlet names, we conduct a crowdsourced study. We recruit five annotators from English-speaking countries representing diverse political orientations (left-leaning, centrist, right-leaning, and far-right). Annotators classify the perceived political stance of summaries generated from identical source articles but prompted with different media outlet names, using a total of 10 articles. Results show that four out of five annotators detect bias perception shifts more frequently than consistent perceptions across outlet-conditioned summaries, validating our findings. Details and complete annotation results are provided in Appendix E.3.

Unexpected Behavior of LLM An unexpected behavior was observed in Llama-3.3$_{\text{70B-Instruct}}$. The model ignored the summarization prompt and appended a note indicating a mismatch between the article's stance and the specified outlet. While unique to this model, the behavior suggests it can distinguish between article content and outlet bias without explicit instruction, warranting further investigation. An example is shown in Appendix E.4.

5 Mitigating Media Outlet Name Bias

5.1 Prompt Optimization Strategies

To develop a model-agnostic and practically applicable method for mitigating media outlet name bias, we adopt an automated prompt optimization framework inspired by Yang et al. (2024b), which treats an LLM as an optimizer. For this process, we extract 10 center-biased articles from the AllSides dataset and apply an initial prompt to each article to compute its corresponding AC, AS, and SIPS scores. These scores serve as the objective signal for the optimizer LLM, GPT-4.1 in our case, which receives the history of previous prompts along with their associated metrics and generates a revised prompt aimed at reducing bias. This iterative refinement continues until either the SIPS score falls below 0.3 or 10 optimization rounds are completed. We then apply the final prompt to original dataset to calculate final SIPS score.

5.2 Results of the Prompt Optimization

Variance of SIPS We confirm that SIPS, AS, and AC scores can be reduced through prompt optimization. Qwen-2.5$_{\text{72B-Instruct}}$, which initially achieved the highest SIPS score of 0.529, reduced to 0.292 after seven iterations, falling below the 0.3 threshold. When applied to the full dataset, the final prompt further reduced SIPS to 0.279, with AS decreasing from 0.439 to 0.385 and AC dropping markedly from 0.605 to 0.088. This indicates that the model became insensitive to media outlet names, with only minor prediction shifts remaining. To evaluate generalizability, we experiment the same prompt on Gemma-2$_{\text{27B-IT}}$, where SIPS fell from 0.510 to 0.362, accompanied by reductions in AC from 0.540 to 0.480 and AS from 0.479 to 0.178. These results demonstrate that SIPS can be minimized through automated prompt refinement, and the method transfers well across models, enabling scalable mitigation of media outlet name bias in article bias prediction. Results of all six major models are presented in the Appendix F.1.

Key Elements of Effective Prompts The final prompt expanded from 550 to 2,969 characters during the optimization process. Two notable trends emerged. First, the prompt increasingly emphasized strict neutrality by assigning the model an explicitly impartial role. Second, rather than asking for a direct bias label, it introduced a structured reasoning framework that led the model to assess article-level bias cues and synthesize a final judgment. While the model does not perform symbolic reasoning, this structure encourages an internal reasoning process that likely reduces bias susceptibility. The full sequence of prompts and corresponding SIPS, AS, and AC scores for each iteration are provided in Appendix F.2.

6 Related Works

6.1 Political Bias in LLMs

Prior work has identified political biases in LLM outputs using surveys and standardized tests, showing a consistent liberal leaning in models like ChatGPT, favoring the US Democratic Party, UK

Labour, and Brazil's Lula over their conservative counterparts (Motoki et al., 2024; Rozado, 2023). These tendencies appear in base models and are often amplified by fine-tuning and alignment processes such as RLHF (Ouyang et al., 2022), which has been shown to induce leftward shifts (Santurkar et al., 2023). Bias intensity also increases with model size (Fulay et al., 2024), suggesting that alignment can inadvertently embed ideological preferences. Similarly, a cross-model analysis by Yang et al. (2024c) showed that political biases in LLMs can intensify with model scale and vary by region of origin, suggesting the need for context-specific bias mitigation. To assess such biases, researchers have applied tools like the Political Compass test and benchmarked outputs against public opinion (Santurkar et al., 2023). Ongoing mitigation efforts include debiasing through fine-tuning (Garimella et al., 2022), prompt-based interventions, and the development of more ideologically balanced alignment datasets, such as the OpenAssistant crowdsourced corpus (Köpf et al., 2023).

However, existing studies focus on general political bias, leaving open questions about how LLMs respond to politically charged context, such as media outlet names. Our work addresses this gap by analyzing bias conditioned on media outlet names, linking general political bias research with context-sensitive evaluation.

6.2 Applications of LLMs in the News Media and Political Science Domain

LLMs are widely used in news workflows, with summarization being a primary application. Instruction-tuned models like GPT-3 show near-human performance in summarization task (Brown et al., 2020; Zhang et al., 2024), though factual consistency remains a challenge. Iterative methods (Zhang et al., 2023) and benchmarks (Laban et al., 2023; Tam et al., 2023) reveal persistent vulnerabilities. Beyond summarization, LLMs assist with headline generation and ideation, enhancing journalist productivity when outputs are curated (Ding et al., 2023; Petridis et al., 2023). However, minimal editorial oversight raises concerns about accuracy, attribution, and confidentiality (Brigham et al., 2024). Generative news recommendation systems further reshape consumption by synthesizing multi-source narratives (Gao et al., 2024).

