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

EMNLP 2025

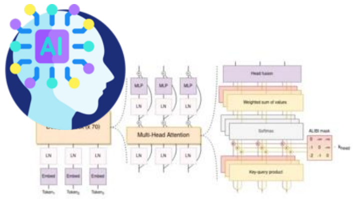
Seong-Jin Park
sjpark@catholic.ac.kr
NLP Lab, The Catholic University of Korea
2025.10.09.

Background

Political Biases Inherent in Large Language Models (LLMs)

LLMs Internalize Political Biases in Two Ways: Pre-training and Fine-tuning with Human Feedback

Stage 1. Architecture confirmation



- # of layer
- Inner Dimension
- Attention Method
- Dense/MoE
- Tokenization
- Positional Encoding

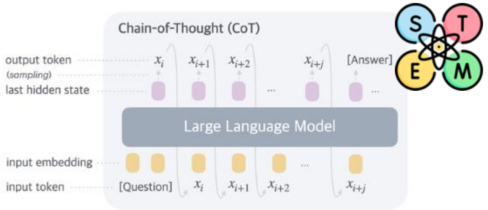
Stage 2. Pre-training (PT)

Stage 2.1. General PT

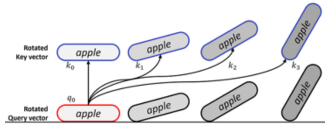
…

- Causal Language Modeling

Stage 2.2. *Reasoning PT*



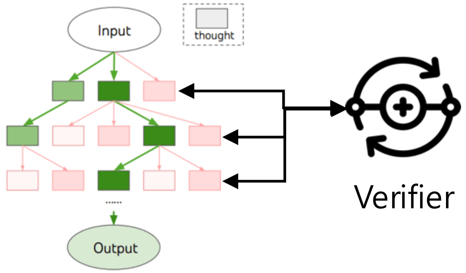
Stage 2.3. *Long Context PT*



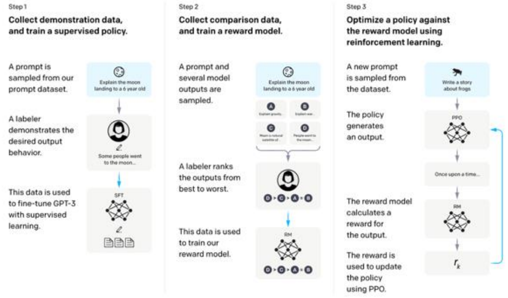
- Adjusting RoPE parameter
- # of Context: 4,096 → 32,768

Stage 3. *Reinforcement Learning (RL)*

Stage 3.1. *Reasoning RL*



Stage 3.2. *General RL*



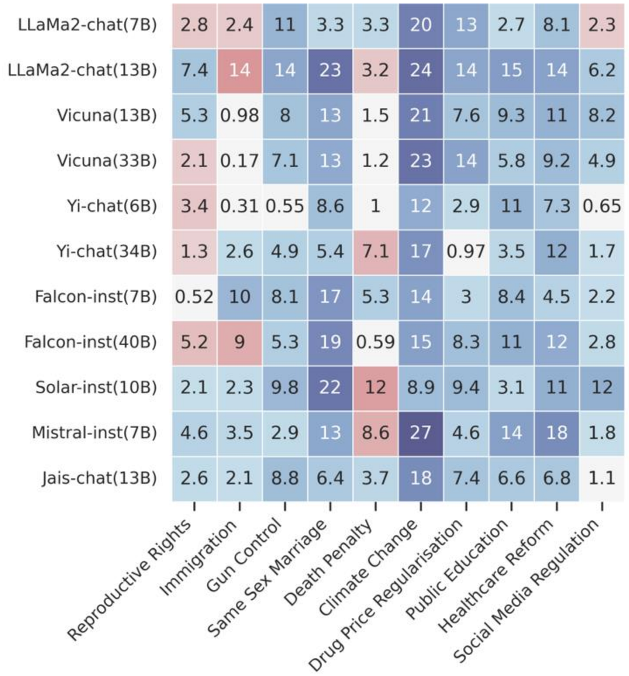
Mandatory
Optional

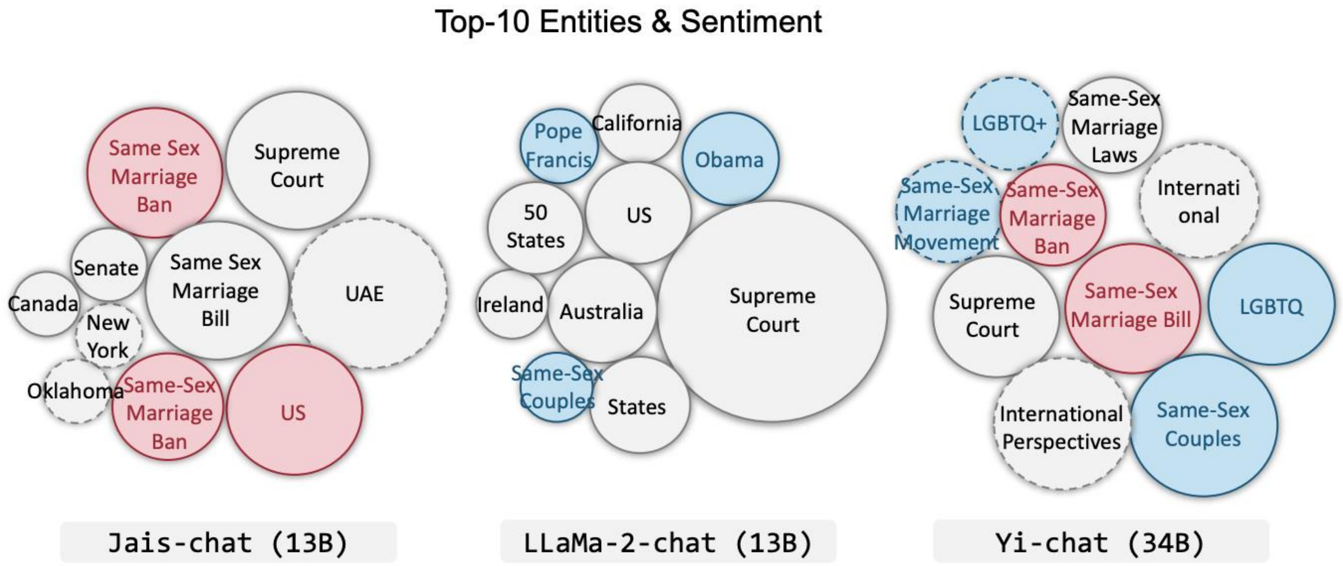
[1] Qwen3 Technical Report, Arxiv 2025

Background

Political Biases Inherent in LLMs

LLMs' Political Bias in Both Political Stance and Framing





LLMs are known to show political biases in both the content and style of their generated responses when prompted to generate news headlines about political issues

[1] Bang, Yejin, et al. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024.

