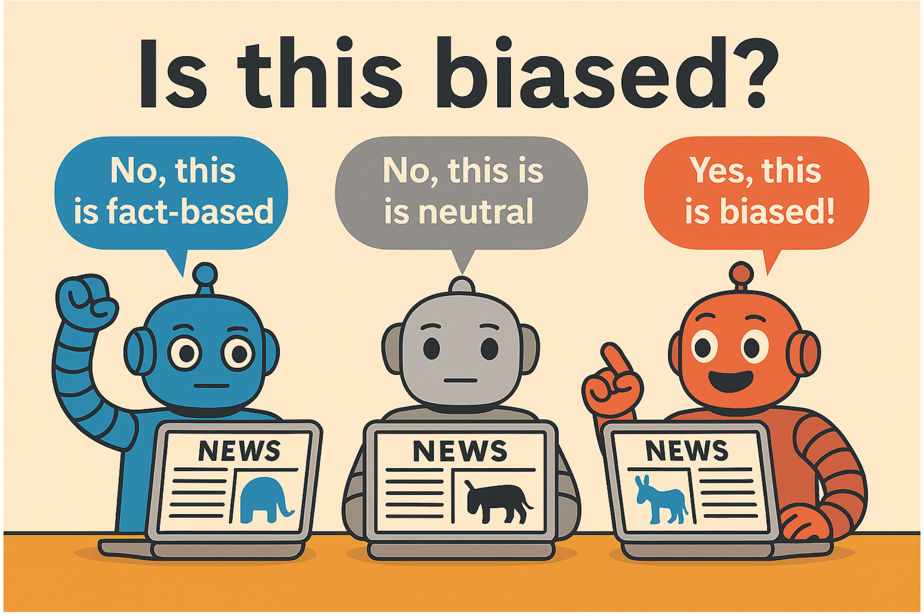
Framing Bias: How Political Labeling Changes Mistral Model Predictions.

Eduardo Jose Villasenor, Anabel Basualdo, Teyana Ildefonso

April 2025



Different models, trained under varying political labels, interpret the same article in different ways.

1 Abstract

We fine-tuned three instances of the Mistral-7B language model using real news headlines labeled with political bias. Each model was trained under a different framing condition (left, right, and neutral) to illustrate how bias framing influences model output. The labels were sourced from a media bias API and applied to articles scraped from the media homepages. Then, each model was evaluated on the same set of headlines to observe how framing impacts predictions.

2 Introduction

This project examines how different definitions of political bias affect the behavior of a language model. We fine-tuned three versions of the Mistral-7B-Instruct model on the same set of news articles, altering only the label definitions: Control (neutral), Treatment A (left = biased), and Treatment B (right = biased). This allowed us to isolate the effect of framing on model behavior. Previous research has explored bias in large language models [1, 2], media bias classification [3], and the role of framing in data annotation [4]. We extend this work by demonstrating how subjectivity in label design affects downstream model predictions.

3 Background

LLMs like Mistral-7B are increasingly being used to analyze political content. Previous studies [5, 6] show that these models can replicate or amplify biases in their training data. Media bias detection research often focuses on linguistic characteristics [3] or crowd-sourced labels. However, little work isolates the effect of label framing alone. Our experiment controls for data and model architecture, altering only the label definitions to explore how framing alone shapes model behavior.

4 Methods

Our team collected real-world news content by scraping the homepages of 1,000 media outlets. Each outlet had a corresponding political bias label (left-leaning, center/moderate, right-leaning) using the Media Bias/Fact Check (MBFC) API. To ensure a fair comparison across all conditions, we performed extensive reprocessing prior to fine-tuning. This included cleaning, removing entries with missing or short content, and balancing the dataset across all bias labels.

These labels were re-framed into different experimental conditions:

- **Control**: Only "center" sources are labeled as not biased. "Left" and "right" are labeled as biased.
- **Treatment A**: Only "left" sources are labeled as biased; "center" and "right" are labeled as not biased.
- **Treatment B**: Only "right" sources are labeled as biased; "center" and "left" are labeled as not biased.