In the political science domain, Li et al. (2024) proposed the Political-LLM framework, which was developed through interdisciplinary collaboration between computer scientists and political scientists to support various tasks including election forecasting, public opinion analysis, voter simulation, and causal inference. Gujral et al. (2024) employed LLMs such as GPT-4 and LLaMA to predict state-level election outcomes in India based on social media data, achieving superior performance compared to traditional polling methods. Yu et al. (2024) simulated U.S. presidential voting behavior by conditioning LLMs on demographically and ideologically representative personas, demonstrating both predictive accuracy and effective bias control.

Despite their utility, LLMs' political biases may influence summarization and framing, subtly shaping public perception. We investigate how such biases emerge through media outlet name attribution and explore mitigation strategies in this context.

7 Conclusion

This study presents a controlled investigation into media outlet name bias in LLMs. We find that, while the degree of bias varies, most models exhibit clear and consistent political bias in response to outlet names, with directionality largely aligned across models. We demonstrate that LLMs exhibit bias toward both real and fictional media names through linguistic cues rather than factual knowledge alone, with training data distributions potentially explaining observed biases as evidenced in our *Associated Press* case study. The proposed SIPS, AS, and AC metrics effectively quantify this bias and can also guide an automated prompt optimization framework that reduces it through prompting alone.

Limitations

Our study has several limitations pointing to future research directions. We focus exclusively on U.S. news media, analyze bias only in prediction and summarization tasks, and lack exploration of architectural mitigation strategies. In addition, our sentiment-based analysis provides limited directional assessment, and our metrics show sensitivity to article selection.

Ethical Considerations

Our study investigates political bias in LLMs with a focus on media outlet name attribution. While the goal is to advance accountability in LLM-based news processing, our findings and methods carry several ethical considerations. We present comprehensive ethical considerations in Appendix A.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was partly supported by the Basic Research Program through a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.2022R1C1C1010317) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25443681).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Mistral AI. 2024. Introducing mistral small. https://mistral.ai/news/mistral-small-3. Accessed: 2025-04-29.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. 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)*, EMNLP '20.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. 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 (Volume 1: Long Papers)*, pages 11142–11159.

David P Baron. 2006. Persistent media bias. *Journal of Public Economics*, 90(1-2):1–36.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bloomberg. 2024. Bloomberg launches gen ai summarization for news content. https://www.bloomberg.com/company/press/bloomberg-launches-gen-ai-summarization-for-news-content/. Accessed: 2025-04-28.

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. 2024. Developing story: Case studies of generative ai's use in journalism. In *Workshop on Socially Responsible Language Modelling Research*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stephen Collinson. 2025. Trump says he would urge putin and zelensky to agree to a ceasefire. https://edition.cnn.com/2025/05/20/politics/trump-putin-russia-ukraine-ceasefire-talks. Accessed: 2025-05-20.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339.

James N Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018.

Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Generative news recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 3444–3453.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319.

R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pratik Gujral, Kshitij Awaldhi, Navya Jain, Bhavuk Bhandula, and Abhijnan Chakraborty. 2024. Can llms help predict elections?(counter) evidence from the world's largest democracy. *arXiv preprint arXiv:2405.07828*.

Felix Hamborg and Karsten Donnay. 2021. Newsmtsc: (multi-)target-dependent sentiment classification in news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Summedits: Measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9662–9676.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, and 1 others. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Mitchell Linegar, Rafal Kocielnik, and R Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science*, 5:1257092.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*.

Riccardo Lunardi, David La Barbera, and Kevin Roitero. 2024. The elusiveness of detecting political bias in language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3922–3926.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. 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: Human Language Technologies (Volume 1: Long Papers)*, pages 4083–4098.

Killian J McCarthy and Wilfred Dolfsma. 2014. Neutral media? evidence of media bias and its economic impact. *Review of Social Economy*, 72(1):42–54.

Uwe Messer. 2025. How do people react to political bias in generative artificial intelligence (ai)? *Computers in Human Behavior: Artificial Humans*, 3:100108.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Jeremy K Nguyen. 2024. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. International Conference on Learning Representations.

Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024b. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Kaiqi Yang, Hang Li, Yucheng Chu, Yuping Lin, Tai-Quan Peng, and Hui Liu. 2024c. Unpacking political bias in large language models: Insights across topic polarization. *arXiv e-prints*, pages arXiv–2412.

Chenxiao Yu, Jinyi Ye, Yuangang Li, Zheng Li, Emilio Ferrara, Xiyang Hu, and Yue Zhao. 2024. A large-scale simulation on large language models for decision-making in political science. *arXiv preprint arXiv:2412.15291*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Summit: Iterative text summarization via chatgpt. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Comprehensive Ethical Considerations

Potential Risks of Misuse. The metrics and techniques proposed in this paper, such as SIPS and automated prompt optimization, could potentially be misused to mask or intentionally amplify specific ideological leanings in LLMs. Although our framework aims to mitigate unintended bias, it could be repurposed to produce content that appears neutral while being subtly slanted, thereby undermining trust in LLM outputs.

