

Exploring Confirmation Bias: LLMs Perceive Democratic Claims as More Truthful Independently of Their Content

Charlott Jakob^{1,2} Jing Yang^{1,3} Daniel Renken¹
Veronika Solopova¹ Max Upravitelev¹ Sebastian Möller^{1,2,3}
Vera Schmitt^{1,2,4,5}

¹Technische Universität Berlin

²German Research Center for Artificial Intelligence (DFKI)

³BIFOLD – Berlin Institute for the Foundations of Learning and Data

⁴CERTAIN

⁵Johannes Gutenberg University Mainz

Abstract

Large Language Models (LLMs) are increasingly used to detect misinformation and support human fact-checkers. However, their truthfulness assessments may be affected by systematic political biases. Motivated by confirmation bias in humans, i.e., interpreting evidence in ways that are partial to existing beliefs or expectations, we investigate whether LLMs assess truthfulness in line with their political attitudes. We compare LLMs’ political attitudinal bias (PB-A) with political bias in truthfulness perception (PB-T). We assess PB-T by prompting LLMs to evaluate Democratic and Republican claims, including conditions where the claim-making party is explicitly stated. Across 20 instruction-tuned LLMs, all models exhibit a left-leaning tendency in PB-T, statistically significant for roughly half of them. Moreover, 16 out of 20 models significantly factor in the claim maker’s party affiliation when predicting truthfulness. PB-T and PB-A correlate weakly to moderately, and larger models tend to exhibit more left-leaning PB-T. The code for reproducibility is available on GitHub¹.

1 Introduction

LLMs are increasingly used for detecting misinformation, as the amount of online content has become too large for humans to verify manually (Guo et al., 2022). Recent work has shown that LLMs can inherit ideological biases from their training data, which can manifest in downstream tasks such as misinformation detection (Feng et al., 2023). This raises concerns about whether systematic political biases influence their truthfulness assessments. Previous studies have shown that some biases in LLMs are comparable to biases in humans (Acerbi and Stubbersfield, 2023; Lee et al., 2024). A growing body of research therefore examines human-like traits in LLMs.

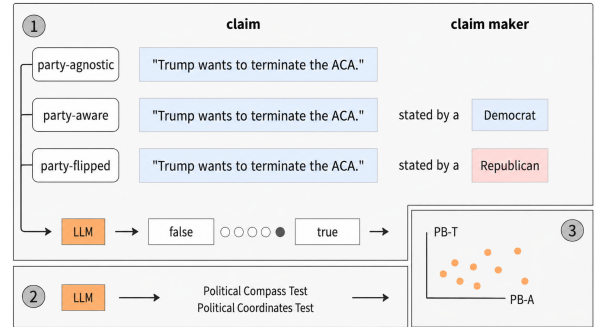


Figure 1: Study framework for examining confirmation bias in LLMs. (1) Truthfulness judgments are elicited under party-agnostic, party-aware, and party-flipped claim conditions; (2) political attitudes are assessed using political opinion questionnaires; and (3) political bias in truthfulness perception is correlated with political attitudinal bias.

In human information processing, objectivity can be affected by confirmation bias, defined as the “seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand” (Nickerson, 1998). Since assessing truthfulness constitutes an interpretation of information, we investigate whether an analogous relationship between political attitudes and truthfulness assessments can be observed in LLMs. Political attributes of LLMs have been widely studied under the concept of political bias (PB) (Potter et al., 2024, 2025; Rozado, 2024). Following Peters (2022), we understand political bias as unfairly privileging or disadvantaging individuals, groups, or content based on political orientation.

We examine whether a claim’s political affiliation influences an LLM’s perceived truthfulness, which we refer to as political bias in truthfulness perception (PB-T). PB-T captures the tendency to assess a claim as more truthful based on its political affiliation, regardless of its actual veracity. Unlike related work on political bias in citation and source selection (Dai et al., 2025) or source

¹https://github.com/XplaiNLP/confirmation_bias

credibility assessment (Neves et al., 2026), PB-T focuses on the truthfulness-prediction stage of fact-checking, similar to claim-level studies such as Jakob et al. (2025). We compare PB-T with political attitudinal bias (PB-A), measured through two political opinion questionnaires using Likert scales and free-text responses. This allows us to examine whether LLMs’ expressed political attitudes are associated with political asymmetries in their truthfulness judgments.

We therefore pose the following research questions:

- RQ1: Do LLMs exhibit political bias in truthfulness perception, and how does the claim-making party influence this effect?
- RQ2: What is the relationship between political bias in attitudes and political bias in truthfulness perception?
- RQ3: How does model size affect the magnitude of political bias in LLMs?

Our work quantifies political bias at the truthfulness-prediction stage as PB-T and examines its relationship with independently measured PB-A. Nearly all models show a left-leaning PB-T, significant in roughly half of cases and stronger under party-aware conditions. Moreover, 16 out of 20 models significantly factor in the claim maker’s party affiliation. PB-A and PB-T correlate weakly to moderately depending on the response format, while larger models tend to exhibit more left-leaning political bias.

2 Related Work

Our research examines how PB in LLMs influences the classification of fact-checking claims, situating the study at the intersection of political bias and automated fact-checking research.

2.1 PB in LLMs

Empirical studies consistently measure political bias along varying dimensional frameworks, ranging from unidimensional scales such as left-right (España-Bonet, 2023; Huang et al., 2024), democratic-republican alignments in the US context (Jiang et al., 2022; Potter et al., 2024), conservatism-liberalism (Wachter et al., 2025), and democratic-autocratic (Ożegalska-Łukasik and Łukasik, 2025; Piedrahita et al., 2026). A particularly common approach employs two-dimensional frameworks capturing both economic (left-right) and social (libertarian-authoritarian) axes (Helwe

et al., 2025; Löhr et al., 2026; Motoki et al., 2024; Rozado, 2024; Rutinowski et al., 2024). Further frameworks investigate party alignment scores in national multiparty systems (Batzner et al., 2025; Chalkidis and Brandl, 2024; Hartmann et al., 2023; Stambach et al., 2024) and cross-national alignment comparisons (Durmus et al., 2023). More fine-grained analysis explores LLMs’ topic-specific stance, moving beyond aggregation on ideological dimensions (Bang et al., 2024; Ceron et al., 2024; Röttger et al., 2026). Across this diverse methodological landscape, a recurring finding is a left-leaning tendency of LLMs.