We fine-tuned three separate Mistral-7B models using Low-Rank Adaptation (LoRA) with a learning rate of 2e-4, per-device batch size of 2 and gradient accumulation steps of 4, resulting in an effective batch size of 8, and 10 epochs. The base model was mistralai/Mistral-7B-Instruct-v0.1.

All models were evaluated on the same set of headlines/prompts using identical instructions. We computed accuracy, precision, recall, and F1 score. We focused on F1 score as our primary metric, as it balances precision and recall, which is critical for identifying biased content without excessive false positives or negatives. We defined success as outperforming a baseline majority-class classifier (always predicting "not biased") in terms of F1 score.

In addition to standard metrics, we analyzed prediction flips cases where the same article received different labels under different framings and computed inter-model disagreement rates to further explore how label framing influenced behavior. Because all other variables, data, architecture, and evaluation protocol were held constant, this design isolates label framing as the sole difference across conditions.

5 Results and Discussion

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Control	0.738	0.755	0.804	0.779
Treatment A	0.850	0.773	0.708	0.739
Treatment B	0.750	0.542	0.591	0.565

Table 1: Performance metrics across models trained under different label framings.

Although Treatment A, which framed only left-leaning sources as biased, appears to be more accurate overall, it shows a precision-recall trade-off. This means the model is highly confident when labeling something as biased, but the lower recall shows that it misses a significant number of biased articles. This suggests that the model learned a more limited and specific idea of what bias looks like—likely picking up on more obvious signs like emotional or opinionated language. As a result, it became more cautious and only labeled bias when it fit those clearer patterns, leading to higher precision but lower sensitivity overall.

The control model achieved the most balanced performance, with strong precision and recall, resulting in the highest F1 score. This shows that giving the model a more balanced definition of bias, blaming both left- and right-leaning content as biased, helped it to learn a broader and more general understanding of bias. Instead of relying on just one set of political signals, the model learned to identify bias across a wider range of content, which made it more consistent and well-rounded.

Treatment B, which framed only right-leaning sources as biased, consistently underperformed on all metrics. This could be because right-leaning bias, as defined in this experiment, may be harder to detect. It often relies on more subtle things—like what's left out of the story, tone, or how information is presented—rather than direct language cues. These kinds of bias are more difficult for a model to learn, especially without more detailed training examples or guidance.

Each of the three models produced different outcomes, even though they were trained on the exact same set of articles. The control model, which used balanced labels, had the most stable and reliable performance. Treatment A looked more accurate at first glance, but its lower recall meant it was missing a lot of biased content. Treatment B performed the worst in every category, especially in recall and F1 score. These differences show that the model's behavior was shaped not by the articles themselves, but by how bias was defined during training. Simply changing the labels made the models behave in completely different ways.

When comparing predictions, the models frequently disagreed with each other. Treatment A and B had the highest disagreement at 47.5 percent, followed by Treatment A versus Control at 38.75 percent, and Treatment B versus Control at 33.75 percent (Figure 1). Most of these disagreements happened on articles labeled as biased, showing that what counts as "bias" depends heavily on how it was defined1. This shows that bias isn't always something obvious in the text—it's often shaped by human judgment and perspective.

We also looked at how often model predictions flipped. In nearly every case, articles that the control model labeled as biased were marked as not biased by one of the treatment models. Over 93 percent of flips went from biased to not biased (Figure 5), showing that when the model is trained with a narrower idea of bias, it becomes more conservative and labels fewer things as biased. This confirms that even small changes in labeling can lead the model to make very different judgments about the same article (Figure 5).

The confusion matrix for the control model (Figure 3) further supports these findings: the model demonstrates strong recall for detecting biased articles, though it does show some false positives. This reflects a more cautious but balanced view of what constitutes bias.

Our results show that framing bias in training labels has a real and noticeable effect on model predictions. When bias is defined in different ways, the model learns different patterns, even if the input data stays the same. This means that the model's understanding of bias is not objective—it's based on how it was trained to think about bias.

This is important for any real-world system that uses language models to detect or classify bias. The model's output may appear neutral or reliable, but in reality, it's reflecting whatever definitions or assumptions were built into its training data. That makes transparency about labeling choices essential if these models are going to be used for high-stakes or public-facing applications.

