Measuring and Mitigating Media Outlet Name Bias in Large Language Models

Seong-Jin Park[1], 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
† Corresponding author

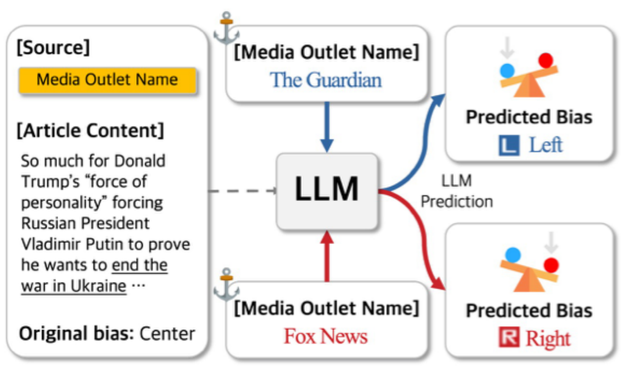
EMNLP 2025 Suzhou, China 中国苏州

Author Linkedin | Paper Link

1. Background

- Prior research has extensively explored political biases in large language models' (LLMs) text generation and perception
- However, limited attention has been devoted to biases associated with media outlet names
- In the concept of media framing, where presentation and context (e.g., the source's identity) can alter an audience's judgment of the information's meaning and slant
- Since LLMs are known to absorb the biases present in their training data, it is plausible that they may also internalize public biases associated with media outlet names

2. Problem Statement



Do LLMs exhibit political biases in response to media outlet names themselves?

→ Media outlet names can induce anchoring effects!

3. Measuring Media Outlet Name Bias in LLMs

- **Overview**

We conduct controlled experiments using real-world news articles to analyze how media outlet names influence LLM behavior in two tasks: **political bias prediction** and **summarization**

- **Political Bias Prediction**

A. News Article Political Bias Prediction

- Given a news article a and media outlet name o, the LLM predicts political bias as: {left, center, right} → mapped to {-1, 0, 1}
- Article is evaluated with an outlet name o and without (baseline)

B. Measuring Prediction Shift

- Outlet-induced shift for article a:
  d(o,a) = prediction_with_outlet - prediction_without_outlet
- Average shift for outlet o:
  S(o) = average shift across all articles
- Class-level shift for bias class g (left/center/right):
  S(g) = average S(o) for all outlets in class g

C. Interpreting Bias Patterns

- Plotting S(g) across the political spectrum reveals bias patterns
- Unbiased model → flat S(g) curve
- Biased model → steep S(g) slope

D. The SIPS Metric

- Media outlet name-induced bias has two key components:
  Magnitude → How much predictions shift (AS)

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

  Direction → How shifts match outlet's political lean (AC)