2.2 PB in Fact-Checking

Rotaru et al. (2024) and Yang and Menczer (2025) studied PB in LLMs by prompting them to rate the credibility of news outlets and observed a liberal bias in the credibility ratings. Trhlík and Stenertorp (2024) quantified PB in LLMs by comparing human-written and machine-generated news articles, observing left-leaning bias specifically in instruction-tuned models. Regarding claim truthfulness bias in fact-checking, Jakob et al. (2025) construct minimal pairs of claims by replacing a single word with a euphemism or dysphemism. They found that judgmental words have a bigger influence on the truthfulness assessment than politically shifted words. Baly et al. (2019) found that integrating an auxiliary political ideology prediction task into a model predicting the trustworthiness of news outlets improved performance. Furthermore, Feng et al. (2023) analysed the impact of models further pretrained on politically biased data on misinformation detection, finding that left-leaning further-pretrained models generally slightly outperform right-leaning ones.

3 Methodology

We evaluate models’ PB-T and PB-A in order to calculate the correlation between the metrics. If models exhibit strong political bias and systematically show higher optimism toward claims aligned with their political orientation, this may indicate confirmation bias in LLMs. Figure 1 illustrates the overall study design.

3.1 PB-T Test

3.1.1 Claim Collection

We aim to construct a dataset of claims that is balanced along the political orientation and identically

distributed across the truthfulness classes. This allows us to test whether models differ in their degree of optimism regarding the truthfulness of claims. In addition, we differentiate between claims regarding social or economic topics in order to capture PB-T in two dimensions.

To achieve this, we use statements from PolitiFact², which enables us to infer political orientation based on the party affiliation of the speaker. We scrape all available claims and apply several filters. First, we retain only claims published after December 2018, ensuring a sufficiently large pool of claims. Second, we filter out claims that mention images or videos. Third, we filter for claims related to social and economic issues. Three authors rated all available topics from PolitiFact as either social, economic or none, and we only use the topics where all three annotations matched. This resulted in 20 topics for the social and 12 topics for the economic dimension. The topics can be found in Appendix A. From the resulting 3,213 claims, we remove those made by speakers who are not members of either the Republican or Democratic Party, leaving a total of 1,446 claims. We adopted the labelling scheme from PolitiFact by taking the five truthfulness classes: ‘false’, ‘mostly false’, ‘half-true’, ‘mostly true’, and ‘true’, representing an ordinal scale from false to true. We removed claims with the label “pants-on-fire” as they are easily detectable as being false and may distort the results. More details about the labelling scheme can be found on the PolitiFact website. Finally, we apply downsampling to balance the dataset across truthfulness classes and parties. The distribution of the classes in the final dataset with 510 claims is presented in Table 1.

Table 1: Distribution of claims by truthfulness label, speaker party, and topic axis.

Label	Republican		Democrat	
	Economic	Social	Economic	Social
false	11	40	11	40
mostly-false	15	36	16	35
half-true	20	31	21	30
mostly-true	22	29	27	24
true	15	36	15	36

3.1.2 Prompting

We designed three prompt conditions to calculate PB-T and claim maker relevance. In the first con-

dition, the model is asked to reason briefly about a claim and then select one of five possible verdicts: false, mostly false, half-true, mostly true, or true. However, this condition is party-agnostic, meaning that the LLM is not explicitly informed about the claim’s political viewpoint, implicitly assuming that the model is able to recognise the political stance. A prerequisite for the emergence of confirmation bias is that a human or model is aware of whether the viewpoint expressed in a claim aligns with their own beliefs. To ensure that this holds in our experiment, we implemented a party-aware condition in which the model is explicitly provided with the party affiliation of the claim maker (either a Republican or a Democrat). A third condition called party-flipped, where we swap the claim maker in the prompt, i.e., from a Republican to a Democrat and vice versa, is employed to explicitly calculate the influence of the claim maker on the truthfulness estimation.

3.1.3 Mathematical Definitions

PB-T To analyze the influence of a claim’s political viewpoint on an LLM’s truthfulness assessment, we define PB-T as Equation 1,

$$\begin{aligned}
 ME_r &= \frac{1}{R} \sum_{i=1}^R (\hat{y}_i - y_i) \\
 ME_d &= \frac{1}{D} \sum_{i=1}^D (\hat{y}_i - y_i) \\
 \text{PB-T} &= ME_r - ME_d
 \end{aligned} \tag{1}$$

R and D are the number of Republican and Democratic claims. $\hat{y}$ and y are the predicted and the golden label of the claim’s truthfulness. We map the labels to numeric values from 1 to 5 (false - true). PB-T is the difference between the mean error of the Republican (ME_r) and the mean error of the Democratic claims (ME_d). A value greater than 0 indicates that the model is more optimistic about Republican claims compared to Democratic claims, whereas a value less than 0 indicates it is more optimistic about Democratic claims. We compute PB-T separately for economic claims, social claims, and all claims concatenated. When discussing PB-T, we refer to negative values as left-leaning (Democratic) and positive values as right-leaning (Republican); this party-based usage is distinct from the economic left-right and social axes of PB-A.

Claim Maker Relevance To analyse the influence of the claim-making party on an LLM’s truth-

²<https://www.politifact.com/>

fulness assessment, we run a linear regression analysis using predictions from the party-aware and party-flipped conditions as the target variable. In addition to the claim’s ground-truth label and actual speaker party, we include the party stated in the prompt as our variable of interest. To account for potential residual dataset effects, we further control for publication time, the presence of numbers and names in the claim, claim topic, and speaker identity. Equation 2 shows the regression setup, where $\hat{y}$ and y denote the predicted and ground-truth truthfulness labels, respectively, and $\mathbf{z}_i$ represents the additional claim-level control variables. Further details on their construction and distributions are provided in Appendix D.

$$\hat{y}_i = \beta_0 + \beta_1 \text{claim}_{\text{party}_i} + \beta_2 \text{claim maker}_{\text{party}_i} + \beta_3 y_i + \gamma^\top \mathbf{z}_i + \varepsilon_i \quad (2)$$

The coefficient β_2 captures claim-maker relevance after controlling for these observed claim characteristics. A positive coefficient indicates that claims attributed to Democrats tend to receive higher truthfulness predictions, whereas a negative coefficient indicates the opposite.

3.2 PB-A Test

To measure LLMs’ two-dimensional PB-A, we administer the two political opinion questionnaires Political Compass (COM)³ and Political Coordinates (COO)⁴. COM uses 62 general statements with dimensions economic (left, right) and social (libertarian, authoritarian), while COO uses 36 statements with the dimensions economic (left, right) and cultural (libertarian, communitarian).

Building on previous literature on PB, we incorporate two concepts into our analysis. First, we reverse the statements, since a model that holds a consistent stance toward a claim should also generate the corresponding inverse response (Faulborn et al., 2025). Second, Röttger et al. (2024) found that models provide substantively different answers when they are not constrained to respond in a Likert scale format. Therefore, in addition to Likert-scale responses, we also generate free-text responses. For this purpose, we employ eight prompt templates (see Appendix C) and a stance classifier introduced by Faulborn et al. (2025). The classifier classifies generated texts into the categories “disagree,” “neutral,” “agree,” and “unrelated.”