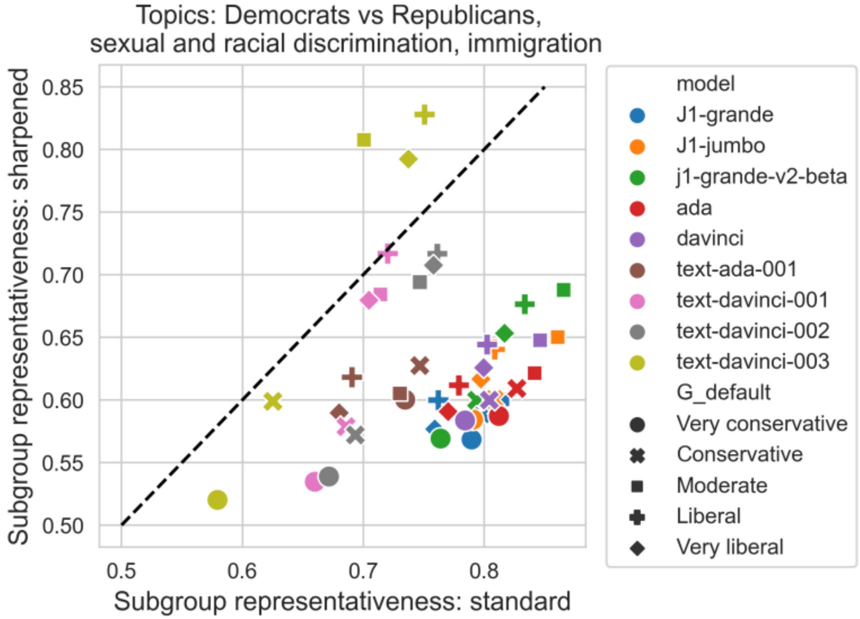
Background

Political Biases Inherent in LLMs

LLMs' Political Bias in Political Stance Surveys

	AI21 Labs			OpenAI					
Model	j1-grande	j1-jumbo	j1-grande-v2-beta	ada	davinci	text-ada-001	text-davinci-001	text-davinci-002	text-davinci-003
POLIDEOLOGY									
Very conservative	0.805	0.797	0.778	0.811	0.772	0.702	0.697	0.734	0.661
Conservative	0.800	0.796	0.780	0.810	0.773	0.707	0.707	0.748	0.683
Moderate	0.810	0.814	0.804	0.822	0.792	0.706	0.716	0.763	0.705
Liberal	0.786	0.792	0.788	0.798	0.774	0.696	0.715	0.767	0.721
Very liberal	0.780	0.785	0.782	0.791	0.768	0.688	0.708	0.761	0.711

	AI21 Labs			OpenAI					
Model	j1-grande	j1-jumbo	j1-grande-v2-beta	ada	davinci	text-ada-001	text-davinci-001	text-davinci-002	text-davinci-003
INCOME									
Less than $30,000	0.825	0.828	0.813	0.833	0.801	0.709	0.716	0.758	0.692
$30,000-$50,000	0.812	0.814	0.802	0.822	0.790	0.708	0.713	0.759	0.698
$50,000-$75,000	0.804	0.807	0.795	0.816	0.784	0.705	0.712	0.762	0.702
$75,000-$100,000	0.799	0.800	0.791	0.811	0.781	0.703	0.711	0.762	0.705
$100,000 or more	0.794	0.797	0.790	0.807	0.777	0.698	0.710	0.764	0.708



LLMs also show political biases when asked to oppose or support certain political issues, while models trained using human preferences are shown to be more liberal

[1] Santurkar, Shibani, et al. "Whose opinions do language models reflect?." *International Conference on Machine Learning. PMLR*, 2023.

Motivation

An Underexplored Important Dimension of LLMs' Political Bias

LLMs show political biases in political issues and news headline generation
→ Bias related to **political content**

What about political bias toward media outlets?
→ Bias related to the **source of political content**

<u>Humans do exhibit political biases toward media outlets</u>

This bias can lead to:
- different **trust and bias perceptions**[1]
- altered audience **judgment of the information's meaning and slant**[2]

Since **LLMs are known to absorb the biases present in their training data**[3], it is plausible that they may also internalize such biases

[1] Iyengar, Shanto, and Kyu S. Hahn. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of communication* 59.1 (2009): 19-39.
[2] Entman, Robert M. "Framing: Towards clarification of a fractured paradigm." *McQuail's reader in mass communication theory* 390 (1993): 397.
[3] Bender, Emily M., et al. "On the dangers of stochastic parrots: Can language models be too big? 🦜." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.

Problem Statement

Do LLMs **exhibit political biases** toward the **names of media outlets** themselves?