Impact on Public Discourse. Given the increasing integration of LLMs into news summarization, recommendation, and generation pipelines, biased outputs, even when subtle, can influence public perception, reinforce echo chambers, or distort information credibility. Our work highlights these risks and advocates for proactive bias detection and mitigation. However, downstream deployment decisions remain outside the scope of our control, and ethical outcomes will depend heavily on how stakeholders implement these tools.

Bias in Ground Truth Labels. This study relies on datasets like AllSides and Hyperpartisan, which use outlet- or article-level annotations as proxies for ground-truth political bias. While widely used, such annotations are themselves subjective and may encode societal or annotator-specific biases.

Generalizability and Representation. We focus exclusively on English-language U.S. news media due to data availability and alignment with LLM pretraining corpora. Consequently, the findings may not generalize to LLM behavior in multilingual, global, or non-Western media contexts. Additionally, by focusing on prominent media outlets, our study may underrepresent the perspectives of smaller or alternative publications.

LLMs as Political Actors. Our work contributes to a growing body of research treating LLMs as semi-autonomous agents that can shape user perceptions through seemingly neutral outputs. We emphasize that these models do not possess intent or ideology but rather reflect statistical patterns in data. Nevertheless, the sociopolitical consequences of these patterns warrant serious attention and continued interdisciplinary oversight.

Transparency and Reproducibility. To support transparency, we release detailed prompt templates, evaluation metrics, and model configurations in the appendices. We release our official codebase. We encourage further open-source benchmarking and community-driven evaluations to verify and extend our results.

B Interpretation of SIPS

Table 4 presents illustrative examples of SIPS, AS, and AC scores across a set of synthetic scenarios designed to highlight key behavioral patterns of the proposed metrics. Each row simulates a hypothetical model response to a given article a, with predictions provided for three input conditions: no media outlet name, and attribution to a left-, center-, or right-leaning outlet.

- **Correct direction**: The model shifts its predictions in alignment with each outlet's political orientation (e.g., -1 for left, +1 for right). This yields perfect agreement coherence (AC = 1.00), moderate absolute sensitivity (AS = 0.33), and a high SIPS score (0.74).
- **No direction, max bias**: All predictions shift to the same extreme regardless of outlet (e.g., all shift to +1), resulting in high AS (1.00) but low AC (0.33), as the shifts do not align directionally. SIPS remains moderate (0.74).
- **Opposite direction**: Shifts are in the reverse of the expected directions (e.g., +1 for left), leading to low AC (0.33) and moderate AS (0.33), with SIPS dropping to 0.33.
- **No shift cases**: The last three rows simulate scenarios where the model's prediction remains fixed regardless of media outlet input. Although there is no shift (AS = 0.00), center tolerance allows AC to remain at 0.33 due to the center outlet's low threshold ($\delta = 0.3$). SIPS reaches its theoretical minimum (≈ 0.24).

These cases serve as interpretive anchors for understanding SIPS behavior in real model outputs, illustrating how AS and AC interact to capture both the strength and directionality of bias induced by media outlet attribution.

C Implementation Details

All experiments were conducted using three NVIDIA A100 80GB GPUs. For each dataset, bias inference for each LLM took approximately 20 hours to complete.

Scenario	$f_\theta(p, \varnothing, a)$	$f_\theta(p, left, a)$	$f_\theta(p, center, a)$	$f_\theta(p, right, a)$	$d(left, a)$	$d(center, a)$	$d(right, a)$	**AS** ↑	**AC** ↑	**SIPS** ↑
Correct dir.	0	-1	0	$+1$	-1	0	$+1$	0.33	1.00	0.74
No dir. & max bias	-1	1	1	1	2	2	2	1.00	0.33	0.74
Opposite dir.	0	$+1$	0	-1	$+1$	0	-1	0.33	0.33	0.33
$\varnothing$ -1 & no shift	-1	-1	-1	-1	0	0	0	0.00	0.33	0.24
$\varnothing$ 0 & no shift	0	0	0	0	0	0	0	0.00	0.33	0.24
$\varnothing$ 1 & no shift	$+1$	$+1$	$+1$	$+1$	0	0	0	0.00	0.33	0.24

Table 4: Illustrative SIPS components for five synthetic scenarios. Each prediction corresponds to a bias score at different political orientations: left, center, and right. AS measures bias-shift magnitude, AC measures directional coherence (center-group tolerance $\delta = 0.3$), and SIPS is their product.

C.1 Entity-wise Summarization Analysis Method

Following Bang et al. (2024), we use an open-source NER model from Hugging Face[9] and the target sentiment classifier proposed by Hamborg and Donnay (2021) to analyze the sentiment of extracted named entities.