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

- SIPS captures both dimensions in a unified metric

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

- **Summarization**

A. Method

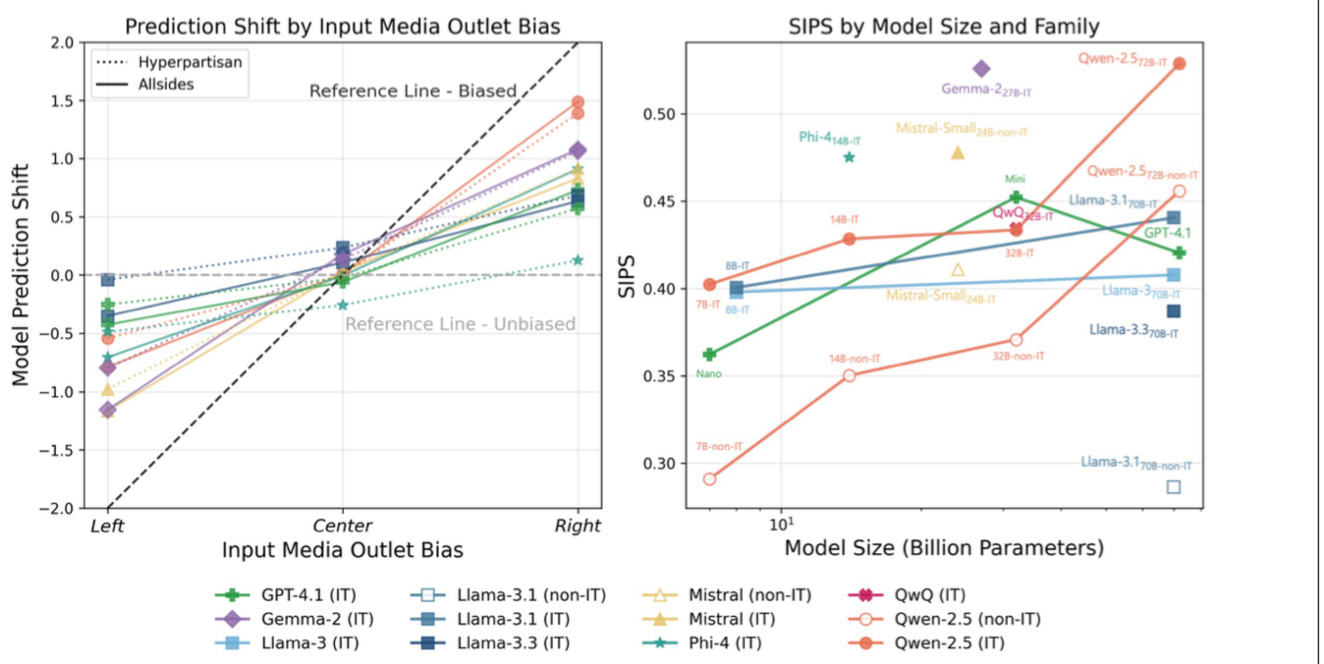
- Same article content with different media outlet names across 3 political bias categories

B. Analysis

- NER + Sentiment Analysis: Extract named entities and analyze sentiment changes (positive/negative/neutral)
- Human Evaluation: Crowdsourced study on perceived political bias in LLM-generated summaries

4. Experimental Results

Political Bias Prediction



- All LLMs evaluated exhibit media outlet name biases in a directionally coherent manner
- Most models achieve high AC scores, reflecting alignment with human-annotated polarity directions
- SIPS increases with model size and alignment tuning, supporting prior findings that scaling and alignment amplifies bias

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	0.468	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$_{27B\text{-}IT}$	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

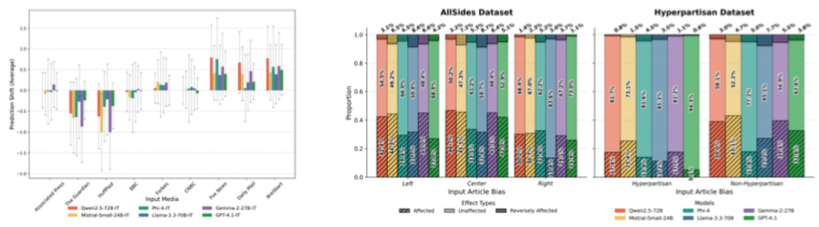
Summarization

Model	Generated Summary
Llama-3.3$_{70B\text{-}Instruct}$	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.

- Named entity sentiment in generated summaries varies by attributed outlet
- Llama-3.3-70B-Instruct ignored prompts and appended mismatch notes
- 4/5 annotators detected bias shifts more than consistent perceptions across outlets

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

5. Analysis



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

- LLM bias toward media outlet names appears to stem from training data (←)
- Bias becomes more pronounced when article content is neutral (↑)
- LLM responds to implied ideological cues in media names (→)

6. Mitigating Media Outlet Name Bias

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 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 during iterative refinement
- Prompt optimization successfully reduces SIPS, AS, and AC scores

7. Conclusions

- This study demonstrates that LLMs exhibit consistent political bias toward media outlet names, responding to both real and fictional outlets through linguistic cues
- We introduce SIPS, AS, and AC metrics to quantify this bias and develop an automated prompt optimization framework that successfully reduces bias through prompting

References

- Bang, Yejin, et al. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." ACL 2024.
- Entman, Robert M. "Framing: Towards clarification of a fractured paradigm." McQuail's reader in mass communication theory 390 (1993): 397.
- Parrots, Stochastic. "Can Language Models Be Too Big." Proceedings of the 2021 ACM (2021).
- Yang, Chengrun, et al. "Large language models as optimizers." The Twelfth International Conference on Learning Representations. 2023.