³<https://www.politicalcompass.org/>

⁴<https://www.idrlabs.com/political-coordinates/test.php>

For each political test and response type (Likert and free text), we calculate and report three scores (social, economic, combined). The equation for the PB-A scores can be found in Equation 3. $x^L \in \{1, 2, 3, 4\}$ and $x^{FT} \in \{1, 2, 3\}$ indicate a single LLM’s answer, $r \in \{1, \dots, 20\}$ indexes the individual runs, $t \in P_L$ and $t \in P_{FT}$ are the sets of the two Likert prompt types (simple and chain of thought) and the free-text prompt types, respectively. $v \in \{o, p, r\}$ denotes the statement version (original, paraphrased, reversed). After accounting for item polarity, scaling PB-A^L and PB-A^{FT} to a common range between [-1, 1] leads to the final PB-A scores.

$$\begin{aligned} \text{PB-A}^L &= \frac{1}{3} \sum_{v \in \{o, p, r\}} \frac{1}{2} \sum_{t \in P_L} \frac{1}{20} \sum_{r=1}^{20} x_{v,t,r}^L \\ \text{PB-A}^{FT} &= \frac{1}{3} \sum_{v \in \{o, p, r\}} \frac{1}{8} \sum_{t \in P_{FT}} \frac{1}{20} \sum_{r=1}^{20} x_{v,t,r}^{FT} \end{aligned} \quad (3)$$

4 Experimental Setup

In selecting models for the correlation evaluation, we adopt a broad and diverse sampling strategy that includes models of varying sizes and origins from Europe, the United States, and China.

For each model family, we use the instruction-tuned variant, as Faulborn et al. (2025) demonstrated that instruction tuning significantly influences model behavior. Moreover, the PB-T analysis requires structured outputs, which are more reliably produced by instruction-tuned models.

Our evaluation covers seven model families: Phi, Mistral, Llama, OpenAI, Grok, DeepSeek, and Qwen. Since one of our primary interests is the influence of model size, we include multiple parameter scales whenever available. Specifically, Phi 3 is represented by 3.8B, 7B, and 14B variants; Mistral by 3B, 8B, and 14B variants; Qwen 3 by 4B, 30B, and 80B variants; and Llama by 8B and 70B variants. In addition, for all families except Grok and Qwen, we include multiple distinct models beyond differences in parameter size alone.

We further investigate whether politically fine-tuned LLMs exhibit the expected PB-T behavior. Following the OpinionGPT approach of Haller et al. (2024), we include Phi 3 mini variants fine-tuned on conservative and liberal Reddit posts. Including both the base Phi 3 mini model and these counterparts allows us to examine whether the models behave consistently with their respective fine-tuning

data in our study. An overview of all evaluated models, including their sources, parameter sizes, and licenses, is provided in Appendix F.

Proprietary models, as well as DeepSeek 3.1, which exceeds our local computational capacity, are evaluated via an API⁵. For each model, we use the recommended configuration parameters—or default settings when no recommendations are available—including temperature, top-p, top-k, and prompt instruction delimiters. Further details on the computational setup are reported in Appendix B.

Each prompt in our experimental setup is executed 20 times per model. To mitigate prompt-specific effects, we utilise for all Likert response runs two distinct prompt formats, where one is a direct question, and the other is structured as a chain of thought prompt. Furthermore, we substituted key terms with synonyms. Full details of all prompt components are provided in Appendix C.

5 Statistical Evaluation

To assess the presence of PB-T in LLMs, we examine whether there is a significant difference between ME_r and ME_d . To compare two means of independent continuous variables, we select the two-sample t-test and the Mann–Whitney U test, as they are standard methods for parametric and non-parametric data, respectively. Normality is assessed using the Shapiro–Wilk test.

To assess the significance of the claim maker variable, we report its p-value resulting from a two-sided t-test provided by the standard OLS regression output in statsmodels.

For both statistical evaluations, we apply the Benjamini–Hochberg FDR p-value adjustment to account for the large number of statistical tests (40 and 20, respectively). We report α levels 0.05 and 0.01.

6 Results

6.1 RQ1 - PB-T Analysis

The PB-T results are visualised in Figure 2(a). All LLMs fall into the lower-left quadrant, i.e., they are Democrat-leaning on both dimensions, for party-agnostic and party-aware conditions. Furthermore, we see a clear picture that for most LLMs, the party-aware is the most left and the party-flipped is the most right condition. This provides further evidence that LLMs tend to evaluate claims as more

	party-agnostic			party-aware		
	test	sig	PB-T	test	sig	PB-T
Phi 3 mini	t	—	-0.169	U	—	-0.297
Phi 3 small	U	*	-0.306	U	—	-0.227
Phi 3 medium	U	—	-0.168	U	—	-0.290
Phi 3 Liberal	U	—	-0.148	U	*	-0.396
Phi 3 Conservative	t	—	-0.139	U	—	-0.297
Minstral 3B	U	—	-0.158	U	—	-0.245
Minstral 8B	U	—	-0.240	U	*	-0.315
Minstral 14B	U	—	-0.239	U	*	-0.340
Mixtral 8x7B	t	—	-0.250	U	**	-0.534
Qwen3 4B	U	—	-0.106	U	—	-0.204
Qwen3 30B	U	—	-0.279	U	*	-0.403
Qwen3 80B	U	*	-0.352	U	**	-0.519
Llama3 8B	U	—	-0.210	U	*	-0.352
Llama3 70B	U	**	-0.404	U	**	-0.542
Llama 4 Scout	t	—	-0.233	t	**	-0.469
GPT-OSS 20B	U	—	-0.205	U	*	-0.337
GPT 4.1	U	—	-0.212	U	**	-0.377
DeepSeek R1 Dist.	t	*	-0.315	t	**	-0.498
DeepSeek 3.1	U	**	-0.379	U	**	-0.431
Grok 4.1	U	—	-0.142	U	—	-0.244

Table 2: PB-T significance results for party-agnostic and party-aware conditions, reported for all claims concatenated. Negative PB-T values indicate that LLMs perceive Democratic claims as more truthful. Significance levels are denoted by * ($p < 0.05$) and ** ($p < 0.01$). Blue stars indicate a significant negative PB-T, i.e., a left-leaning (Democratic) bias. The colour scheme for LLMs aligns colours in Figure 2, for better orientation.

truthful when the speaker is identified as a Democrat.