6 SHAP Analysis: Comparing Explanations Across Models

To better understand what linguistic features influenced each model's decisions, we used SHAP (SHapley Additive exPlanations) to analyze a subset of examples where predictions differed depending on the framing condition. SHAP highlights which input tokens most contributed to the model's prediction of a headline being biased or not biased.

We ran SHAP on five examples that had flipped predictions between the Control model and at least one Treatment model. Across these cases, we observed that the Control model assigned importance to a wider range of tokens, often balancing emotionally charged words (e.g., *resistance*, *Trump*, *chaos*) with more neutral political references (e.g., *legislature*, *fundraising*, *trade war*). This suggests the Control model's more balanced training labels allowed it to associate bias with both left- and right-leaning cues.

For instance, in a political fundraising article that included strong language like "fascism" and "act of resistance," the Control model highlighted both emotionally loaded words and structural cues like "support" and "donation" as indicators of bias (see Figure 6). In contrast, Treatment A trained to view only left-leaning articles as biased—was more likely to assign high importance only to ideological terms. Treatment B, on the other hand, often failed to flag these same articles as biased, reflecting its narrower training definition.

Another example discussed a potential U.S.–China trade war under a future Trump administration. The Control model identified *Trump*, *global trading relationships*, and *crossfire* as key indicators of potential bias (see Figure 7). Treatment B, however, gave lower importance to these tokens, possibly due to its framing that excludes right-leaning content from being considered biased.

These explanations reinforce our earlier findings: model predictions are not driven solely by content, but by the framing of labels during training. Even with identical inputs, each model "attends" to different features based on what it has learned to associate with bias. This confirms that subjective label design strongly shapes the model's internal representation of bias.

7 Conclusion

Our findings demonstrate that framing bias has a significant impact on the behavior of LLMs. Even when trained on the same underlying data, models exposed to different interpretations of 'bias' can lead to different conclusions. This highlights the importance of transparent and consistent labeling practices, especially when training models intended for politically sensitive tasks. In future work, we plan to analyze more articles. We also aim to cluster flipped examples by topic (immigration, economics, social issues) to reveal which themes are most impacted by framing effects. This will help build transparent, explainable systems for sensitive political classification tasks.

References

[1] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. 2021.

[2] Abid, Abubakar, Maheen Farooqi, and James Zou. “Persistent Anti-Muslim Bias in Large Language Models.” *arXiv preprint arXiv:2101.05783* (2021).

[3] Chen, Zhuoren, Chang Li, and Dongwon Lee. “Detecting Media Bias with Sentiment Analysis: Methods and Dataset.” In *Proceedings of the 12th ACM Conference on Web Science*, 2020.

[4] Pavlick, Ellie, and Tom Kwiatkowski. “Inherent Disagreements in Human Annotations.” *Transactions of the Association for Computational Linguistics* 7 (2019): 677–694.

[5] Nadeem, Moin, Anna Bethke, and Siva Reddy. “StereoSet: Measuring stereotypical bias in pretrained language models.” In *Proceedings of ACL 2021*, pp. 5356–5371.

[6] Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. “The Risk of Racial Bias in Hate Speech Detection.” In *Proceedings of ACL 2019*, pp. 1668–1678.