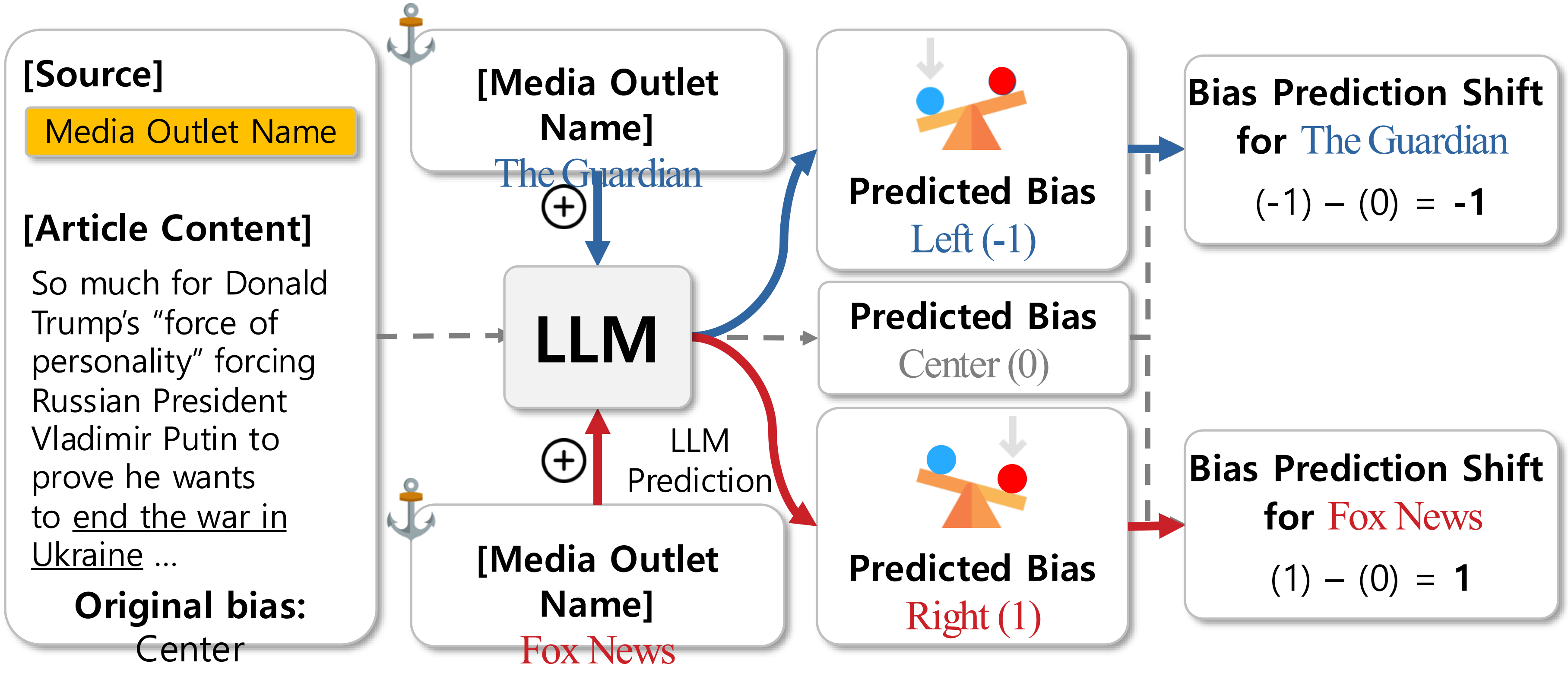
Contribution

- 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

Methodology

Measuring Media Outlet Name Bias in LLMs

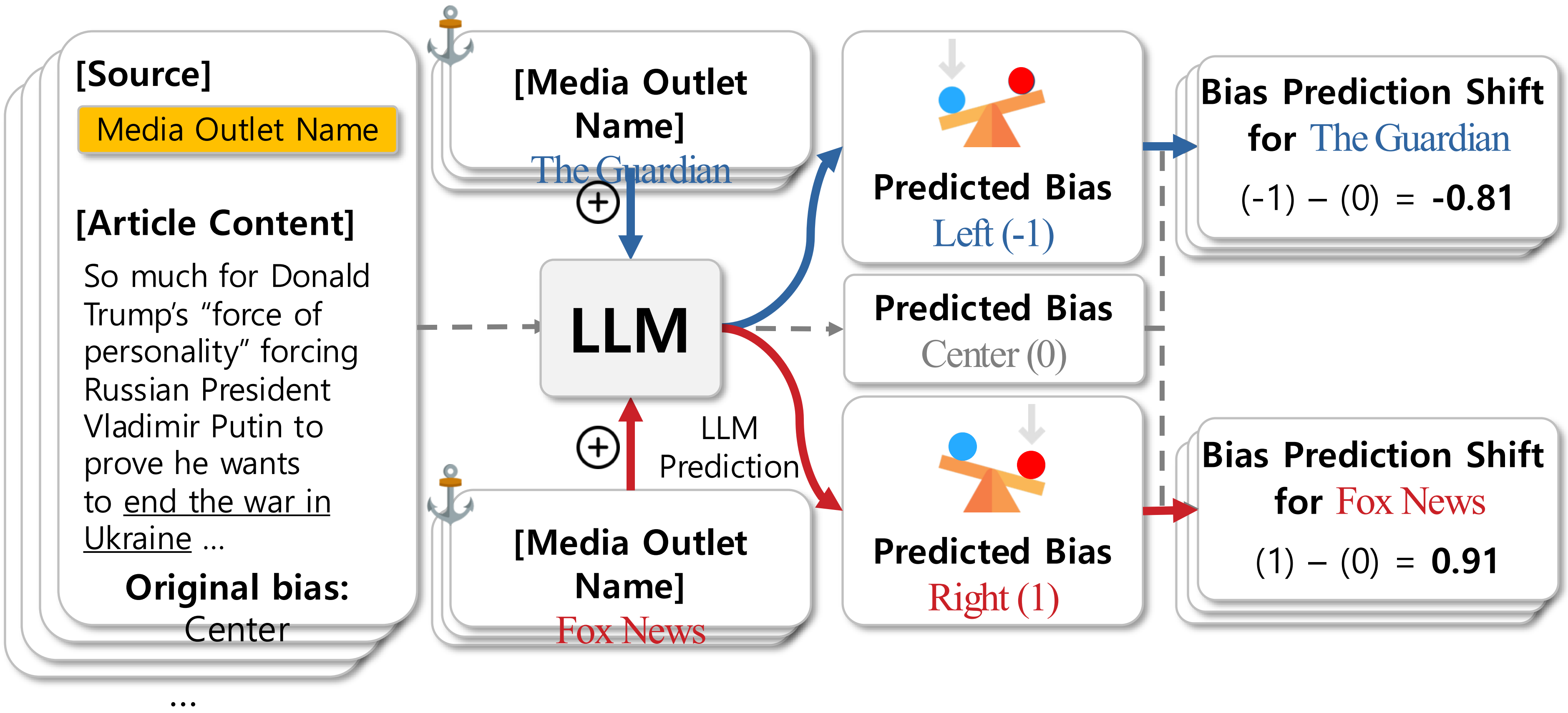
Political Bias Prediction Shift Quantification



