Bridging Human and Model Perspectives: A Comparative Analysis of Political Bias Detection in News Media Using Large Language Models

Shreya Adrita Banik, Niaz Nafi Rahman, Tahsina Moiukh, Farig Sadeque
Department of Computer Science and Engineering, BRAC University
{shreya.adrita.banik, niaz.nafi.rahman, tahsina.moiukh}@g.bracu.ac.bd farig.sadeque@bracu.ac.bd

Abstract

Detecting political bias in news media is a complex task that requires interpreting subtle linguistic and contextual cues. Although recent advances in Natural Language Processing (NLP) have enabled automatic bias classification, the extent to which large language models (LLMs) align with human judgment still remains relatively underexplored and not yet well understood. This study aims to present a comparative framework for evaluating the detection of political bias across human annotations and multiple LLMs, including GPT, BERT, RoBERTa, and FLAN. We construct a manually annotated dataset of news articles and assess annotation consistency, bias polarity, and inter-model agreement to quantify divergence between human and model perceptions of bias. Experimental results show that among traditional transformer-based models, RoBERTa achieves the highest alignment with human labels, whereas generative models such as GPT demonstrate the strongest overall agreement with human annotations in a zero-shot setting. Among all transformer-based baselines, our fine-tuned RoBERTa model acquired the highest accuracy and the strongest alignment with human-annotated labels. Our findings highlight systematic differences in how humans and LLMs perceive political slant, underscoring the need for hybrid evaluation frameworks that combine human interpretability with model scalability in automated media bias detection.

1 Introduction

Political bias in news discourse shapes public opinion and influences democratic decision-making. While recent advances in transformer-based language models have improved automated bias detection, the extent to which these models *interpret* bias in ways that align with human reasoning remains unclear. Most prior work focuses on training classifiers or analyzing ideological framing, treating human annotation as a fixed ground truth. However, this overlooks a key question: *do models and humans rely on similar cues when identifying political bias?*

In this work, we introduce a comparative framework to evaluate human–model alignment in political bias annotation. We analyze how different model families like BERT, RoBERTa, XLM-R, FLAN-T5, and GPT-5 assign bias labels under zero-shot, guideline-conditioned, and few-shot prompting regimes. Using a human-annotated corpus of news text, we examine not only model accuracy, but also the interpretive strategies underlying model decisions. Our analysis shows clear differences, like relying too much on word sentiment cues and struggling to tell apart quoted speech from authorial stance.

The results indicate that while fine-tuned transformers perform well on explicit bias signals, generative models like GPT-5 show better zero-shot alignment with human annotation patterns. Yet across all model families, alignment weakens for implicit or context-dependent cases. This suggests that current systems detect *entity-level sentiment* rather than the *discourse-level framing* used in human judgment. This gap highlights the need for bias detection methods that include contextual inference, speaker attribution, and narrative structure.

Our contributions are threefold:

1. **Human-Annotated Bias Corpus.** We construct a human-annotated corpus of news articles labeled for political bias which involves a multi-step annotation process with adjudicated disagreement resolution. This approach offers an interpretable and consistently coded benchmark for examining human–model alignment in bias perception.

2. **Unified Evaluation Framework.** We present a unified framework that compares annotation behavior across multiple model families (BERT, RoBERTa, XLM-R, FLAN-

T5, and GPT-based models) under zero-shot, guideline-conditioned, and few-shot prompting configurations. The framework separates the effects of model architecture from those of inference strategy on bias classification.

3. **Qualitative Divergence Analysis.** We provide a comparative analysis that shows systematic divergences between human reasoning and model predictions, including: (i) reliance on lexical polarity rather than contextual framing, (ii) failure to differentiate between speaker attribution from author stance, and (iii) entity-centered heuristics that overshadow multi-actor narrative interpretation.

Taken together, our findings show that higher model accuracy does not imply human-aligned interpretation. Current models perform bias classification primarily through sentiment- and entity-level heuristics, whereas human annotators rely on discourse structure, contextual justification, and pragmatic inference. This shows a significant difference between pattern-matching behavior and interpretive reasoning in detecting political bias.

2 Related Work

2.1 Political Bias Detection in NLP

Political bias detection has been a central topic in computational social science and NLP, aiming to identify ideological slant and framing in textual media. Moverover, modern Language Models have political learnings that reinforce polarization ((Feng et al., 2023)). Early research employed linguistic and source-level features to classify bias in news reporting (Baly et al., 2020; Fan et al., 2019; Morstatter et al., 2014). However media bias is a multi-dimensional phenomenon manifested through framing, agenda-setting, and information selection, yet it is often operationalized in computational research as a unidimensional text-classification task ((Rodrigo-Ginés et al., 2024)). Recent transformer-based models have improved robustness by leveraging contextualized embeddings for political ideology prediction. However, these approaches rely heavily on human-labeled data, and the subjectivity inherent in political labeling limits the generalizability and interpretability of the resulting models.