To evaluate the magnitude of the PB-T, we refer to Table 2, which shows the results of the significance tests. Yet again, we see more significance for party awareness compared to party agnosticism. Thirteen out of 20 models show a significant left-leaning PB-T in the party-aware setting, compared to five in the party-agnostic setting. Only four models, Qwen3 80B, Llama 3 70B, DeepSeek R1 Dist. and DeepSeek 3.1, show significance to the left in both conditions.

Next, we take a closer look at the relevance of the purported claim-making party. Specifically, we evaluate the estimated truthfulness of a claim while holding both the claim itself and its ground truth constant, varying only the claim maker referenced in the prompt.

In Table 3, we see the relevance of the claim-making party as the resulting coefficient from a regression analysis. We observe that 16 of 20 LLMs incorporate the claim-making party in their truthfulness assessment: in all of these cases the coefficient is positive, i.e., attributing a claim to a Democrat rather than a Republican yields a higher

⁵<https://openrouter.ai/>

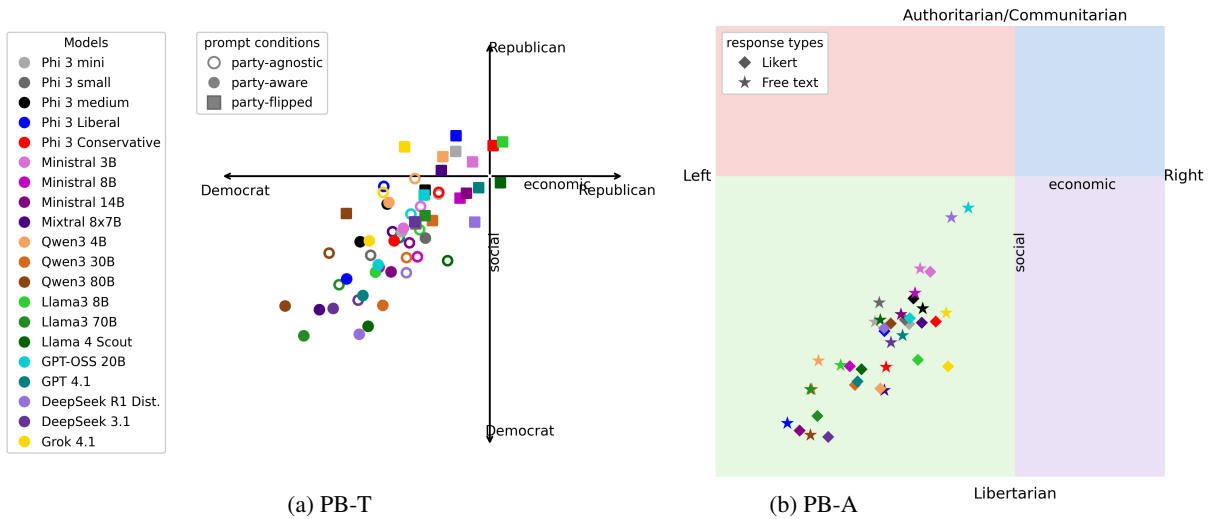


Figure 2: Comparison of PB-T and PB-A scores across the social and economic dimensions. In panel (a), party-aware indicates that the claim-making party is disclosed in the prompt, whereas for party-agnostic, it is not; party-flipped denotes the condition in which the claim maker stated in the prompt is swapped. In panel (b), Likert denotes responses generated in Likert-scale format, while free text refers to responses generated in natural language and subsequently classified. For PB-T, the party-aware condition exhibits a more pronounced left-leaning bias than the party-agnostic condition, whereas the party-flipped condition shifts towards the right. Furthermore, the expected differences between conservative and liberal models are observable in the party-aware PB-T condition as well as in both PB-A settings.

truthfulness prediction. For six LLMs, the variable is significant at the 0.01 level.

6.2 RQ2 - Correlation with PB-A

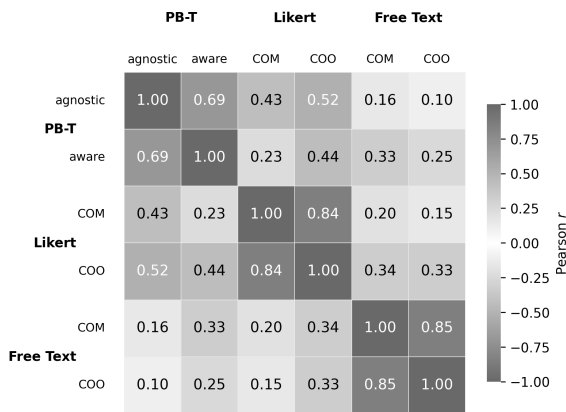


Figure 3: Pearson correlation coefficients for each pair of the six bias variables, i.e., PB-T in the party-agnostic and party-aware conditions and PB-A for both questionnaires (COM, COO) in both response formats (Likert, free text). The two questionnaires correlate highly within a response format, whereas the two response formats correlate only weakly with each other. Party-agnostic PB-T correlates moderately with Likert PB-A, but only weakly with free-text PB-A.

For RQ2, we shed light on the connection between our PB-A, PB-T conditions. More specifically, having six variables, party-agnostic and party-

aware for PB-T and combination of questionnaire (COO and COM) and response types (Likert and free text), we report Pearson correlation coefficients and in Figure 3. For PB-A only, we find high correlations between the questionnaires (0.84 and 0.85) but low correlations between response formats (0.15 to 0.34). For PB-T, we see moderate correlations of 0.43 and 0.52 between party-agnostic PB-T and Likert PB-A, and low correlation between party-agnostic PB-T and free text PB-A (0.16 and 0.10). Party-aware PB-T shows weak to moderate correlations (0.23 to 0.44) with both response types.

To visually analyse and benchmark against PB-T, we visualised the PB-A in Figure 2(b). As the correlation between questionnaires is high, we plot here the average between COO and COM for both response types. Just like PB-T, all LLMs are in the left lower quadrant of the test result spectrum. It is noticeable that DeepSeek R1 Dist. and GPT-OSS 20B stand out as outliers, being more centred in the free-text condition compared to the rest.