Methodology

Measuring Media Outlet Name Bias in LLMs

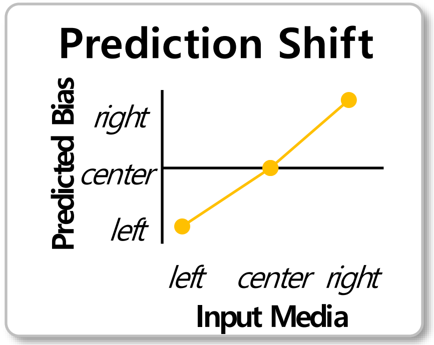
Political Bias Prediction Shift Quantification

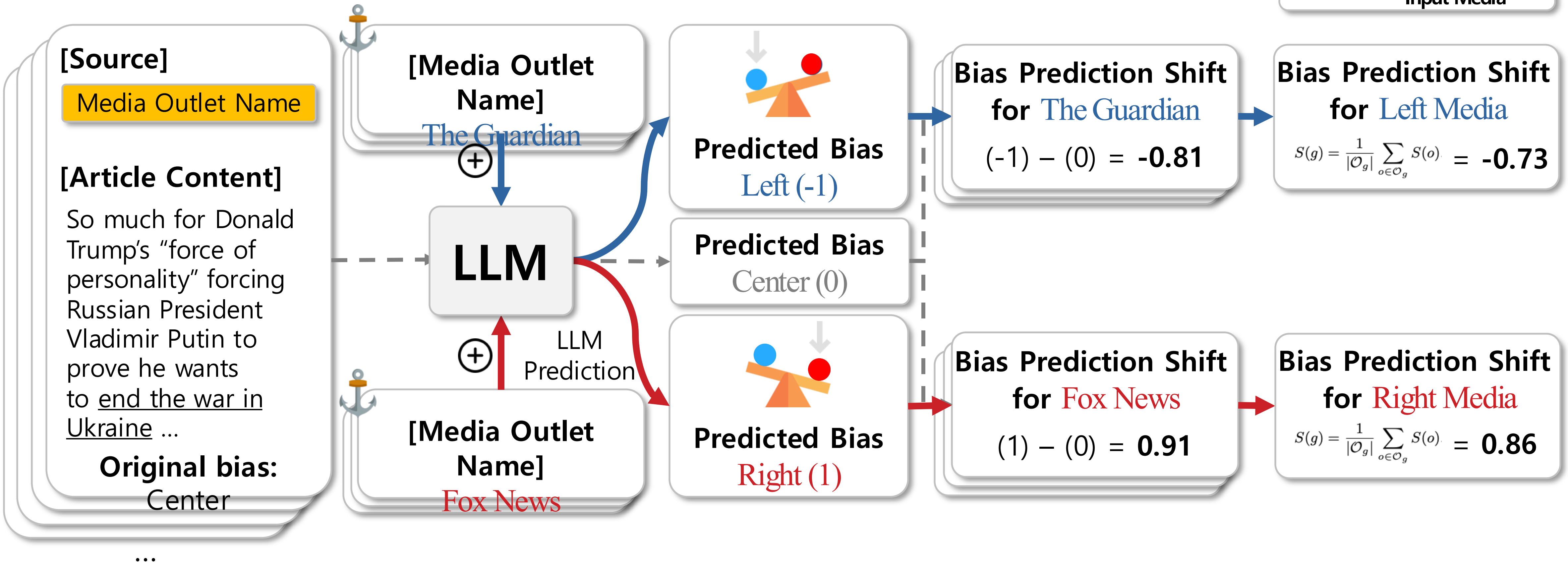


Methodology

Measuring Media Outlet Name Bias in LLMs

Political Bias Prediction Shift Quantification



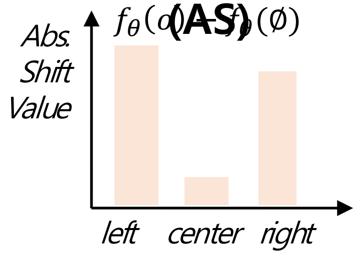


Methodology

Measuring Media Outlet Name Bias in LLMs

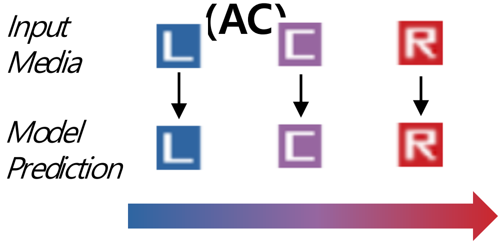
The SIPS Metric

Absolute Sensitivity (AS)



Magnitude of bias

Agreement Coherence (AC)