2.2 Bias and Fairness in Large Language Models

Large Language Models (LLMs) have transformed NLP evaluation, revealing both their potential as detectors of social bias and their own representational biases. Jaremko et al. (2025) show that GPT-4 and LLaMA-3 achieve near-human performance in identifying implicitly abusive language. This indicates that LLMs capture nuanced linguistic features correlated with implicit bias. Similarly, Movva et al. (2024) evaluate annotation alignment between GPT-4 and human annotators across conversational safety datasets and their findings include that GPT-4 often matches or exceeds the median human–human correlation. Münker et al. (2024) demonstrate that zero-shot prompt-based annotation using models such as FLAN-T5 or mT0 can achieve results comparable to fine-tuned classifiers on multilingual political datasets. These findings collectively suggest that LLMs can serve as capable, if imperfect, substitutes for human annotators, particularly for complex or subjective linguistic phenomena.

2.3 Human–Machine Annotation Alignment

Recent studies have increasingly examined how closely large language model (LLM) annotations align with human judgment, particularly for subjective NLP tasks. Prior work highlights substantial variability among human annotators themselves, as well as between humans and models (Plank, 2022b; Davani et al., 2022). Chiang and Lee (2023) and Gilardi et al. (2023) find that GPT-based models can replicate or even outperform crowdworkers in text evaluation tasks, with GPT-3.5 producing higher-quality annotations than MTurk at significantly lower cost. Similarly, Movva et al. (2024) report that GPT-4 achieves higher correlation with average human ratings than the median human annotator. (Bojic et al., 2025) find GPT-4 yields greater inter-rater reliability than humans on political stance and sentiment annotation. Prompt-based models such as FLAN-T5 and mT0 have also demonstrated competitive zero-shot annotation performance in political discourse and topic classification (Jaremko et al., 2025; Münker et al., 2024).

Despite these advances, alignment remains imperfect. Santurkar et al. (2023) show that LLM outputs often reflect a homogenized ideological bias (e.g., left-leaning), while Lee et al. (2023) observe

that current models fail to fully capture the diversity of human opinions. In the fairness domain, multilingual models trained on stereotype benchmarks like CrowS-Pairs and BBQ show reduced bias and improved generalization compared to monolingual counterparts (Nie et al., 2024). These findings collectively motivate our study, which systematically compares human and model annotations in political bias detection, quantifying divergence across multiple model families, including GPT, BERT, RoBERTa, XLM-R, and FLAN.

3 Dataset

3.1 Dataset Overview

We use the **FIGNEWS 2024** dataset, released as part of the ArgMining shared task on news media narratives (FIGNEWS, 2024; Zaghouani et al., 2024). The dataset consists of news headlines and Facebook posts related to the Israel–Palestine conflict, published between *October 7, 2023* and *January 31, 2024*. It includes data in five languages—English, Arabic, French, Hebrew, and Hindi—with machine translations to English provided by the organizers. For consistency and comparability, all entries were analyzed in English.

After removing irrelevant metadata and incomplete entries, our final dataset contains approximately **15,000 samples** with 11 attributes, later reduced to 5 core columns (`source`, `type`, `text`, `translation`, and `bias`). Each instance was manually annotated for bias into one of seven labels: *Unclear*, *Biased Against Palestine*, *Biased Against Israel*, *Unbiased*, *Biased Against Both*, *Biased Against Others*, and *Not Applicable*. The data was divided into **70% training**, **10% validation**, and **20% test** splits.

3.2 Human Annotation Process

The annotation was performed manually by three researchers following a structured annotation guideline inspired by the FIGNEWS shared task objectives. Each annotator labeled entries independently, after which discrepancies were resolved through group discussion. Inter-annotator agreement (IAA) was calculated using Cohen's κ, yielding substantial agreement scores of **0.65–0.68**.

Annotators assessed whether each text was relevant to the conflict, objective or subjective, and if biased, toward which entity the bias was directed. Weekly consensus meetings ensured that ambiguous or borderline cases were discussed and resolved. The resulting manually labeled set serves as the *human benchmark* for the comparative analysis presented in later sections.

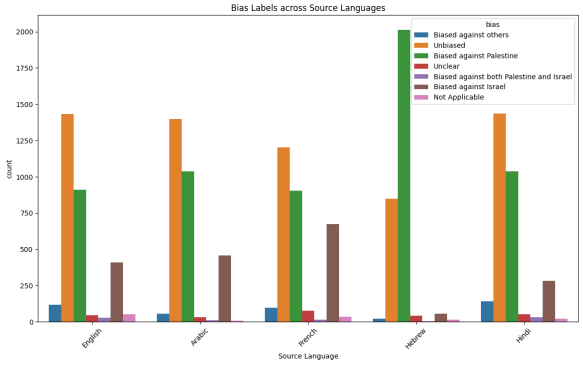

Figure 1: Distribution of political bias categories in the human-annotated FIGNEWS subset.