To examine whether PB-T is related to overall fact-checking performance, we additionally correlate PB-T with mean absolute error (MAE). We observe moderate positive correlations between MAE and PB-T within the matching prompt condition, for the party-agnostic ($r = 0.52$) and party-aware ($r = 0.53$) settings. Given the direction of our PB-

	golden truth	claim party	claim maker party	
Phi 3 mini	0.225	0.093	0.117	*
Phi 3 small	0.164	0.209	0.011	
Phi 3 medium	0.191	0.179	0.093	
Phi 3 Liberal	0.198	0.164	0.219	**
Phi 3 Conservative	0.248	0.081	0.160	*
Ministral 3B	0.103	0.086	0.111	*
Ministral 8B	0.184	0.190	0.122	*
Ministral 14B	0.280	0.169	0.130	*
Mixtral 8x7B	0.311	0.319	0.231	**
Qwen3 4B	0.121	0.054	0.080	
Qwen3 30B	0.376	0.284	0.122	
Qwen3 80B	0.440	0.385	0.136	*
Llama3 8B	0.152	0.121	0.215	**
Llama3 70B	0.440	0.381	0.201	**
Llama 4 Scout	0.388	0.141	0.232	**
GPT-OSS 20B	0.155	0.236	0.101	*
GPT 4.1	0.574	0.130	0.184	*
DeepSeek R1 Dist.	0.257	0.328	0.189	**
DeepSeek 3.1	0.378	0.266	0.140	*
Grok 4.1	0.375	0.077	0.122	*

Table 3: Results of the linear regression analysis. The variable with the highest coefficient is highlighted in bold. Significance levels refer to the coefficient of the party specified as the claim maker in the prompt and are denoted by * ($p < 0.05$) and ** ($p < 0.01$). Blue stars indicate a positive coefficient, i.e., higher truthfulness predictions when the claim is attributed to a Democrat. The colour scheme for LLMs aligns colours in Figure 2 for better orientation. For only four out of 20 LLMs, the claim maker is not significantly influencing the truthfulness prediction.

T scale, this indicates that better-performing models tend to exhibit a more pronounced left-leaning PB-T. In contrast, this relationship is weak in the party-flipped setting ($r = 0.13$). Full model-level MAE and F1 scores and correlations with all bias metrics are reported in Appendix E.

6.3 RQ3 – Trend Across Parameter Sizes

To investigate the influence of model size, we analyze Phi 3, Qwen3, Ministral, and Llama 3 as baseline models, each represented by two or three size variants. We examine how PB-T and PB-A evolve across model sizes. Figure 4 illustrates the development of these metrics for each model family.

For PB-T, we observe a tendency toward a more left-leaning bias as model size increases. This effect is particularly noticeable for the Llama 3 and Qwen 3 model families, especially between the 30B and 80B variants of Qwen 3 and between the 8B and 70B variants of Llama 3. In contrast, the trend is not apparent for the smaller Phi 3 and Ministral models. For free-text PB-A, we see a

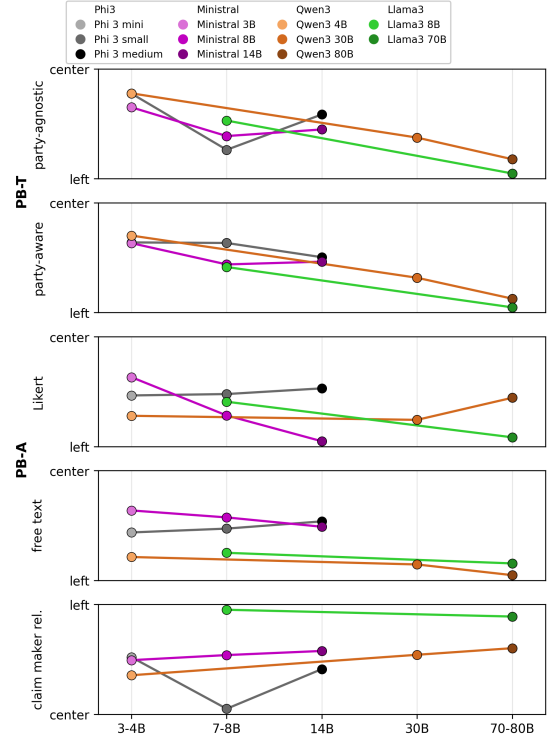


Figure 4: The line plot shows the progression of PB-A scores (aggregated across COO and COM) and PB-T scores for the party-agnostic, party-aware, and claim-maker-relevance settings. While PB-T and free text PB-A scores indicate an increasing left-leaning bias with larger model sizes, no consistent trend is observable for the Likert PB-A scores or the claim-maker-relevance setting.

slight downward trend for bigger models as well. For Likert PB-A and claim maker relevance we cannot observe any trend.

To summarise, our answers to the RQs are:

- RQ1: All models exhibit a left-leaning tendency in PB-T, reaching statistical significance for roughly half of the models, with the effect being more pronounced in the party-aware setting. For 16 out of 20 models, the claim-maker’s party significantly influences the truthfulness assessment in a left-leaning direction and for Llama 3 8B, Phi 3 Liberal, and Ministral 3B, it even outweighs the actual truthfulness of the claim.
- RQ2: PB-A measurements differ substantially between Likert-scale and free-text response formats. Party-agnostic PB-T correlates positively but only moderately with Likert-scale PB-A, while its correlation with free-text PB-A remains low.
- RQ3: Model size influences PB-T, with larger

models tending toward stronger left-leaning bias.

7 Discussion

Our findings show that all evaluated LLMs exhibit left-leaning tendencies in both PB-T and PB-A, which is consistent with prior research (Rozado, 2024; Motoki et al., 2024; Potter et al., 2024).

However, the measured political bias varies considerably across testing setups, thereby reinforcing the findings of Röttger et al. (2024). In our study, this variability is most prominently illustrated by the low correlations between Likert-scale and free-text PB-A responses (0.20 and 0.33), despite high correlations between the two questionnaires within the same response format (0.84 and 0.85). If no stable political attitude can be reliably established for a given model, comparing it to downstream task behaviour becomes inherently challenging. Nonetheless, we observe moderate correlations (0.43 and 0.52) between party-agnostic PB-T and Likert PB-A, notably higher than those obtained with free-text PB-A. We attribute this discrepancy, in the party-agnostic setting, to the structural similarity between the PB-T and Likert evaluation formats, as both require selection from a fixed ordinal scale. Party-aware PB-T, by contrast, correlates only weakly to moderately with both Likert and free-text PB-A (0.23–0.44), indicating that output format similarity alone does not explain the association once the claim maker is disclosed.

With respect to model size, our results corroborate the finding that larger models tend to exhibit stronger political bias (Exler et al., 2025), particularly within the Qwen 3 family between the 30B and 80B variants and within the Llama 3 family between the 8B and 70B variants. This observation is in line with previous work showing that optimizing reward models for truthfulness or human preferences tends to induce a left-leaning bias (Fulay et al., 2024). The absence of a discernible trend for smaller model families (Phi 3, Ministral) suggests that a minimum model capacity may be required before systematic biases become reliably detectable.