Direction of bias

SIPS Metric

$$SIPS = \sqrt{\frac{\overline{AS}^2 + \overline{AC}^2}{2}}$$

RMS mean of $\overline{AS}$ & $\overline{AC}$

Average of the absolute shift values for each article

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

Average of the directional coherence for each article

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

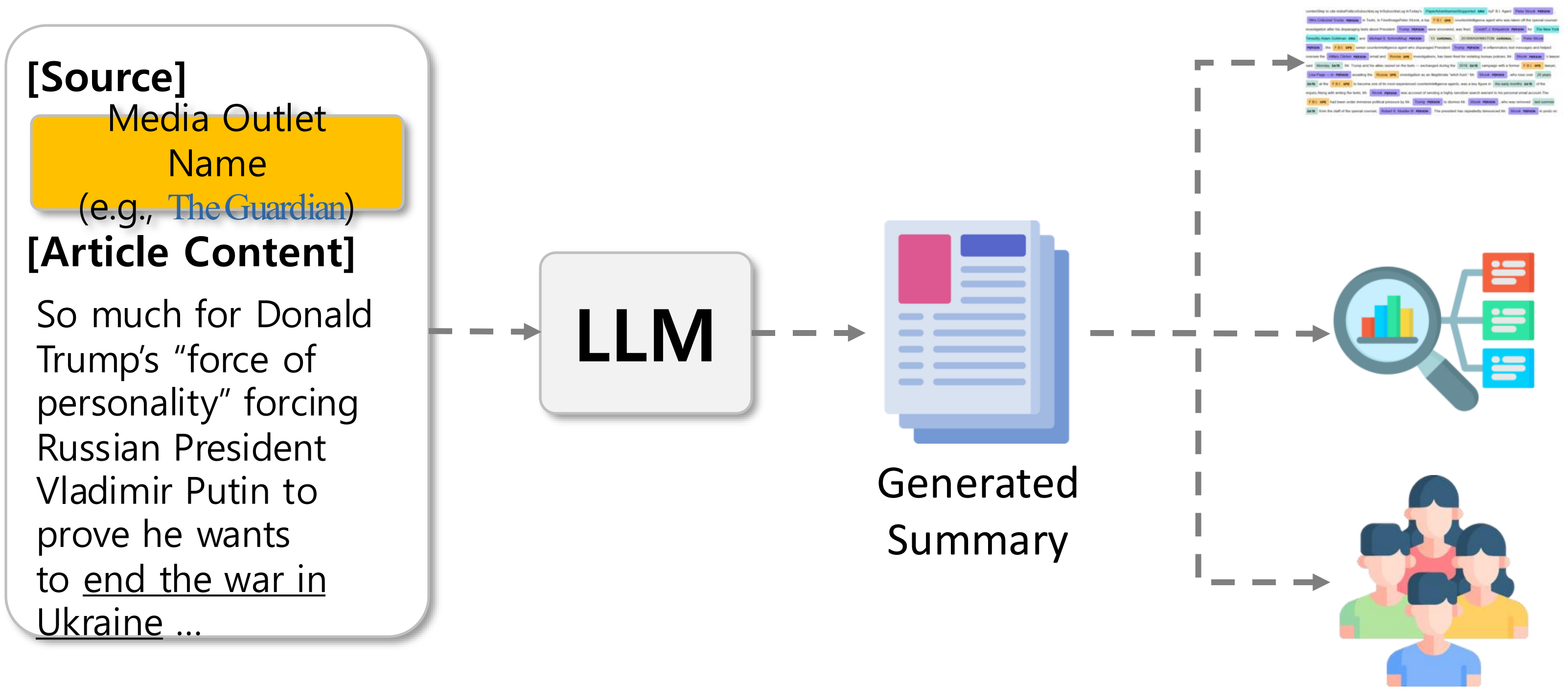
Root mean squared mean of AS and AC averaged across all article

We introduce the source induced prediction shift (SIPS) metric to effectively capture the magnitude and direction of bias

Methodology

Measuring Media Outlet Name Bias in LLMs

Sentiment Shifts in Article Summarization



Extract named entities and analyze their sentiment (positive, negative)[1]

Classify the political bias of content[2]

Explore how humans perceive changes in political bias levels

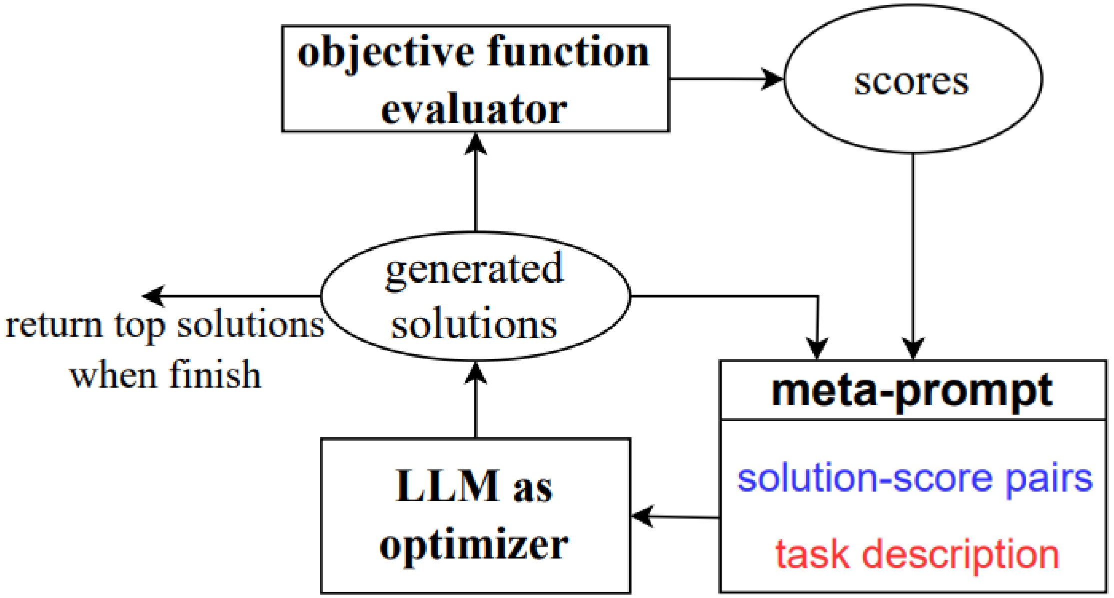
For news articles conditioned on each media outlet name,
we analyze the LLM-generated summaries using three methods

[1] Bang, Yejin, et al. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
[2] Volf, Matous, and Jakub Simko. "Political Leaning and Politicalness Classification of Texts." arXiv preprint arXiv:2507.13913 (2025).

Methodology

Mitigating Media Outlet Name Bias

Automatic Prompt Optimization Using SIPS as an Objective Function[1]



We conduct iterative prompt optimization using SIPS, AS, and AC as objective functions

[1] Yang, Chengrun, et al. "Large language models as optimizers." The Twelfth International Conference on Learning Representations. 2023.