C.2 Detailed Criteria for LLM Selection

To focus our analysis on widely adopted LLMs, we employed the following selection procedure. We began by identifying models labeled as “Text Generation” on Hugging Face[10], and retrieved their total download counts via the Hugging Face Hub API. Download statistics were aggregated at the model family level, and the ten most downloaded families were selected for inclusion.

We then applied a set of exclusion criteria. Models specialized for reasoning tasks were omitted, as they were designated for separate experiments. We also excluded models exceeding 100B parameters, which were intended to be evaluated via API access. In addition, we removed any model families that lacked at least one model above 10B parameters or did not include an instruction-tuned variant. For families with multiple versions, only the most recent release was retained. All artifacts we used were free to use for research.

C.3 Detailed LLM Prompts

Table 6 presents the prompts used in each experiment.

C.4 Detailed LLM Generation Configuration

We presented multiple bias class options prefixed with alphabetical labels (e.g., A, B, C) and determined the model’s choice by analyzing the logits of the first generated token. This approach minimized generation instability and improved prediction accuracy. For closed-source LLMs, where direct access to logits was unavailable, we employed greedy decoding with a constraint to generate only a single token and extracted the bias prediction from the output. For reasoning-specialized LLMs, we adopted the recommended sampling configurations and required outputs to be returned in a structured JSON format.

D Additional Analysis Results for Article Bias Prediction

D.1 Detailed Analysis by Media Bias Labels and Names

Figure 6 presents the analysis of input media bias and predicted bias across both datasets.

D.2 Generated & Formulated Media Outlet Names

The following lists contain the fictitious media outlet names used in our experiments to test whether LLMs are influenced by the political connotations of outlet names, even when those names are fabricated.

Formulated Left-Leaning Outlets *Guardian Slate*, *Beast Nation*, *Intercept Times*, *Semafor Nation*, and *HuffGuardian*.

Formulated Right-Leaning Outlets *BreitWire*, *Federal Caller*, *Conservative Examiner*, *Millennial Times*, and *Spectator Wire*.

Generated Left-Leaning Outlets *The Liberty Press*, *Progressive Voice*, *New Dawn Journal*, *Equity Times*, and *True Progress News*.

Generated Right-Leaning Outlets *Great America Times*, *Patriot’s Chronicle*, *Freedom Sentinel*, *Constitution Today*, and *True Valor News*.

[9] https://huggingface.co/dslim/bert-large-NER

[10] https://huggingface.co/

D.3 Saliency Analysis Details

We utilize the captum[11] package in Python to compute gradient-based saliency scores. Due to hardware limitations on our end and BFloat computation support issues, we exclude models over 70B parameters and models requiring BFloat support. Instead, we analyze two models: Mistral-Small$_{\text{24B-Instruct}}$ and Phi-4$_{\text{14B}}$ as referenced in our paper. Table 5 presents the saliency analysis results.

E Additional Analysis Results for Article Summarization

E.1 Detailed Result of Entity-wise Summarization Analysis

Table 7 presents the sentiment distribution of model-generated summaries across different summary lengths and media bias categories.

E.2 Detailed Result of Content-wise Summarization Analysis

Table 8 presents average predicted political bias scores across five models, grouped by input media outlet bias classification. The results demonstrate varying degrees of bias amplification across models. Qwen-2.5$_{\text{72B-Instruct}}$ shows clear directional shifts (left: 0.87, center: 0.72, right: 0.92), while Mistral-Small$_{\text{24B-Instruct}}$ exhibits more subtle variations (left: 0.46, center: 0.39, right: 0.47). Notably, Gemma-2$_{\text{27B-IT}}$, which achieves the second-highest SIPS score in our main analysis, also displays aligned directional bias in this task.

Qualitative Analysis We identify systematic linguistic adjustments that align with outlet-specific framing patterns. Representative examples include:

- **Obama Speech Coverage:** Gemma-2$_{\text{27B-IT}}$ shifts from "cast Trump as a threat" (*HuffPost*) to "painted Trump as a threat" (*BBC*) to omitting the reference entirely (*Breitbart*).
- **Richard Spencer Speech:** The same model evolves from "highlights the ongoing debate" (*Associated Press*) to "Students and faculty are divided" (*Forbes*) to "divided the student body" (*Fox News*).
- **Border Wall Emergency:** Qwen-2.5$_{\text{72B-Instruct}}$ changes from "not receiving pay" (*Guardian*) to "800,000 workers" (*CNBC*) to "remain unpaid" (*Daily Mail*).

Discussion The analysis confirms that LLMs generate summaries with varying political stances depending on input media outlet names. However, the extent of bias and proportionality between input outlet stance and output vary considerably across models. These variations warrant deeper investigation into the mechanisms through which outlet names influence model behavior and the factors determining effect magnitude.

E.3 Human Evaluation Details

We conduct a small-scale crowdsourced study via Positly[12] to assess whether human annotators perceive political stance differences in outlet-conditioned summaries.