3.3 Annotation Guidelines

The annotation schema defines seven mutually exclusive bias categories:

- **Unbiased**: Objective reporting that represents all sides fairly and avoids loaded language.
- **Unclear**: Texts lacking sufficient context to determine direction of bias.
- **Biased Against Israel**: Negative framing or delegitimization of Israel or its institutions.
- **Biased Against Palestine**: Negative framing or stereotypes directed at Palestinians or Palestinian groups.
- **Biased Against Both**: Simultaneous negative portrayal of both sides.
- **Biased Against Others**: Bias directed toward external actors (e.g., USA, UN, Iran).
- **Not Applicable**: Texts unrelated to the Israel–Palestine conflict.

Full annotation guidelines with examples for each class are provided in the Appendix A.1.

3.4 Preprocessing

We conducted extensive data preprocessing to ensure data consistency. All text was lowercased and stripped of URLs, emojis, and non-ASCII characters. Unlike traditional NLP pipelines, we retained stop words and inflections to preserve contextual richness which helps transformer-based models capture subtle rhetorical cues. Texts were normalized and tokenized using the `Hugging Face` tokenizer corresponding to each model.

The preprocessed dataset was then inspected through exploratory data analysis (EDA), including

frequency plots and word clouds to find key thematic terms and demographic distributions in the dataset.

3.5 Data Augmentation

The dataset exhibited substantial class imbalance, with two dominant categories (*Unbiased*, *Biased Against Palestine*) overshadowing the rest. In order to mitigate this, we employed GPT-4o-mini via the OpenAI API to paraphrase minority-class examples. For each underrepresented label, 3–5 semantically equivalent variants were generated per instance ensuring that the original meaning and bias polarity were preserved. This increased minority-class coverage and improved model robustness during fine-tuning.

All augmentation outputs were manually verified to ensure semantic consistency and avoid label drift.

4 Methodology

4.1 Overview

We employ both recurrent and transformer-based neural architectures to detect and quantify political bias in news articles. Our approach first establishes a sequence-modeling baseline using a BiLSTM network and subsequently fine-tunes several transformer models like BERT, RoBERTa, and LLaMA-3.1 on the annotated dataset. Finally, we extend the framework by comparing human annotations with transformer-generated labels to assess cross-annotator consistency and model interpretability.

4.2 Model Architectures

4.2.1 Baseline: BiLSTM

To capture sequential dependencies and linguistic context, we implemented a Bidirectional Long Short-Term Memory (BiLSTM) network as a baseline. The model processes text in both forward and backward directions which helps it to preserve contextual information across variable-length sequences. Word embeddings were initialized using pre-trained *GloVe* and *FastText* representations. The network comprises two stacked BiLSTM layers (512 and 256 units) followed by a dense layer with a softmax activation for seven-way classification. Dropout layers and early stopping were applied to prevent overfitting, and optimization was performed using the *Adam* optimizer with a learning rate of 0.001.

4.2.2 Transformer Models

For context-aware language representation, we employed transformer-based architectures—BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)—as the core models for bias classification. Both models were fine-tuned for sequence classification using the Hugging Face `Transformers` library. Special tokens (`[CLS]`, `[SEP]`) were used to encode sentence boundaries, and outputs from the `[CLS]` token were passed through a classification head. We additionally experimented with *LLaMA-3.1* (8B variant), fine-tuned using parameter-efficient tuning (QLoRA) and 4-bit quantization to evaluate the scalability of large language models under resource-constrained settings.

4.3 Fine-Tuning Procedure

Each model was fine-tuned using the same preprocessed dataset described in Section 3. For transformers, we used a batch size of 64 and trained for three epochs with early stopping based on validation loss. Optimization was performed with *AdamW* and a weight decay of 0.01. Learning rates were set to 2×10^{-5} for transformers and 1×10^{-3} for the BiLSTM. All experiments were conducted on an NVIDIA RTX 4090 GPU. Cross-entropy loss was used for multi-class classification, and model checkpoints were saved based on the highest macro F1 score on the validation set.

4.4 Evaluation Setup

Model performance was evaluated on the held-out test set using accuracy, precision, recall, and macro F1-score. We also performed a paired t-test to assess whether the difference between BERT and RoBERTa performance was statistically significant ($p > 0.05$). Confusion matrices were used to analyze class-level misclassifications and identify systematic prediction errors.

4.5 Human vs. Machine Annotation Framework

To evaluate how closely transformer-based models align with human perception of political bias, we designed an annotation framework comparing human-labeled data with automatically generated annotations from both open-source and proprietary large language models (LLMs). This framework explores three inference regimes—**zero-shot**, **zero-shot with guideline conditioning**, and **few-shot**

with guideline examples—across multiple model architectures.