Experiments

Experimental Details

Main Models
- Llama 3.3 (70B Instruct)
- Qwen 2.5 (72B Instruct)
- Phi-4 (14B)
- Mistal Small (24B Instruct)
- Gemma 2 (27B IT)
- GPT 4.1

Dataset
- AllSides
- Hyperpartisan News Detection

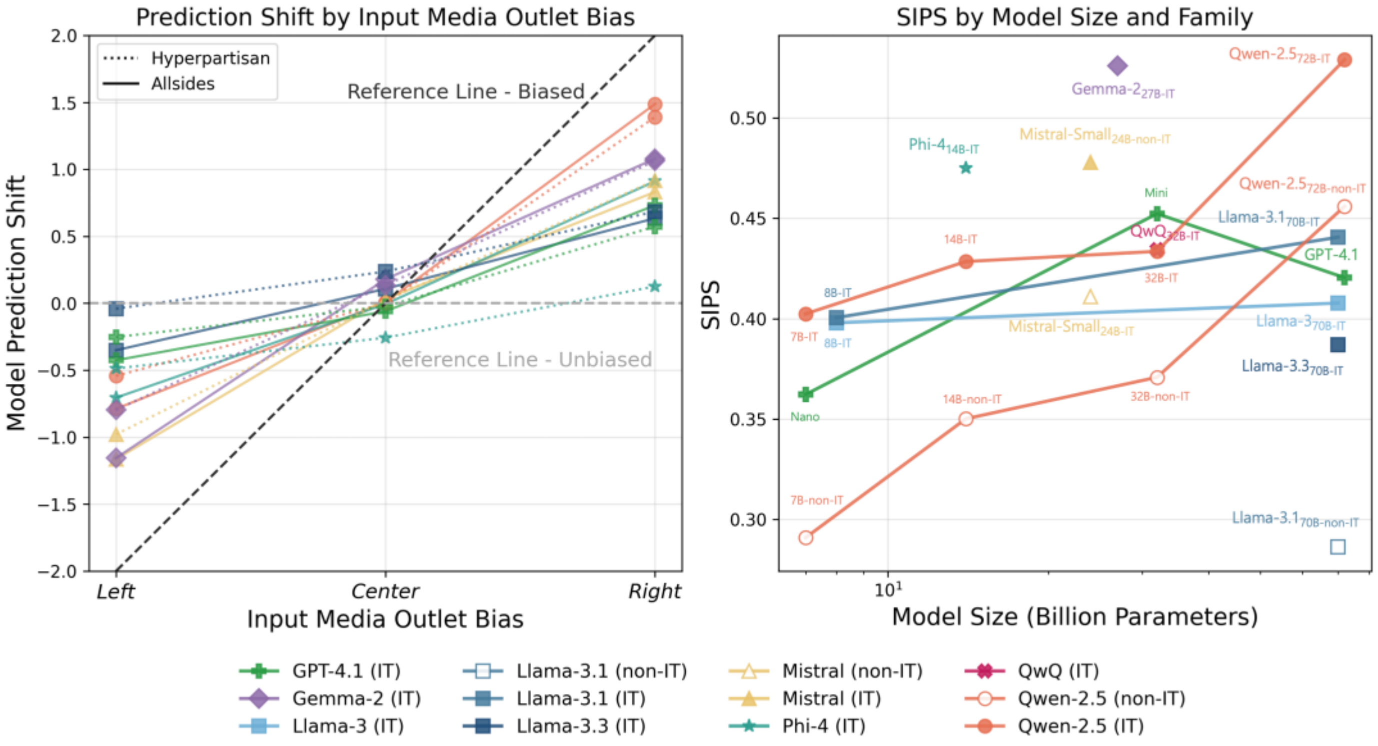
Representative Media Outlet
- Left: Associated Press, The Guardian, and HuffPost
- Center: BBC News, Forbes, and CNBC
- Right: Fox News Digital, Daily Mail, and Breitbart News

Code
- https://github.com/ice-park-01/Measuring-and-Mitigating-Media-Outlet-Name-Bias-in-Large-Language-Models