Methodology We recruit five annotators from English-speaking countries who self-identify their political orientations: one left-leaning, one centrist, one right-leaning, and two far-right. Annotators classify the perceived political stance (left, center, or right) of summaries generated from identical source articles but prompted with left-, center-, or right-leaning media outlet names. The summaries used are identical to those in content-wise analysis. Each annotator receives $15 compensation with an average completion time of 14.7 minutes.

Annotator Instructions Below, we report the instructions presented to annotators before and during the human evaluation process.

Political Bias Classification Instructions. We are conducting a study to measure the degree of political bias present in AI-generated news summaries. You will be shown three summaries. For each one, select the political bias that best fits based on its language and framing: - Left**: Generally progressive or liberal viewpoints; emphasis on social equity, environmental concerns, or support for government intervention. - Center**: Neutral or balanced tone; avoids taking a strong political stance or presents multiple sides fairly. - Right**: Generally conservative viewpoints; emphasis on tradition, market freedom, national security, or limited government. If uncertain, choose the category that the overall tone and language lean toward the most.

Task description screen provided to annotators during the human evaluation process is described in Figure 7 and the actual annotation interface provided to the annotators is described in Figure 8.

[11] https://captum.ai/

[12] https://www.positly.com/

Model	Media Outlet Tokens Saliency Mean	Media Outlet Tokens Saliency Std	Content Tokens (Reference) Saliency Mean	Content Tokens (Reference) Saliency Std
Mistral-Small$_{\text{24B-IT}}$	2.6093	0.8385	0.8493	0.3292
Phi-4	2.8368	0.7992	0.8325	0.3291

Table 5: Detailed saliency analysis results.

Detailed Results Table 9 presents the frequency of bias perception shifts versus consistent perceptions for each annotator. Four out of five annotators detect political stance differences across outlet-conditioned summaries more frequently than consistent perceptions, suggesting human evaluators perceive media outlet name-induced bias shifts.

Qualitative Analysis Table 10 illustrates representative annotation patterns for summaries generated from the same source article with different outlet prompts. In some cases, summaries prompted with left- or right-leaning outlet names are perceived as more politically extreme than those with center-leaning outlets. For example, multiple annotators perceive leftward shifts when right-leaning outlet names are used, suggesting complex interaction patterns between outlet bias and perceived content stance.

E.4 Example of LLM Political Bias Perception

Table 11 presents an instance where the Llama-3.3$_{\text{70B-Instruct}}$ model, during summarization, incidentally generated content that reflects its perception of the article's political bias.

F Additional Prompt Optimization Results

F.1 Detailed Prompt Optimization Results Across 6 Major Models

Table 12 provides prompt optimization results using SIPS across the 6 major models. Consistent and effective reductions in SIPS, AS, and AC scores are observed across all models.

F.2 Prompts Generated by Iterative Optimization

Table 13 presents the changes in SIPS, AS, and AC scores of Qwen-2.5$_{\text{72B-Instruct}}$ across seven rounds of iterative prompt optimization using 10 articles. Table 14 shows the final optimized prompt.