Annotation Models We employed five representative language models covering a range of architectures and parameter scales:

- **BERT-MNLI (base)** a bidirectional transformer fine-tuned on the MultiNLI entailment task, used as a lightweight baseline for entailment-based classification.
- **RoBERTa-MNLI (large)** a robust transformer trained with improved optimization and dynamic masking, serving as the primary zero-/few-shot classifier.
- **XLM-RoBERTa (large)** a multilingual transformer enabling cross-lingual evaluation of the FIGNEWS dataset.
- **FLAN-T5 (base)** a sequence-to-sequence model fine-tuned with instruction-following objectives.
- **ChatGPT 5.0 (API)** a multimodal large language model with improved reasoning and context understanding

Annotation Regimes Each model was prompted under three distinct configurations:

1. **Zero-shot:** Models classified texts into one of the seven bias categories without any external guidance.
2. **Zero-shot + Guidelines:** Models received concise natural language definitions for each bias label (e.g., "Biased against Palestine" or "Unbiased"), adapted from our human annotation manual.
3. **Few-shot + Guidelines:** Models were provided with 2–3 human-annotated examples per label, alongside guidelines serving as contextual anchors.

Inference Pipeline. For open-source models, inference was conducted via the Hugging Face `transformerszero-shot-classification` pipeline, using the `roberta-large-mnli` and `bert-base-mnli` checkpoints. Each text was evaluated against all seven candidate hypotheses derived from the guideline definitions. The model produced entailment scores for each label, and the highest-scoring label was selected as the predicted bias category. To mitigate spurious predictions, additional decision rules were applied: low-confidence predictions ($p < 0.22$) or minimal label margins ($\Delta p < 0.03$) were assigned as *Unclear*, and texts deemed irrelevant to the Israel–Palestine conflict (via a secondary NLI-based relevance classifier) were labeled as *Not Applicable*. A dual-threshold heuristic ($s_{BAI}, s_{BAP} \geq 0.30$, $|s_{BAI} - s_{BAP}| \leq 0.06$) was used to identify instances "Biased against both Palestine and Israel."

Guideline Integration. In guideline-conditioned setups, each hypothesis included the label's definition, and for few-shot settings, short in-context examples drawn from the human-annotated training data. This combination improved interpretability and encouraged models to emulate human reasoning during bias labeling. All prompts followed a unified structure:

> *"You are an expert annotator specializing in bias detection for social media news content about the Israel-Palestine conflict. Classify the [TEXT] into one of seven categories using these detailed guidelines [GUIDELINES]"*

Outputs and Agreement Evaluation. Each model produced a machine-annotated dataset with predicted labels and confidence probabilities for each class. These were aligned with human annotations to compute agreement metrics including Cohen's κ, accuracy, precision, recall, and F1-score. Further qualitative evaluation examined systematic divergences between human and model decisions (e.g., sarcasm, implicit bias, or culturally dependent framing). This comparative framework enables a comprehensive assessment of how annotation consistency varies by model family, prompting regime, and the degree of human conditioning.

5 Results

5.1 Performance of Fine-Tuned Models

Table 1 summarizes the classification performance of all fine-tuned models on the seven-way bias detection task. The transformer-based architectures substantially outperform the recurrent baseline, underscoring the benefit of contextual embeddings for political bias detection.

Across all models, RoBERTa achieved the highest macro F1-score (0.774), narrowly surpassing BERT (0.764). The improvements in precision and recall demonstrate the effectiveness of dynamic masking and larger pretraining data in RoBERTa. While LLaMA-3.1 achieved competitive results given its quantized and parameter-efficient setup, it

Model	Accuracy	Precision	Recall	F1	Loss
BiLSTM	0.567	0.565	0.567	0.562	1.36
BERT	0.762	0.772	0.762	0.764	0.84
RoBERTa	**0.772**	**0.785**	**0.772**	**0.774**	0.81
LLaMA-3.1	0.693	0.691	0.693	0.691	**0.79**

Table 1: Performance of fine-tuned models on the FIGNEWS dataset. Metrics are reported on the held-out test set. RoBERTa achieves the best overall performance across all metrics, while BiLSTM serves as a non-contextual baseline.

underperformed transformer baselines due to limited fine-tuning epochs and smaller batch sizes due to lack of computational resources. The BiLSTM model lagged behind, highlighting the importance of contextual representation learning for capturing subtle ideological cues.

A paired t-test between BERT and RoBERTa outputs revealed that the performance difference was not statistically significant ($p > 0.05$), confirming both models' strong and comparable reliability. However, performance differences across classes suggest that models identify overt polarity, but struggle with implicit or multi-actor framing, as shown by confusion concentrated among *Unbiased*, *Biased Against Palestine*, and *Biased Against Israel* labels. This indicates that the task is not purely lexical sentiment classification, but requires contextual interpretation that current models partially approximate.