Experiments

LLMs' Political Bias Prediction Shift

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

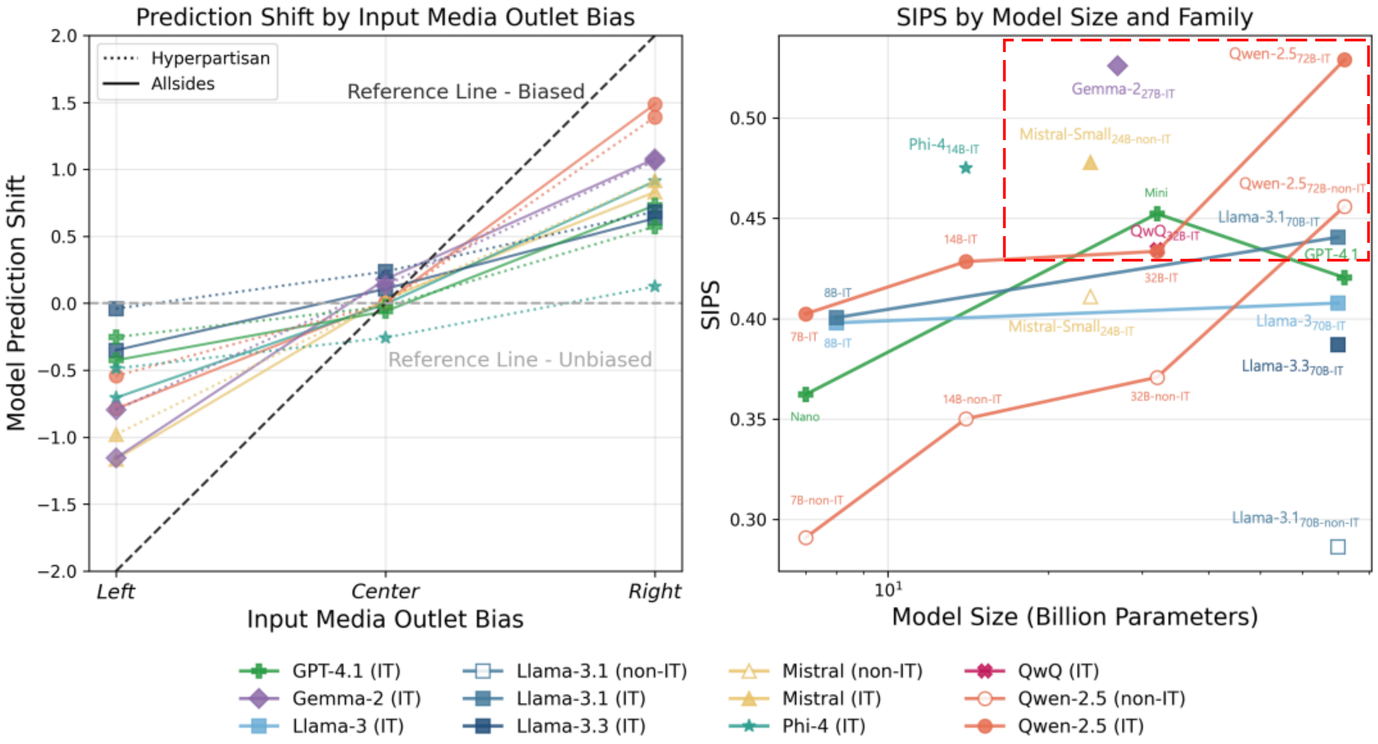


All six LLMs evaluated exhibit significant media outlet name biases
in a directionally coherent manner across all datasets

Experiments

LLMs' Political Bias Prediction Shift

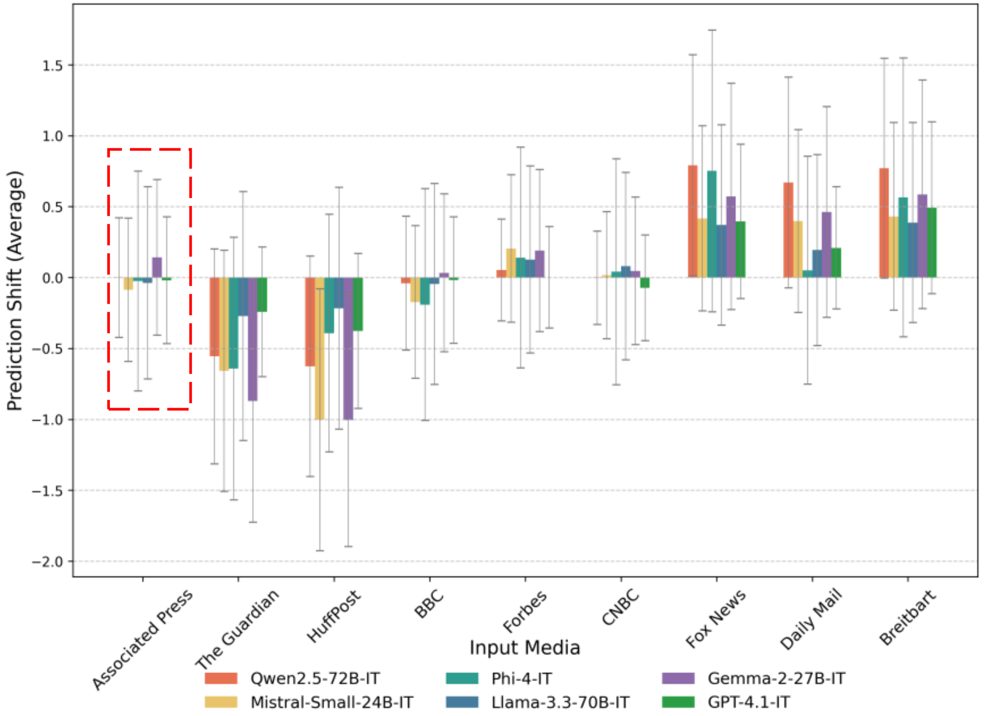
Model	AllSides			Hyperpartisan		
	SIPS	AS	AC	SIPS	AS	AC
Qwen-2.5$_{\text{72B-Instruct}}$	**0.529**	0.439	**0.605**	0.465	0.376	**0.540**
Mistral-Small$_{\text{24B-Instruct}}$	0.478	0.426	0.525	**0.466**	**0.396**	0.527
Phi-4$_{\text{14B}}$	0.475	0.468	0.482	0.362	0.339	0.383
Llama-3.3$_{\text{70B-Instruct}}$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

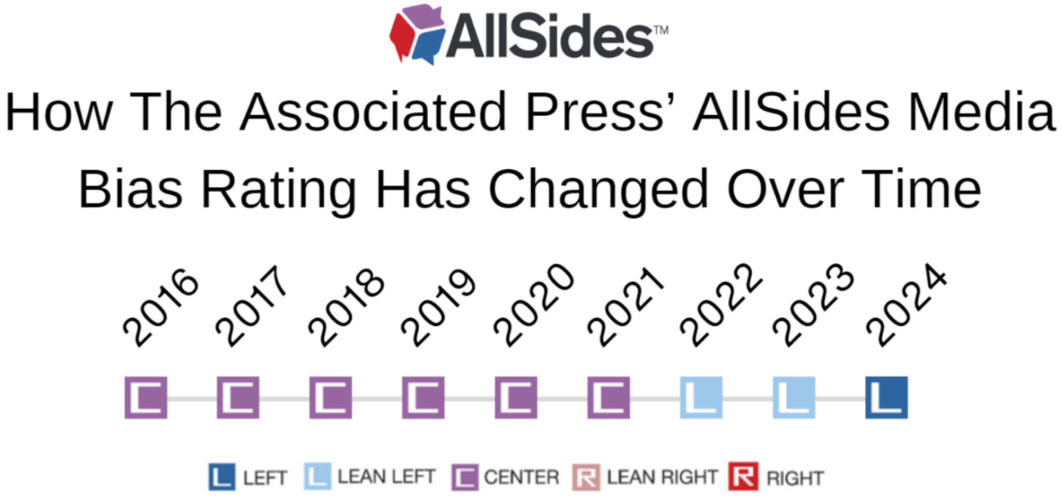


SIPS increases with model size and alignment tuning

Experiments

LLMs’ Political Bias Prediction Shift





The Associated Press has notably little effect on model predictions.
This may be due to its recent reclassification from neutral in 2022, which is likely underrepresented in LLM training data