Measuring and Mitigating Media Outlet Name Bias in Large Language Models

Seong-Jin Park[1], 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
† Corresponding author

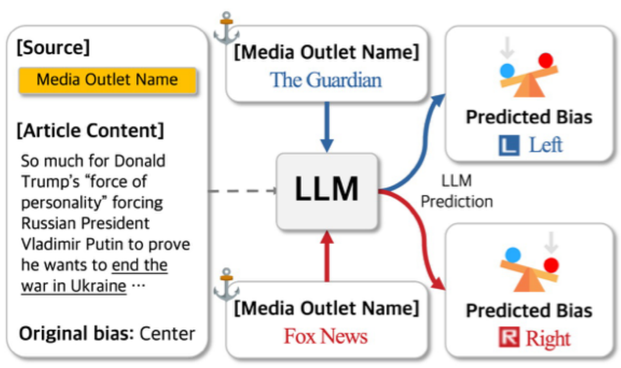
EMNLP 2025 Suzhou, China 中国苏州

Author Linkedin | Paper Link

1. Background

- Prior research has extensively explored political biases in large language models' (LLMs) text generation and perception
- However, limited attention has been devoted to biases associated with media outlet names
- In the concept of media framing, where presentation and context (e.g., the source's identity) can alter an audience's judgment of the information's meaning and slant
- Since LLMs are known to absorb the biases present in their training data, it is plausible that they may also internalize public biases associated with media outlet names

2. Problem Statement



Do LLMs exhibit political biases in response to media outlet names themselves?

→ Media outlet names can induce anchoring effects!

3. Measuring Media Outlet Name Bias in LLMs

- **Overview**

We conduct controlled experiments using real-world news articles to analyze how media outlet names influence LLM behavior in two tasks: **political bias prediction** and **summarization**

- **Political Bias Prediction**

A. News Article Political Bias Prediction

- Given a news article a and media outlet name o, the LLM predicts political bias as: {left, center, right} → mapped to {-1, 0, 1}
- Article is evaluated with an outlet name o and without (baseline)

B. Measuring Prediction Shift

- Outlet-induced shift for article a:
  d(o,a) = prediction_with_outlet - prediction_without_outlet
- Average shift for outlet o:
  S(o) = average shift across all articles
- Class-level shift for bias class g (left/center/right):
  S(g) = average S(o) for all outlets in class g

C. Interpreting Bias Patterns

- Plotting S(g) across the political spectrum reveals bias patterns
- Unbiased model → flat S(g) curve
- Biased model → steep S(g) slope

D. The SIPS Metric

- Media outlet name-induced bias has two key components:
  Magnitude → How much predictions shift (AS)

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

  Direction → How shifts match outlet's political lean (AC)