5.2 Human vs. Machine Annotation Results

Unlike the fine-tuned classifiers, zero-shot models, especially GPT-based, demonstrated moderate alignment with human judgments ($\kappa = 0.338$), improving substantially when label definitions were included (guideline-conditioned prompting). However, Zero-shot RoBERTa and BERT tended to over-predict "Unbiased", suggesting a default neutrality stance when uncertainty is high. While FLAN-T5 showed higher sensitivity but low specificity, often detecting bias where humans did not. This demonstrates that model behavior varies more by inference strategy (zero-shot vs. guided) than by model size alone.

Quantitative Agreement. Table 2 presents agreement metrics across models and prompting regimes. Among the non-generative models, **RoBERTa-MNLI (zero-shot + guideline)** achieved the strongest alignment with human labels (Accuracy $= 0.419$, $\kappa = 0.133$), along with the lowest Jensen–Shannon divergence, indicating that guideline conditioning not only improves classification agreement but also calibrates the predicted label distribution toward human-like judgments. This steady improvement across architectures shows that *explicit operationalization of bias categories reduces ambiguity and promotes more stable decision boundaries*.

Label-wise Trends. All models had highest agreement with human annotators for *Unbiased* and *Biased against Palestine* labels and consistently struggled with *Biased against Both* and *Unclear* labels. This reflects the broader challenge of detecting implicit stance and multi-actor framing that require contextual reasoning beyond lexical or entity sentiment cues (Spinde et al., 2021).

Interpretation. Collectively, the results show that *accuracy alone obscures significant differences in annotation reasoning*. Guideline-conditioned prompting encourages models to adopt more human-like labels that reduce both over-neutral and over-biased predictions. However, the decline in performance in the few-shot setting especially for RoBERTa suggests that non-instruction-tuned models may treat in-context examples as noise when they are not aligned with their pretraining objectives. Meanwhile, FLAN-T5's degradation under few-shot prompting indicates a tendency toward surface-level pattern matching rather than conceptual generalization. These findings reinforce that human–machine alignment is shaped not only by model capacity but by how task semantics are represented and communicated during inference.

6 Result Analysis

6.1 Overview

The results reveal systematic differences in how models interpret political bias. Fine-tuned transformer architectures, particularly RoBERTa, achieved higher supervised performance than BiLSTM and zero-shot baselines. However, improvements in accuracy and macro-F1 were primarily driven by recognition of *overt sentiment polarity* rather than discourse-level stance reasoning. Even the strongest models struggled with implicit, multi-actor, and context-dependent bias, indicating that current approaches approximate bias via lexical cues rather than interpretive inference.

Model	Setup	Accuracy	Matthews corrcoef	κ	Jensen–Shannon divergence
BERT-MNLI	Zero-shot	0.094	0.068	0.015	0.601
BERT-MNLI	Zero-shot +Guideline	0.155	0.101	0.033	0.496
BERT-MNLI	Few-shot +Guideline	0.188	0.090	0.039	0.524
RoBERTa-MNLI	Zero-shot	0.128	0.105	0.049	0.581
RoBERTa-MNLI	Zero-shot +Guideline	**0.419**	**0.198**	**0.133**	**0.179**
RoBERTa-MNLI	Few-shot +Guideline	0.062	0.067	0.009	0.656
FLAN-T5	Zero-shot +Guideline	0.296	0.179	0.055	0.294
FLAN-T5	Few-shot +Guideline	0.322	0.177	0.055	0.233
XLM-R (large)	Zero-shot +Guideline	0.097	0.069	0.013	0.614
GPT-5.0	Zero-shot	**0.553**	**0.350**	**0.338**	**0.137**

Table 2: Agreement between human and model annotations. Metrics computed on the shared training set (N=2961). JS divergence measures label distribution distance (lower is better).

6.2 Analysis of Fine-Tuned RoBERTa Behavior

RoBERTa consistently outperformed BERT and BiLSTM in accuracy (0.772 vs. 0.762 and 0.567, respectively) and macro-F1 (0.774 vs. 0.764 and 0.562), confirming the benefits of contextualized embeddings and dynamic masking. Nevertheless, qualitative inspection revealed several recurring weaknesses.

Lexical and Framing Sensitivity RoBERTa frequently associated emotionally charged terms (e.g., *"terrorist"*, *"occupation"*) with directional bias regardless of narrative framing. Statements such as *"There are terrorists in my house"* were misclassified as *Biased against Palestine*, demonstrating reliance on lexical polarity rather than contextual attribution.

Entity-Centric Bias Inference Bias direction was often inferred from which actor appeared in negative proximity. Mentions of Israeli state actions skewed predictions toward *Biased against Palestine*, while references to Hamas skewed toward *Biased against Israel*. Balanced statements with explicit dual stances were frequently collapsed into singular polarity.

Implicit and Contextual Bias RoBERTa failed to infer bias expressed indirectly through questioning, critique, implication, or omission. For example, *"Gaza, a ground offensive doomed to failure?"* was labeled as *Biased against Israel* despite being an internal policy critique rather than ideological stance.