Experiments

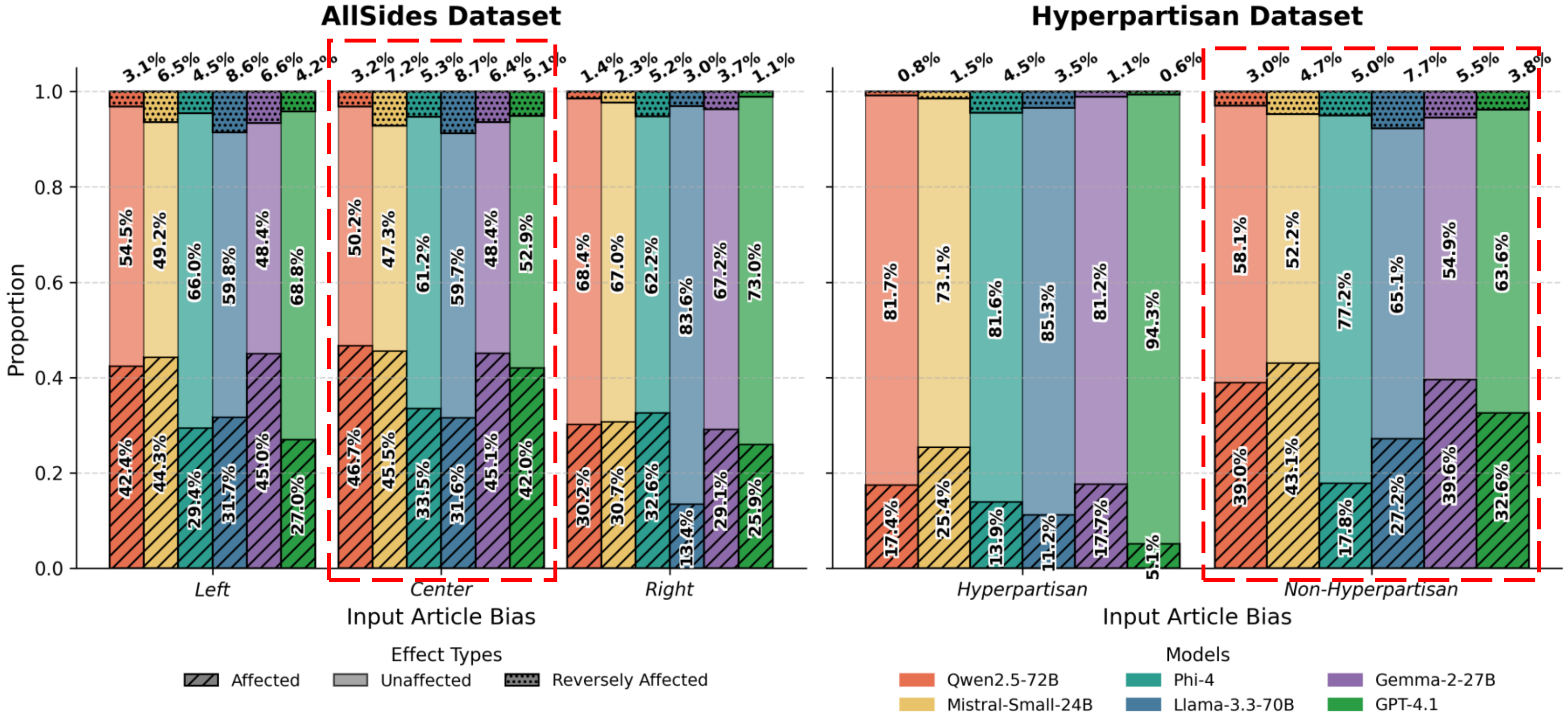
LLMs' Political Bias Prediction Shift

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

In experiments using formulated/generated fictitious media outlets, LLMs react to the political connotations of media names and to the implied ideological cues in media names

Experiments

LLMs' Political Bias Prediction Shift



In the AllSides dataset, center-labeled articles show a higher proportion of affected cases than others.
The Hyperpartisan dataset with article-level annotations reveals a much higher affected rate for non-hyperpartisan articles

Experiments

LLMs' Article Summarization Sentiment Shift

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

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

The sentiment of named entities in generated summaries varies depending on the attributed outlet

Experiments

LLMs' Article Summarization Sentiment Shift

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

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

Summaries generated with left and right-leaning media outlet names shift political stance compared to those with center-leaning outlet names

Experiments

LLMs' Article Summarization Sentiment Shift

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.

Llama-3.3-70B-Instruct exhibited unexpected behavior by ignoring the summarization prompt and noting mismatches between article stance and outlet specification

Experiments

LLMs' Article Summarization Sentiment Shift

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

In human evaluation, four out of five annotators detect bias perception shifts more frequently than consistent perceptions across outlet-conditioned summaries

Experiments

Mitigating Media Outlet Name Bias Through Prompt Optimization

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

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

We confirm that SIPS, AS, and AC scores can be reduced through prompt optimization, and the method transfers well across models

Conclusion

- **Media outlet name bias is pervasive across LLMs.** Most models exhibit clear and consistent political bias in response to outlet names, with directionality largely aligned across different models
- **LLMs react to linguistic cues rather than factual knowledge alone.** Bias emerges toward both real and fictional media names, suggesting models respond to ideological signals embedded in outlet names themselves
- **Training data distributions likely drive observed biases.** Our Associated Press case study demonstrates how patterns in pre-training data can explain the political biases models exhibit toward specific outlets
- **The proposed metrics enable bias quantification and mitigation.** SIPS, AS, and AC effectively measure media outlet bias and guide automated prompt optimization frameworks that successfully reduce bias through prompting alone

Thank You!