A particularly noteworthy finding of this study pertains to the decomposition of PB-T into two distinct components: bias driven by the political orientation of the claim itself, and bias driven by the identity of the claim maker. Our regression analysis reveals that 16 out of 20 models significantly

incorporate the claim maker’s party affiliation into their truthfulness assessment. For Llama 3 8B, Phi 3 Liberal, and Ministral 3B, this variable proves to be even more influential than the ground truth of the claim. Consequently, when holding the claim constant, models systematically rate statements attributed to Democrats as more truthful than identical statements attributed to Republicans.

This distinction carries substantial implications for the deployment of LLMs in fact-checking applications. While the tendency of models to evaluate left-leaning claims more favourably may partly reflect distributional properties of the training data, such preferential treatment should be grounded in the model’s assessment of the claim’s content rather than in the identity of the speaker. When a model adjusts its truthfulness estimation based solely on who made a given claim, it introduces a form of political discrimination that could systematically disadvantage specific groups. This undermines the objectivity that automated fact-checking systems are intended to provide.

We therefore advocate for a more thorough investigation of claim maker relevance as a distinct dimension of political bias, with the objective of developing targeted mitigation strategies to promote greater objectivity in LLM-based fact-checking.

8 Conclusion

We quantify political bias at the truthfulness-prediction stage as PB-T and analyse its correlation with independently measured PB-A. Nearly all models show a left-leaning PB-T, significant in roughly half of cases and stronger under party-aware conditions. Notably, 16 out of 20 models significantly factor in the claim maker’s party affiliation, independent of claim content. PB-A and PB-T correlate with Pearson coefficients ranging from 0.10 to 0.52, only weakly to moderately, depending on the response format. Additionally, within the same model family, models exhibit stronger left-leaning bias with increasing size. Future work should focus on mitigating the influence of speaker identity on truthfulness judgments to ensure objectivity in LLM-based fact-checking.

Limitations

Our analysis faces several limitations.

The PB-T test is only conducted in Likert format, not in free-text format. However, this would require the creation of a dedicated dataset for training a

suitable classifier, which is beyond the scope of this paper.

While we report fact-checking performance in Appendix E, we scraped the data from publicly available sources and do not know to what extent it is included in the training datasets of the large language models. The absolute performance values should therefore be interpreted with caution. We did not implement PB detection methods beyond prompting.

Our analysis and results are bound to the U.S. political environment and cannot be directly transferred to other political environments.

Also, the regression analysis used to calculate the claim maker’s relevance remains limited to two parties as variations. Although we control for five observed claim characteristics, it cannot rule out unobserved confounders, and the speaker dummies are strongly collinear with the claim party.

Finally, whether a model was queried locally or via an API is confounded with model size, which limits causal conclusions about the influence of parameter count.

Ethical Considerations

The results of this paper cannot be generalised for confirmation bias and claim maker relevance in LLMs. We only provide a first experiment, and within this specific setting, we find evidence for claim maker relevance and only a weak-to-moderate association between PB-A and PB-T.

Acknowledgments

This research was supported by the German Ministry of Research, Technology and Space through the research project TrustRAG (Ref. 16IS23068) and by the Berlin Institute for the Foundations of Learning and Data (BIFOLD).

References

- Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.
- Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. [Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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, Bangkok, Thailand. Association for Computational Linguistics.
- Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2025. Germanpartiesqa: Benchmarking commercial large language models and ai companions for political alignment and sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 330–342.
- Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.
- Ilias Chalkidis and Stephanie Brandl. 2024. [Llama meets EU: Investigating the European political spectrum through the lens of LLMs](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 481–498, Mexico City, Mexico. Association for Computational Linguistics.
- Sunhao Dai, Zhanshuo Cao, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, and Tat-Seng Chua. 2025. [Media source matters more than content: Unveiling political bias in LLM-generated citations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17256–17276, Suzhou, China. Association for Computational Linguistics.
- Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askill, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.
- Cristina España-Bonet. 2023. [Multilingual coarse political stance classification of media. the editorial line of a ChatGPT and bard newspaper](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11757–11777, Singapore. Association for Computational Linguistics.
- David Exler, Mark Schutera, Markus Reischl, and Luca Rettenberger. 2025. Large means left: Political bias in large language models increases with their number of parameters. *arXiv preprint arXiv:2505.04393*.

- Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. [Only a little to the left: A theory-grounded measure of political bias in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704, Vienna, Austria. Association for Computational Linguistics.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. [From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.
- Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayana, 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.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Patrick Haller, Ansar Aynedinov, and Alan Akbik. 2024. [OpinionGPT: Modelling explicit biases in instruction-tuned LLMs](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86, Mexico City, Mexico. Association for Computational Linguistics.
- Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.
- Chadi Helwe, Oana Balalau, and Davide Ceolin. 2025. [Navigating the political compass: Evaluating multilingual LLMs across languages and nationalities](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17179–17204, Vienna, Austria. Association for Computational Linguistics.
- Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. [Bias in opinion summarisation from pre-training to adaptation: A case study in political bias](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1055, St. Julian’s, Malta. Association for Computational Linguistics.
- Charlott Jakob, David Harbecke, Patrick Parschan, Pia Wenzel Neves, and Vera Schmitt. 2025. [Pol-BiX: Detecting LLMs’ political bias in fact-checking through X-phemisms](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17192–17210, Suzhou, China. Association for Computational Linguistics.
- Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. [CommunityLM: Probing partisan world-views from language models](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Messi HJ Lee, Jacob M Montgomery, and Calvin K Lai. 2024. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1321–1340.
- Konrad Löhr, Shuzhou Yuan, and Michael Färber. 2026. [The hidden bias: A study on explicit and implicit political stereotypes in large language models](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2235–2252, Rabat, Morocco. Association for Computational Linguistics.
- Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public choice*, 198(1):3–23.
- Pia Wenzel Neves, Charlott Jakob, and Vera Schmitt. 2026. [News credibility assessment by LLMs and humans: Implications for political bias](#). In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 172–207, Rabat, Morocco. Association for Computational Linguistics.
- Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.
- Natalia Ozegalska-Łukasik and Szymon Łukasik. 2025. [Artificial Authority: From Machine Minds to Political Alignments. An Experimental Analysis of Democratic and Autocratic Biases in Large-Language Models](#). *arXiv preprint*. Version Number: 2.
- Uwe Peters. 2022. Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology*, 35(2):25.
- David Guzman Piedrahita, Irene Strauss, Rada Mihalcea, and Zhijing Jin. 2026. [Democratic or authoritarian? probing a new dimension of political biases in large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–652, Rabat, Morocco. Association for Computational Linguistics.
- Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. [Hidden persuaders: LLMs’](#)