Ambiguity and Attribution Quoted speech was often treated as authorial opinion. Headlines attributing statements to political actors (*"Netanyahu: Israel will return to fighting..."*) were labeled as biased, indicating absence of speaker–narrator distinction. The model also defaulted to "Biased against both" under uncertainty, suggesting conservative over-generalization. These errors point to the need for discourse-aware objectives and attribution-sensitive fine-tuning.

Summary RoBERTa's fine-tuning improved explicit bias detection but exposed enduring weaknesses in neutrality recognition and mixed-stance comprehension. Its improvements reflect stronger lexical precision and not genuine stance reasoning. The model fails to interpret multi-source narratives, attribution, or how framing works, which are key parts of political bias.

6.3 Human–Machine Divergence (RoBERTa-MNLI)

Comparison between human and RoBERTa-MNLI (zero-shot with guidelines) annotations show the persistent gap between rule-conditioned models and human inference. The model achieved 41.9% accuracy and $kappa$ = 0.13 which is only slight agreement beyond chance. It under-detected bias in over half of "Biased against Palestine" cases and 42% of "Biased against Israel" cases, often labeling them "Unbiased." Conversely, it over-predicted bias in one-third of truly neutral examples that highlight asymmetry between sensitivity and specificity. Distributional analysis showed overuse of "Unbiased" (57% vs. 41% in human labels) and underuse of "Unclear" or "Not Applicable," with a Jensen–Shannon divergence of 0.046. This imbalance reflects over-confidence in assigning directional bias even under semantic uncertainty.

Qualitative Divergence Representative examples reveal systematic misalignment: (1) factual event descriptions misinterpreted as biased due to

lexical triggers; (2) overtly partisan hashtags dismissed as "Unclear"; and (3) off-topic statements forced into directional labels. These errors reinforce that zero-shot transformers rely on surface cues instead of evaluative intent which confirms findings from prior alignment work (Plank, 2022a; Movva et al., 2024).

6.4 GPT-5 Zero-Shot Performance

The GPT-5 model achieved 55.3% accuracy ($kappa$ = 0.34, weighted-F1 = 0.57), outperforming RoBERTa-MNLI and XLM-R on the same task. Without explicit supervision, GPT-5 effectively captured broad sentiment polarity and overt bias cues, particularly for high-frequency categories such as "Unbiased" (F1 = 0.62) and "Biased against Palestine" (F1 = 0.62). However, the model exhibited a critical tendency to default to the "Unbiased" classification when presented with nuanced or low-resource classes (like "Biased against Others" or "Unclear"). This preference for neutrality under conditions of uncertainty, rather than making a categorical commitment, reflects the model's underlying mechanism of probabilistic averaging, a behavior likely rooted in its instruction-tuning safety optimizations.

Cross-Model Comparison GPT-5 was better at maintaining contextual consistency when dealing with instances that involved multiple actors than RoBERTa, which tended to rely too much on basic rules for identifying entities. However, GPT-5's overall interpretation of the text varied more widely. While RoBERTa performed well due to its specialized training for precise word choice, GPT-5's main advantage was its ability to perform tasks it hadn't been explicitly trained on (zero-shot adaptability) and its broad applicability across different subjects. Nevertheless, GPT-5's lack of specific training for particular tasks meant it struggled to be consistent when handling unclear or ambiguous situations.

6.5 GPT-5 Variant Analysis

Model	Acc.	M-F1	W-F1	κ	Sup.
GPT-5 (ZS)	0.68	0.12	0.77	0.09	877
GPT-5 (ZS+GL)	0.63	0.13	0.72	0.08	656
GPT-5 (FS+GL)	0.58	0.13	0.69	0.07	714

Table 3: Performance comparison of GPT-5 variants on a sample subset. **ZS** = Zero-Shot, **GL** = Guideline, **FS** = Few-Shot.

Only a subset of 600–700 samples was tested across GPT-5 configurations. It is important to note that the evaluated subset was highly imbalanced, with the majority of samples labeled as *Unbiased* or *Unclear*, while the explicitly biased categories were sparsely represented. Zero-shot GPT outperformed guideline- and few-shot variants, suggesting that additional instruction constrained its broader representational inference and introduced label conservatism. This reflects a known effect of safety-aligned instruction tuning, where models are discouraged from strong interpretive claims unless high certainty is present.

6.6 Where Humans and Models Diverge

Across all models and inference settings, three systematic divergence patterns emerged between human and machine bias perception (Table 4). These patterns indicate that even when label agreement appears moderate, the *underlying reasoning strategies* differ substantially.