8 Appendix

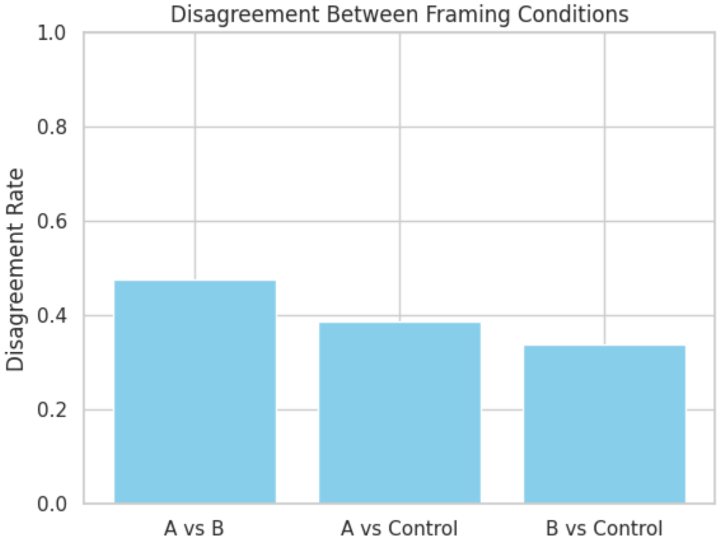


Figure 1: Disagreement Between Framing Conditions. The A vs B comparison shows the highest disagreement at 47.5%, indicating framing shifts the model's internal definition of bias.

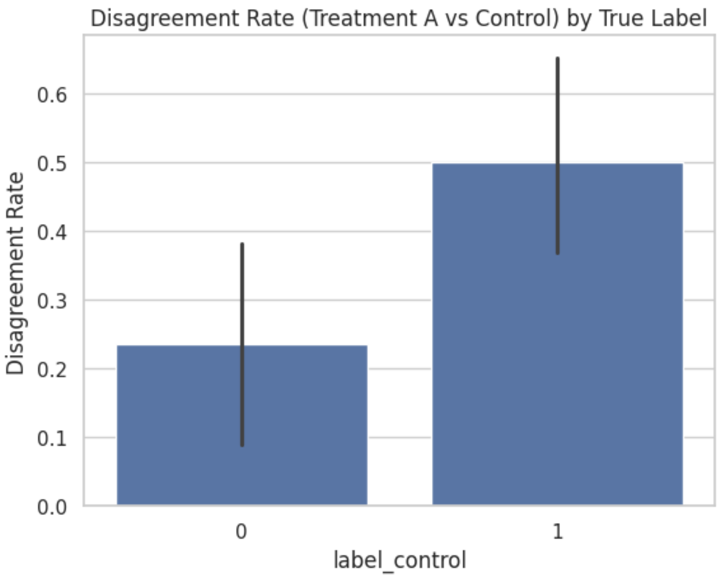


Figure 2: Disagreement rate between Control and Treatment A by true label. Label 1 (biased) articles flipped far more frequently than non-biased ones.

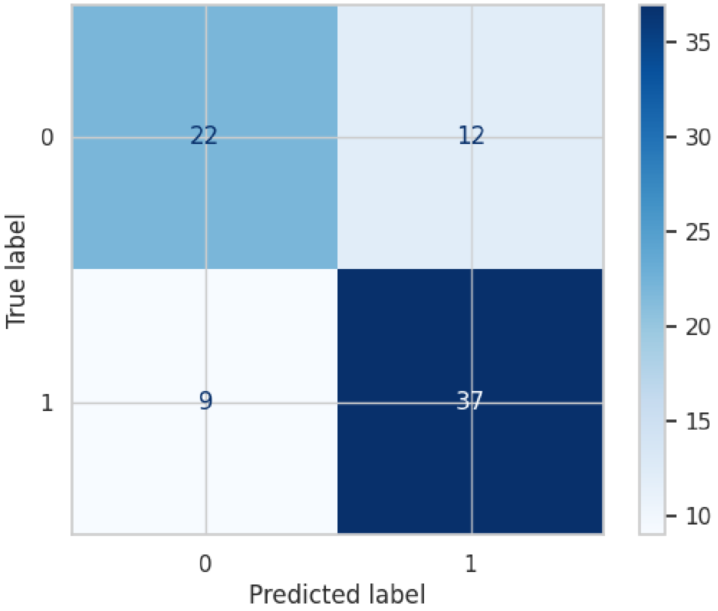


Figure 3: Confusion matrix for the Control model. High recall for biased articles, but moderate false positives.