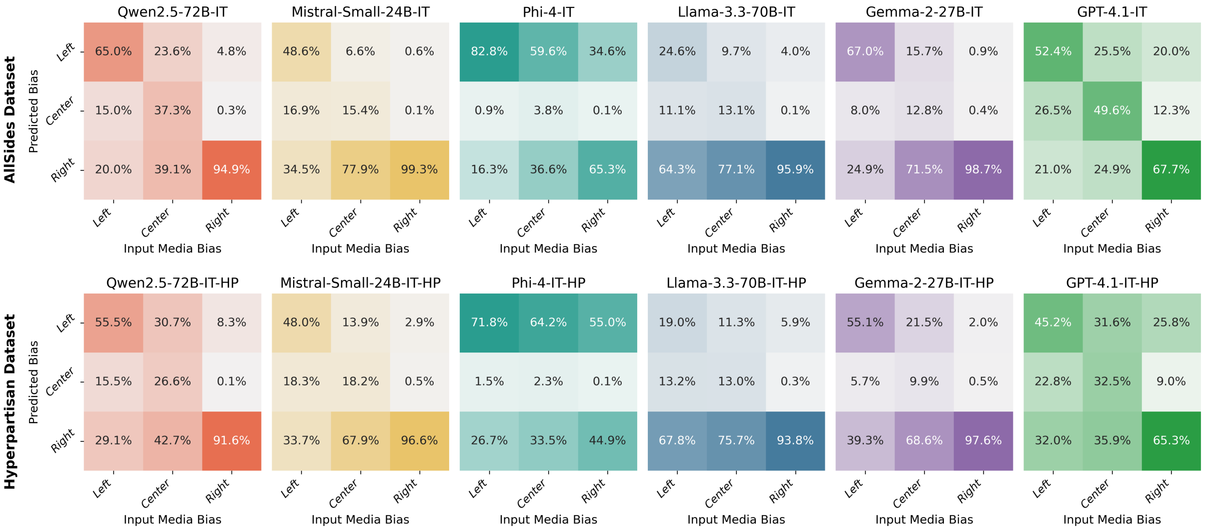


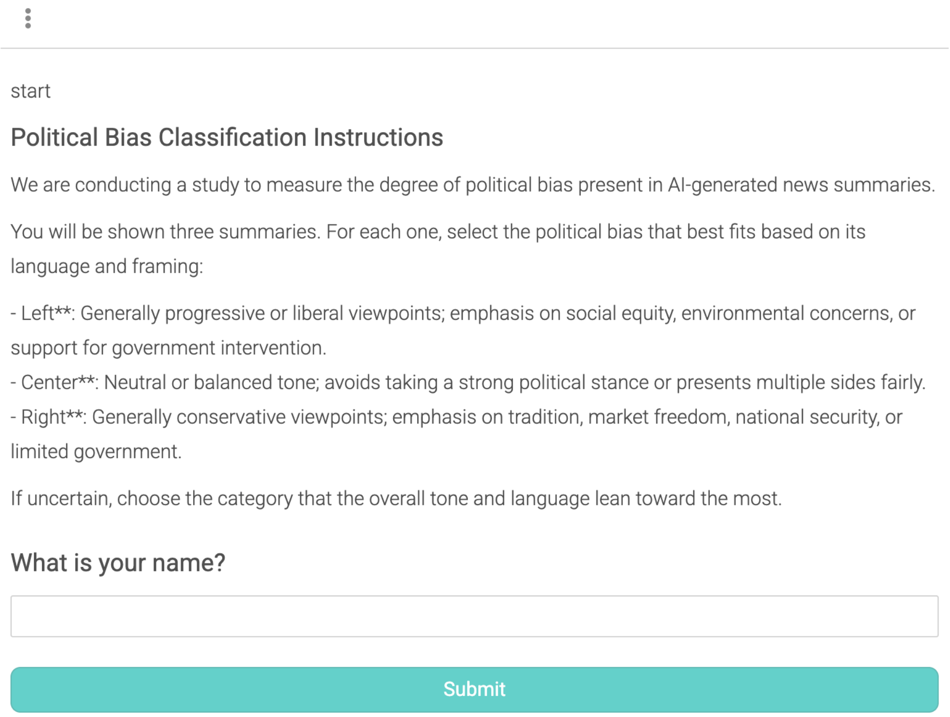
Figure 6: Extended results of predicted bias heatmaps by input media bias across models. Predicted bias is calculated by averaging model outputs and mapping them to discrete classes, with scores in the range (–0.3, 0.3) classified as center. IT refers to instruction tuned models.

Task	Prompts
Bias Prediction	You are an expert in political media analysis. Your task is to determine the political bias of the following news text based on its language, tone, framing, and implicit political messages. Classify the text into one of the following categories: A. Left B. Center C. Right Your final output must consist of only one uppercase letter (A, B or C) corresponding to your classification. Do not include any explanation or extra text. Read the content between "###Content starts." and "###Content ended.", and determine the political bias of the content.
Summarization	Article: {article} Summarize the article in {summary_length} sentences.

Table 6: The actual prompt used in the experiment is shown.

		Left Biased Media			Right Biased Media			Center Biased Media		
Model	Summary Length	Positive ER	Negative ER	Neutral ER	Positive ER	Negative ER	Neutral ER	Positive ER	Negative ER	Neutral ER
Qwen-2.5$_{\text{72B-Instruct}}$	3	0.0535	0.0947	0.1136	0.0540	0.0939	0.1149	0.0576	0.0967	0.1127
	5	0.0580	0.1147	0.1223	0.0503	0.1166	0.1230	0.0554	0.1177	0.1291
	10	0.0524	0.1253	0.1385	0.0526	0.1217	0.1297	0.0523	0.1221	0.1320
Mistral-Small$_{\text{24B-Instruct}}$	3	0.0590	0.1211	0.1466	0.0677	0.1314	0.1525	0.0627	0.1166	0.1477
	5	0.0807	0.1592	0.1776	0.0903	0.1579	0.1852	0.0825	0.1589	0.1836
	10	0.0721	0.1487	0.1628	0.0669	0.1376	0.1630	0.0665	0.1469	0.1641
Phi-4$_{14B}$	3	0.0638	0.1249	0.1474	0.0625	0.1282	0.1437	0.0606	0.1275	0.1375
	5	0.0556	0.1094	0.1296	0.0503	0.1176	0.1353	0.0548	0.1262	0.1398
	10	0.0613	0.1166	0.1177	0.0562	0.1227	0.1194	0.0539	0.1230	0.1194
Llama-3.3$_{\text{70B-Instruct}}$	3	0.0787	0.1395	0.1681	0.0765	0.1505	0.1684	0.0775	0.1466	0.1789
	5	0.0669	0.1420	0.1693	0.0558	0.1416	0.1595	0.0629	0.1391	0.1644
	10	0.0702	0.1340	0.1533	0.0633	0.1304	0.1485	0.0772	0.1416	0.1607
Gemma-2$_{\text{27B-IT}}$	3	0.0721	0.1108	0.1290	0.0744	0.1134	0.1362	0.0733	0.1070	0.1264
	5	0.0563	0.1357	0.1388	0.0561	0.1228	0.1356	0.0582	0.1263	0.1313
	10	0.0542	0.1037	0.1037	0.0526	0.1109	0.1077	0.0592	0.1058	0.1056

Table 7: Changes in entity sentiment proportions during summarization for each model, by input media bias and summary length.



start

Political Bias Classification Instructions

We are conducting a study to measure the degree of political bias present in AI-generated news summaries.