Divergence Type	Model Behavior	Human Interpretation
Lexical sentiment cues	Bias inferred from emotionally loaded words (e.g., *terrorist"*, *occupation"*) regardless of context	Bias judged based on *source attribution, intent*, and *narrative framing*, not single-word polarity
Entity association heuristics	Bias direction often tied to which actor receives negative descriptive language	Humans evaluate speaker identity, quoted sources, and rhetorical stance before inferring bias
Implicit / multi-actor framing	Ambiguous or balanced texts frequently labeled as *Unbiased*	Humans infer stance through emphasis, omission, and relational framing across actors

Table 4: Systematic reasoning differences between human and model bias judgments.

These divergences show that current models primarily rely on surface-level lexical and entity-association cues, whereas human annotators use contextual, pragmatic, and discourse-level reasoning.

Thus, even when overall accuracy is high, models fail to detect genuine political bias. **Ultimately, models are not detecting political bias itself, they are detecting local sentiment polarity toward named entities.**

This methodological discrepancy explains several systematic failure modes observed in model performance:

- misclassification of neutral but emotionally toned statements,
- failure on multi-actor narratives, and
- over-reliance on words rather than narrative framing.

This analysis strongly motivates the need for a new generation of models capable of *representing political stance as a comprehensive discourse structure*, moving decisively beyond mere token-level sentiment analysis.

6.7 Comparative Insights

Synthesizing across models, several trends emerge: fine-tuned transformer models offer stability but exhibit only surface-oriented performance, while Large Language Models (LLMs) are more adaptable yet prone to interpretive drift. The core differences between human and machine analysis typically revolve around the models' inability to handle subtle contextualization, accurately manage quotation handling, and detect implicit bias. Ultimately, increasing model scale only improves the coverage or breadth of detection, not the interpretive fidelity. Closing this critical gap requires developing hybrid frameworks that integrate human-level contextual reasoning, advanced discourse-level modeling, and specific, bias-aware objectives to ensure a robust and trustworthy system for political-bias detection.

7 Limitations

While the present study offers meaningful insights into human–model divergence in bias perception, several limitations should be noted. First, the research is focused exclusively on the Israel-Palestine conflict, which is highly complex and emotional, meaning the results might not apply to other political topics. Second, since political bias is subjective, the human-assigned labels used represent a consensus among annotators, not an absolute, single truth. Third, computational constraints restricted fine-tuning experiments to medium-scale models; larger architectures may exhibit different alignment behavior. Finally, the study only examined textual data, neglecting how bias is often amplified or conveyed through multimodal elements like images and overall network effects. Addressing these factors presents a valuable direction for future research.

8 Future Work

Future work will expand this framework beyond a single conflict domain to more diverse datasets and languages, testing cross-cultural robustness. With greater computational resources, larger LLMs and ensemble approaches can be evaluated to examine the effect of model scale on human–machine alignment. We also plan to incorporate discourse-aware objectives and entity-relation modeling to capture implicit and multi-actor bias. Integrating explainability and human-in-the-loop calibration could further improve interpretability. Finally, multimodal extensions combining textual, visual, and social-network signals may offer a richer understanding of media bias dynamics.

9 Conclusion

This study provides a comparative analysis of political bias perception across human annotators and multiple language model architectures. While RoBERTa achieved the highest supervised performance and GPT-based models demonstrated strong zero-shot adaptability, our analysis shows that alignment in labels does not equate to alignment in reasoning. Transformer models primarily detect lexical and sentiment cues, whereas human annotators interpret bias through contextual framing, attribution, and implied stance. These divergences suggest that improving automated bias detection requires moving beyond surface-level text patterns toward discourse-level reasoning and narrative structure modeling. Future work should focus on incorporating multi-actor relational context, rhetorical framing cues, and interpretability-driven training objectives to bridge the gap between human and model bias interpretation.

10 Ethics Statement

All data were obtained from the publicly available FIGNEWS 2024 corpus under its research license. No personally identifiable information was used. Annotators were briefed on ethical considerations and potential emotional impact due to conflict-related content.

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 4982–4991, Online. Association for Computational Linguistics.

Ljubisa Bojic, Olga Zagovora, Asta Zelenkauskaite, Vuk Vukovic, Milan Cabarkapa, Selma Veseljević Jerkovic, and Ana Jovančevic. 2025. Evaluating large language models against human annotators in latent content analysis: Sentiment, political leaning, emotional intensity, and sarcasm. *Preprint*, arXiv:2501.02532.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15607–15631. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fan Fan, Jiaqi Zhou, Yifan Wu, and Chenliang Li. 2019. Hierarchical neural network for stance and bias detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3414. 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. *Preprint*, arXiv:2305.08283.

FIGNEWS. 2024. Fignews 2024: News media narratives dataset. `https://sites.google.com/view/fignews/home`. Part of the The Second Arabic Natural Language Processing Conference (ArabicNLP 2024) Co-located with ACL 2024.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. volume 120. Proceedings of the National Academy of Sciences.

Julia Jaremko, Dagmar Gromann, and Michael Wiegand. 2025. Revisiting implicitly abusive language detection: Evaluating LLMs in zero-shot and few-shot settings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3879–3898, Abu Dhabi, UAE. Association for Computational Linguistics.