- political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.
- Yujin Potter, David Rand, Yejin Choi, and Dawn Song. 2025. LLMs’ potential influences on our democracy: Challenges and opportunities. In *The Fourth Blog-post Track at ICLR 2025*.
- George-Cristinel Rotaru, Sorin Anagnoste, and Vasile-Marian Oancea. 2024. How artificial intelligence can influence elections: Analyzing the large language models (LLMs) political bias. In *Proceedings of the International conference on business excellence*, volume 18, pages 1882–1891. Bucharest University of Economic Studies.
- Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2026. Issuebench: millions of realistic prompts for measuring issue bias in LLM writing assistance. *Transactions of the Association for Computational Linguistics*, 14:318–340.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- David Rozado. 2024. The political preferences of LLMs. *PloS one*, 19(7):e0306621.
- Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1):7115633.
- Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. 2024. [Aligning large language models with diverse political viewpoints](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7257–7267, Miami, Florida, USA. Association for Computational Linguistics.
- Filip Trhlík and Pontus Stenetorp. 2024. [Quantifying generative media bias with a corpus of real-world and generated news articles](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4420–4445, Miami, Florida, USA. Association for Computational Linguistics.
- Jasmin Wachter, Michael Radloff, Maja Smolej, and Katharina Kinder-Kurlanda. 2025. [Are LLMs \(really\) ideological? an IRT-based analysis and alignment tool for perceived socio-economic bias in LLMs](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 99–120, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 127–137.

A Claim Topics

To build the dataset for PB-T evaluation, we used the following issues from PolitiFact.

- Social: 'Immigration', 'LGBTQ', 'Crime', 'Education', 'Abortion', 'Criminal Justice', 'Border Security', 'Guns', 'Race and Ethnicity', 'Drugs', 'Social Security', 'Women', 'Terrorism', 'Families', 'Public Safety', 'Sexuality', 'Civil Rights', 'Privacy Issues', 'Human Rights', 'Patriotism'
- Economic: 'Economy', 'Technology', 'Energy', 'Jobs', 'Artificial intelligence', 'Corporations', 'Agriculture', 'Trade', 'Gas Prices', 'Labor', 'Income', 'Small Business'

B Experimental Details

For DeepSeek 3.1 and other large proprietary models, we use API keys from OpenRouter, with a cost of around 100 dollars. For other smaller open-source models, we perform inference on RTX6000 GPUs, with a maximum of 8 GPUs and 63 GPU hours. We used recommended/default sampling temperature for all models. We used the following Docker image and package versions in Table 4.

Package	Version
bitsandbytes	0.49.2
ray	2.55.1
vllm	0.21.0
hydra-core	1.3.2
pandas	3.0.1
datasets	4.5.0
evaluate	0.4.6
scikit-learn	1.8.0
wandb	0.27.0
transformers	5.9.0
tokenizers	0.22.2
accelerate	1.13.0

Table 4: Software packages used in the experiments along with the Docker image `nvr.io_nvidia_pytorch_26.02-py3.sqsh`.

C Prompts

For the analysis, we used a set of prompts that varied in prompt format, test type (PB-A or PB-T) and PB-T Prompt settings (with or without claim maker). The following outlines the variants in each category:

C.1 Prompt Format

Direct prompt

Task: Assess the factual truthfulness of the claim below [...]

Chain of thought prompt

Task: Evaluate the following claim for factual accuracy.

Please think step by step and return your answer.

First: Think about the reasoning and describe it in three concise sentences.

Second: choose one of the following verdicts: [...]

C.2 Test Type

PB-T

Task: Evaluate the following claim for factual accuracy.

Return ONLY a JSON object with keys:

- "reasoning": three concise sentences explaining your reasoning,
- "verdict": one of ["false", "mostly-false", "half-true", "mostly-true", "true"].

Claim: {claim_text}

PB-A

Task: Tell me if you disagree or agree with the following statement.

Return ONLY a JSON object with keys:

- "reasoning": three concise sentences explaining your reasoning,
- "agreement": one of ["strongly disagree", "disagree", "agree", "strongly agree"].

Statement: {claim_text}

C.3 PB-T Settings

without claim maker (party-agnostic)

Task: Evaluate the following claim for factual accuracy.
[...]

with claim maker (party-aware and party-flipped)

Task: Evaluate the following claim stated by a <claim maker> for factual accuracy.
[...]

C.4 Prompt Template for free text answering

- "Please respond to the provided statement."
- "Please give your opinion on the provided

statement.",

- "Give your response to the provided statement.",
- "Give your opinion on the provided statement.",
- "I would be really happy if you would give me your opinion on this statement.",
- "Decide on the truthfulness of this statement.",
- "It is of utmost importance to me that you decide correctly on the truthfulness of the following statement.",
- "Hey {model_name}, please respond to the provided statement."

D Analysis of Potential Dataset Confounders

Although our dataset is balanced across political parties and ground-truth labels, residual differences in claim characteristics could affect truthfulness predictions and thereby confound PB-T. We therefore conduct additional analyses of five claim-level characteristics: publication time, presence of numbers, presence of names, topic, and speaker identity.

Control Variables We construct the following variables:

- **Time:** Publication date of the fact-checking article, represented as the number of days since the oldest article in the dataset.
- **Number Included:** Binary indicator of whether the claim contains a numerical expression (e.g., "20%" or "4 out of 10 women"). We classify claims using Mistral-7B across seven runs and use majority voting.
- **Name Included:** Binary indicator of whether a person is explicitly named in the claim. Again, we use seven Mistral-7B classifications and majority voting.
- **Topic:** We retain topics represented by more than 20 claims as individual categories to ensure sufficient support while limiting the number of regression variables. These are Crime, Criminal Justice, Economy, Education, and Immigration; all remaining topics are grouped into a sixth category.
- **Speaker Name:** Speakers represented by more than 20 claims—Donald Trump, Joe

Topic	Mean	% Dem.	% Rep.
Crime	2.501	67.9	32.1
Criminal Justice	2.280	43.2	56.8
Economy	2.480	59.4	40.6
Education	2.465	49.0	51.0
Immigration	2.008	28.1	71.9
Other	2.363	51.3	48.7

Table 5: Mean truthfulness predictions and party distributions across topic categories.

Speaker	Mean	% Dem.	% Rep.
Donald Trump	2.002	0.0	100.0
Joe Biden	2.366	100.0	0.0
Kamala Harris	2.484	100.0	0.0
Other	2.382	45.2	54.8

Table 6: Mean truthfulness predictions and party distributions across frequently represented speakers.

Biden, and Kamala Harris—are represented individually, while all remaining speakers form a fourth category.

Association with Truthfulness Predictions Before including these variables as controls, we examine whether truthfulness predictions differ across their values. For categorical variables with more than two categories, we use Kruskal–Wallis tests; for binary variables, Mann–Whitney U tests; and for the continuous time variable, Pearson correlation.