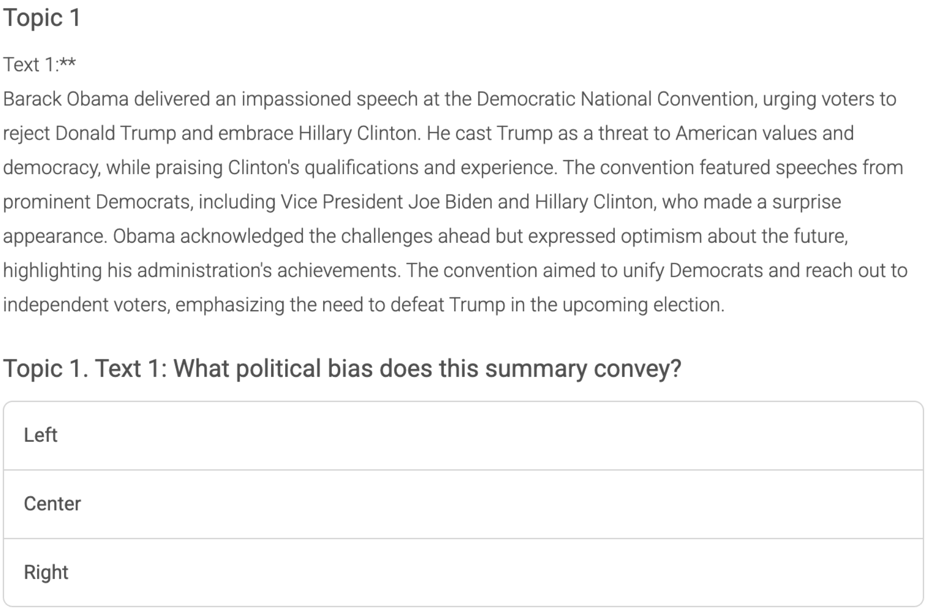
You will be shown three summaries. For each one, select the political bias that best fits based on its language and framing:

- Left**: Generally progressive or liberal viewpoints; emphasis on social equity, environmental concerns, or support for government intervention.
- Center**: Neutral or balanced tone; avoids taking a strong political stance or presents multiple sides fairly.
- Right**: Generally conservative viewpoints; emphasis on tradition, market freedom, national security, or limited government.

If uncertain, choose the category that the overall tone and language lean toward the most.

What is your name?

Figure 7: Task description screen provided to annotators during the human evaluation process.



Topic 1

Text 1:**
Barack Obama delivered an impassioned speech at the Democratic National Convention, urging voters to reject Donald Trump and embrace Hillary Clinton. He cast Trump as a threat to American values and democracy, while praising Clinton's qualifications and experience. The convention featured speeches from prominent Democrats, including Vice President Joe Biden and Hillary Clinton, who made a surprise appearance. Obama acknowledged the challenges ahead but expressed optimism about the future, highlighting his administration's achievements. The convention aimed to unify Democrats and reach out to independent voters, emphasizing the need to defeat Trump in the upcoming election.

Topic 1. Text 1: What political bias does this summary convey?

Figure 8: Actual evaluation screen provided to annotators during the human evaluation process.

Model	Bias of Input Media Outlet	Avg. Bias Score
Qwen-2.5$_{\text{72B-Instruct}}$	Left	0.8667
	Center	0.7222
	Right	0.9222
Mistral-Small$_{\text{24B-Instruct}}$	Left	0.4556
	Center	0.3889
	Right	0.4667
Phi-4$_{\text{14B}}$	Left	0.8000
	Center	0.7444
	Right	0.7778
Llama-3.3$_{\text{70B-Instruct}}$	Left	0.7778
	Center	0.7333
	Right	0.7667
Gemma-2$_{\text{27B-IT}}$	Left	0.6889
	Center	0.7222
	Right	0.7556

Table 8: The average predicted political bias score by the classifier for summaries generated by the model, grouped by the input media outlet's bias class. For example, if the predictions for four summaries are 0 (Left), 0 (Left), 1 (Center), and 2 (Right), the Avg. Bias Score would be 0.75.

Annotator	Political Orientation of Annotator	# of Bias Perception Shifts (Post-Summary)	# of Bias Perception Consistent (Post-Summary)
Coder 1	Moderate	7	3
Coder 2	Conservative	5	5
Coder 3	Liberal	9	1
Coder 4	Very conservative	7	3
Coder 5	Very conservative	6	4

Table 9: The frequency of bias perception shifts versus consistent perceptions for each annotator.