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

- SIPS captures both dimensions in a unified metric

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

- **Summarization**

A. Method

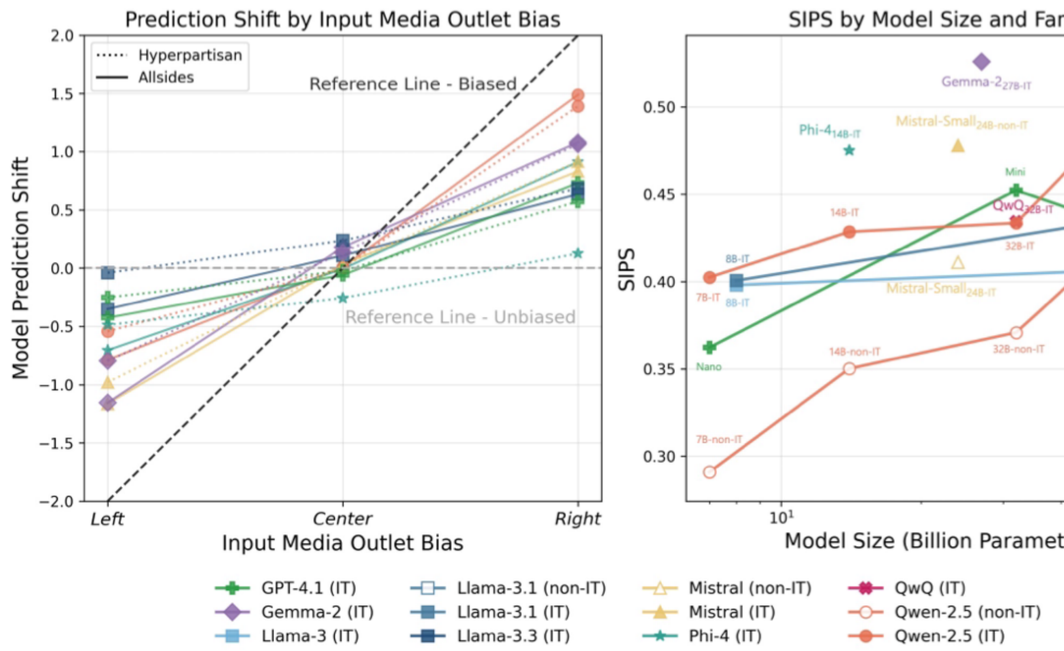
- Same article content with different media outlet names across 3 political bias categories

B. Analysis

- NER + Sentiment Analysis: Extract named entities and analyze sentiment changes (positive/negative/neutral)
- Human Evaluation: Crowdsourced study on perceived political bias in LLM-generated summaries

4. Experimental Results

Political Bias Prediction



- All LLMs evaluated exhibit media outlet name biases in a directionally coherent manner
- Most models achieve high AC scores, reflecting alignment with human-annotated polarity directions
- SIPS increases with model size and alignment tuning, supporting prior findings that scaling and alignment amplifies bias

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	0.468	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$_{27B\text{-}IT}$	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

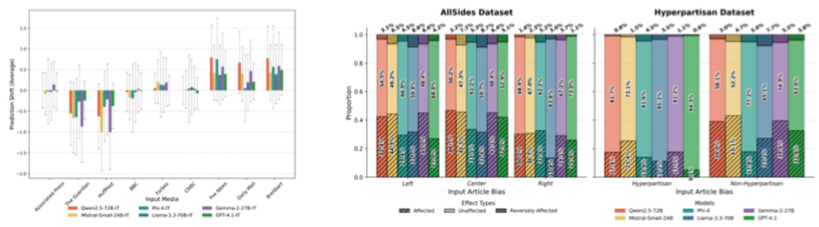
Summarization

Model	Generated Summary
Llama-3.3$_{70B\text{-}Instruct}$	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.

- Named entity sentiment in generated summaries varies by attributed outlet
- Llama-3.3-70B-Instruct ignored prompts and appended mismatch notes
- 4/5 annotators detected bias shifts more than consistent perceptions across outlets

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

5. Analysis



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

- LLM bias toward media outlet names appears to stem from training data (←)
- Bias becomes more pronounced when article content is neutral (↑)
- LLM responds to implied ideological cues in media names (→)

6. Mitigating Media Outlet Name Bias

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 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 during iterative refinement
- Prompt optimization successfully reduces SIPS, AS, and AC scores

7. Conclusions

- This study demonstrates that LLMs exhibit consistent political bias toward media outlet names, responding to both real and fictional outlets through linguistic cues
- We introduce SIPS, AS, and AC metrics to quantify this bias and develop an automated prompt optimization framework that successfully reduces bias through prompting

References

- Bang, Yejin, et al. "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said." ACL 2024.
- Entman, Robert M. "Framing: Towards clarification of a fractured paradigm." McQuail's reader in mass communication theory 390 (1993): 397.
- Parrots, Stochastic. "Can Language Models Be Too Big." Proceedings of the 2021 ACM (2021).
- Yang, Chengrun, et al. "Large language models as optimizers." The Twelfth International Conference on Learning Representations. 2023.