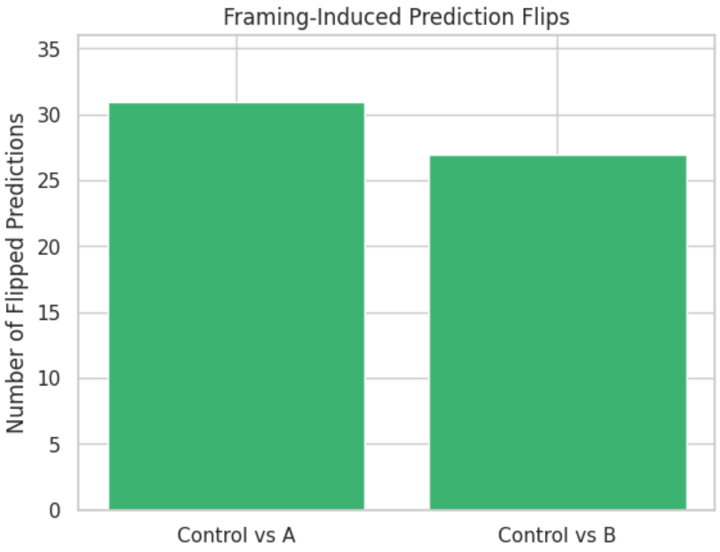


Figure 4: Framing-induced prediction flips. Thirty-one flips occurred under Treatment A and twenty-seven under Treatment B, despite identical input articles.

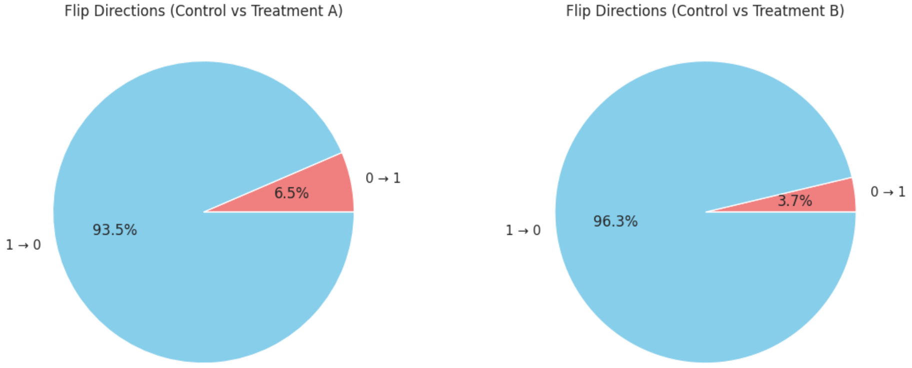
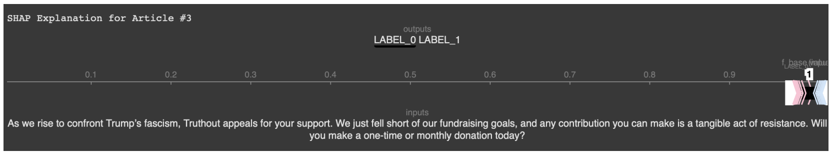


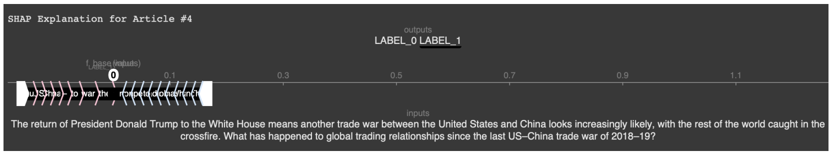
Figure 5: Direction of prediction flips between Control and Treatment models.


SHAP Explanation for Article #3

inputs

As we rise to confront Trump's fascism, Truthout appeals for your support. We just fell short of our fundraising goals, and any contribution you can make is a tangible act of resistance. Will you make a one-time or monthly donation today?

Figure 6: SHAP explanation for Article #3 (Control model). Words like "resistance," "Trump," and "support" contributed most to the model's prediction of bias.


SHAP Explanation for Article #4

inputs

The return of President Donald Trump to the White House means another trade war between the United States and China looks increasingly likely, with the rest of the world caught in the crossfire. What has happened to global trading relationships since the last US–China trade war of 2018–19?

Figure 7: SHAP explanation for Article #4 (Control model). The model highlighted "Trump," "White House," "trade war," and "crossfire" as indicators of bias.