Annotator	Response for Summary 1 (left media outlet name added)	Response for Summary 2 (center media outlet name added)	Response for Summary 3 (right media outlet name added)
Coder 1	Left	Left	Center
Coder 2	Center	Center	Center
Coder 3	Right	Center	Left
Coder 4	Center	Center	Left
Coder 5	Center	Center	Left

Table 10: Each annotator's responses to three summaries generated from the same source article, based on prompts containing left-, center-, and right-leaning media outlet names.

Model	Generated Summary
Llama-3.3$_{\text{70B-Instruct}}$	President Trump held a contentious press conference at Trump Tower where he defended his original claim that both sides were to blame for the violence in Charlottesville, Va. He insisted that there were "very fine people" on both sides, including the neo-Nazis and white supremacists, and that the "alt-left" protesters were also culpable for the violence. The president's comments were widely criticized and praised by white nationalists, including former Ku Klux Klan leader David Duke, Note: The article is not from Breitbart News, it seems to be from a liberal or left-leaning news source, given the tone and content of the article.

Table 11: Generated article summary with unexpected media outlet bias recognition.

Model	SIPS (Before Mitigation)	SIPS (After Mitigation)	AS (Before Mitigation)	AS (After Mitigation)	AC (Before Mitigation)	AC (After Mitigation)
Qwen-2.5$_{\text{72B-Instruct}}$	0.529	0.279	0.439	0.385	0.605	0.088
Mistral-Small$_{\text{24B-Instruct}}$	0.478	0.356	0.426	0.133	0.525	0.441
Phi-4$_{\text{14B}}$	0.475	0.366	0.468	0.228	0.482	0.330
Llama-3.3$_{\text{70B-Instruct}}$	0.387	0.363	0.358	0.209	0.414	0.399
Gemma-2$_{\text{27B-IT}}$	0.510	0.362	0.479	0.178	0.540	0.480
GPT-4.1	0.421	0.293	0.266	0.094	0.532	0.364

Table 12: Prompt optimization results using SIPS for the 6 major models.

Round	SIPS	AS	AC
0	0.499	0.311	0.633
1	0.425	0.278	0.533
2	0.437	0.311	0.533
3	0.362	0.211	0.467
4	0.311	0.078	0.433
5	0.334	0.189	0.433
6	0.321	0.133	0.433
7	0.292	0.100	0.400

Table 13: SIPS, AS, and AC scores of Qwen-2.5$_{\text{72B-Instruct}}$ across prompt optimization rounds. Optimization terminated after Round 7 when SIPS dropped below the threshold of 0.3.

Round	Generated Prompt
7	You are a hyper-rigorous political bias classification engine. Your exclusive function is to assign a political bias label to a news text segment strictly and solely by analyzing explicit, quantifiable linguistic features within the section marked “###Content starts.” and “###Content ended.” Absolutely disregard all external knowledge, author intent, context, or any inferred meaning. Apply this exhaustive, step-by-step analytical framework: 1. Decomposition and Tagging: a. Break the text into sentences. b. For each sentence, tag every instance of: - Evaluative adjectives/adverbs directly describing political entities, actions, or policies. - Verbs denoting approval, disapproval, endorsement, or criticism concerning political topics. - Nouns or noun phrases acting as partisan identifiers, slogans, or loaded terminology associated with left, center, or right ideologies. - Direct quotations or paraphrases explicitly supporting or denouncing any political position. 2. Quantitative Coding: a. For each tagged instance, assign: - Bias Direction: Left-favoring, Right-favoring, or Neutral/Balanced. - Intensity Score: 1 (Mild/suggestive), 2 (Moderate/clear), 3 (Strong/forceful/dogmatic). - Position Weight: Headline/title (4x), lead/first paragraph (3x), main body (2x), conclusion/summary (2x), supporting quote (1x). - Framing Weight: Standalone assertion (2x), attribution to source (1x), hypothetical/conditional (0.5x). b. Calculate the Weighted Bias Score for each instance: Intensity × Position Weight × Framing Weight. 3. Aggregation: a. Sum Weighted Bias Scores for Left-favoring, Right-favoring, and Neutral/Balanced separately. b. Compute each direction’s percentage of the total weighted score (sum Left, Right, Neutral). 4. Dominance Assessment: a. If one direction’s percentage exceeds both others by at least 20 percentage points, and both opposing directions individually account for less than 25% of the total, assign that direction as dominant. b. If none meets this threshold, or if Neutral/Balanced is highest but not dominant by threshold, default to Center. 5. Strict Output Protocol: - Output ONLY a single uppercase letter according to the following mapping: A: Left – Left-favoring cues are quantitatively and proportionally dominant by the criteria above. B: Center – No direction meets dominance criteria, or Neutral/Balanced has the highest score. C: Right – Right-favoring cues are quantitatively and proportionally dominant by the criteria above. - Do NOT include any explanation, reasoning, punctuation, or additional text. Any ambiguous, implied, sarcastic, or context-dependent content MUST be excluded from consideration unless it contains unmistakable, explicit partisan markers. Your decision must be based exclusively on the systematic, mathematical analysis of explicit linguistic evidence within the provided boundaries.

Table 14: Final prompt generated after 7 rounds of iterative optimization.