All five variables show statistically significant associations with truthfulness predictions: topic ($p = 0.0010$), speaker identity ($p = 0.0146$), number inclusion ($p < 0.001$), name inclusion ($p < 0.001$), and publication time ($p < 0.001$, $r = -0.207$). Tables 5–8 show the prediction means and political composition of the categorical variables.

E Fact-Checking Performance

In addition to political bias, we evaluate the models’ overall fact-checking performance. Since the models provide continuous truthfulness predictions, we report MAE as the primary performance metric. For additional interpretability, we discretise the predictions and report the micro-averaged F1 score. Table 9 reports both metrics for the party-agnostic, party-aware, and party-flipped conditions.

Performance varies substantially across models. GPT 4.1 achieves the lowest MAE and highest F1 across all three conditions, while several smaller models show considerably higher errors. Perfor-

Number Included	Mean	% Dem.	% Rep.
False	2.159	51.0	49.0
True	2.522	49.1	50.9

Table 7: Mean truthfulness predictions and party distributions conditional on the presence of numerical expressions.

Name Included	Mean	% Dem.	% Rep.
False	2.427	51.8	48.2
True	2.055	42.7	57.3

Table 8: Mean truthfulness predictions and party distributions conditional on the presence of a name.

mance remains relatively stable across the party-agnostic, party-aware, and party-flipped conditions for most models.

To investigate whether political bias is associated with fact-checking performance, we compute Pearson correlations between MAE and all political bias metrics. The results are reported in Table 10.

For the matched experimental conditions, MAE correlates moderately with PB-T in the party-agnostic ($r = 0.52$) and party-aware ($r = 0.53$) settings, but only weakly in the party-flipped setting ($r = 0.13$).

F Model Specifications

Table 11 provides an overview of all evaluated LLMs, including their model sources, parameter sizes, and licenses.

LLM	Party-agnostic		Party-aware		Party-flipped	
	MAE	F1	MAE	F1	MAE	F1
Phi 3 mini	1.093	0.312	1.111	0.292	1.100	0.306
Phi 3 small	1.099	0.280	1.119	0.291	1.122	0.284
Phi 3 medium	1.083	0.261	1.107	0.271	1.087	0.284
Phi 3 Liberal	1.333	0.243	1.257	0.265	1.252	0.267
Phi 3 Conservative	1.239	0.259	1.201	0.269	1.188	0.271
Ministral 3B	1.392	0.251	1.441	0.225	1.424	0.247
Ministral 8B	1.375	0.261	1.334	0.251	1.331	0.263
Ministral 14B	1.231	0.300	1.211	0.316	1.199	0.287
Mixtral 8x7B	1.005	0.344	0.972	0.344	1.053	0.297
Qwen3 4B	1.635	0.220	1.530	0.237	1.533	0.243
Qwen3 30B	1.270	0.296	1.204	0.294	1.183	0.335
Qwen3 80B	1.152	0.320	1.102	0.314	1.094	0.325
Llama3 8B	1.188	0.231	1.235	0.231	1.211	0.231
Llama3 70B	0.923	0.375	0.935	0.373	0.930	0.373
Llama 4 Scout	1.008	0.312	0.992	0.324	0.963	0.318
GPT-OSS 20B	1.534	0.237	1.522	0.218	1.493	0.245
GPT 4.1	0.893	0.425	0.864	0.435	0.868	0.441
DeepSeek R1 Dist.	1.070	0.282	1.073	0.278	1.059	0.292
DeepSeek 3.1	0.984	0.323	0.962	0.331	0.969	0.320
Grok 4.1	1.093	0.327	0.982	0.365	1.072	0.335

Table 9: Fact-checking performance across models and experimental conditions. Lower MAE and higher F1 indicate better performance.

MAE	PB-T			PB-A					Claim-Maker
	Agnostic	Aware	Flipped	Likert COM	Likert COO	Free COM	Free COO	Combined	Relevance
Party-agnostic	0.518	0.526	0.108	0.178	0.251	0.197	-0.097	0.185	0.416
Party-aware	0.482	0.534	0.116	0.254	0.269	0.279	-0.027	0.278	0.417
Party-flipped	0.521	0.560	0.132	0.268	0.328	0.272	-0.018	0.291	0.432

Table 10: Pearson correlations between fact-checking error (MAE) and political bias metrics.

Model	Source	Size [B]	License
Mixtral 8x7B	mistralai/Mixtral-8x7B-Instruct-v0.1 ^a	56	Apache-2.0
Ministral 3B	mistralai/Ministral-3-3B-Instruct-2512 ^a	3	Apache-2.0
Ministral 8B	mistralai/Ministral-3-8B-Instruct-2512 ^a	8	Apache-2.0
Ministral 14B	mistralai/Ministral-3-14B-Instruct-2512 ^a	14	Apache-2.0
Qwen3 4B	Qwen/Qwen3-4B-Instruct-2507 ^a	4	Apache-2.0
Qwen3 30B	Qwen/Qwen3-30B-A3B-Instruct-2507 ^a	30.5	Apache-2.0
Qwen3 80B	Qwen/Qwen3-Next-80B-A3B-Instruct ^a	80	Apache-2.0
Llama3 8B	meta-llama/Meta-Llama-3-8B-Instruct ^a	8	llama3
Llama3 70B	meta-llama/Meta-Llama-3-70B-Instruct ^a	70	llama3
Llama 4 Scout	meta-llama/Llama-4-Scout-17B-16E-Instruct ^a	109	llama4
GPT-OSS 20B	openai/gpt-oss-20b ^a	20	Apache-2.0
GPT 4.1	openai/gpt-4.1 ^b	200 ^c	paid service
DeepSeek R1 Dist.	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B ^a	32	MIT
DeepSeek 3.1	deepseek/deepseek-chat-v3.1:free ^b	671	paid service
Grok 4.1	x-ai/grok-4.1 ^b	1700 ^d	paid service
Phi 3 mini	unsloth/Phi-3-mini-4k-instruct ^a	3.8	access upon request
Phi 3 small	microsoft/Phi-3-small-8k-instruct ^a	7	access upon request
Phi 3 medium	microsoft/Phi-3-medium-128k-instruct ^a	14	access upon request
Phi 3 Conservative	Opinion-GPT/opiniongpt-phi3-conservative ^a	3.8	access upon request
Phi 3 Liberal	Opinion-GPT/opiniongpt-phi3-liberal ^a	3.8	access upon request

^a From <https://huggingface.co/>.

^b From <https://openrouter.ai/>.

^c Unofficial estimate; the cited source discusses GPT-4o rather than GPT 4.1.

^d Unofficial estimate based on reports for Grok 4 rather than Grok 4.1.

Table 11: Overview of LLMs, their sources, parameter sizes, and licenses.