Cinoo Lee, Shibani Santurkar, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Mind the opinion gap: Understanding what llms miss in capturing human disagreement. *arXiv preprint arXiv:2310.01980*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased? assessing the representativeness of twitter's streaming api. *Preprint*, arXiv:1401.7909.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing llm and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062. Association for Computational Linguistics.

Simon Münker, Kai Kugler, and Achim Rettinger. 2024. Zero-shot prompt-based classification: topic labeling in times of foundation models in german tweets.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

Barbara Plank. 2022a. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank. 2022b. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, and et al. 2024. The fignews shared task on news media narratives. *arXiv preprint arXiv:2407.18147*.

A Appendix

A.1 Guidelines

Unbiased: Content labeled "Unbiased" presents information objectively, avoiding favoritism or prejudice. While some word choices might hint at the author's per spective, the content remains unbiased if these choices are contextually relevant and don't significantly distort the overall narrative. An article is considered unbiased if it offers fair representation of all sides, avoids misleading language, and doesn't express opinions favoring a particular viewpoint. For example, the statement "Both parties have expressed their willingness to nego tiate, according to international mediators" presents the viewpoints of both sides equally, without favoring one over the other. Attributing the information to an ex 21 ternal source (international mediators) further enhances neutrality. Similarly, the sentence "The Palestinian death toll in Gaza from the war between Israel and the territory's Hamas rulers has soared past 25,000, the Gaza Health Ministry said" reports the death toll neutrally, avoiding blame or bias. Attributing the information to the Hamas-controlled health ministry ensures that subjective content is sourced externally. The statement "South Africa has filed an application at the International Court of Justice to begin proceedings over allegations of genocide against Israel" is also neutral, as it attributes the information to the International Court of Justice. The facts are presented without bias, allowing the audience to form their own conclusions.

Unclear: The label "Unclear" is applied to those articles where it is difficult to determine the bias due to vague language, incomplete information or insufficient context. These articles often discuss sensitive issues without clearly identifying which side of the argument they are supporting or opposing. This can happen in media coverage of conflicts where the language avoids direct attribution on purpose. For instance, the quote " The story of the battlefield... in the words of ZEE NEWS The great victory of 'Operation Ajay' Watch the time fixed to meet the soil of 'Gaza' #Deshhit LIVE with #ChandanSingh #IsraelHamasWar #Gaza #IsraelArmy #OperationAjay #ZeeNews #ZeeLive" mentions a battlefield and a success ful operation, but doesn't say who was involved, or the broader context. It's difficult to determine whether the statement carries bias or not or who it might be biased against. Additionally, texts or titles that are too short, such as "Losing support" or "He was a real hero" are labeled as UNCLEAR. These texts may have links to other media or contain too little information to evaluate properly.

Biased against Palestine: A text was labeled biased against Palestine if the text unfairly omitted or downplayed Palestinian viewpoints, had only one sided rep resentation, used emotionally charged or pejorative language against Palestinians such as referring them as terrorists, exhibited facts that support a negative image of Palestine while ignoring justification and legitimate grievances, focused on negative stereotypes and generalizations of the Palestinians. For example: "LIVE: Presi dent Joe Biden Responds to Hamas Terrorist Attack on Israel. . . ". Here the label "terrorist" without providing context for Palestinian grievances reflects a one-sided portrayal, leading to bias. Another instance: "The singer Idan Raichel attacked: "The Palestinians are not rebelling against Hamas- therefore most of them should be treated as involved."" "They could have been brave, entered the tunnels tonight and revolted," added Reichel". In this text all Palestinians have been generalized as complicit for not rebelling against Hamas which unfairly stereotypes and dehuman izes the broader population.

Biased against Israel: This label has been used for texts which portrayed Israel in a negative or distorted manner such as Zionist or occupation, used emotionally charged language to describe Israel without context or explanation, highlighted Is raeli military actions without justification, portrayed Israelis in a way that dehuman izes them or justifies violence against them, disregarded Israel's legitimate concerns regarding security and unfairly portrayed Israeli actions as purely aggressive. Ex 22 ample: "Israeli gunboats violate the truce and bomb the coast of Khan Yunis". This text focuses solely on Israeli military actions, portraying them as aggressive without context or explanation of potential security concerns. Another example: "Urgent | "Osama Hamdan, a leader in Hamas":- "The Israeli aggression has written its mili tary and political end"- "The occupation leaders will not see their prisoners held by the resistance alive until after a comprehensive cessation of the aggression."". This text refers to Israel's actions as "aggression" and "occupation" without acknowledg ing Israel's security context, dehumanizing Israeli leadership and justifying violence against them.

Biased against both Palestine and Israel: The label "Biased against both Palestine and Israel" is

used for articles that paint both sides of the conflict in an unfavorable light, often ridiculing or undermining the actions of both Palestine and Israel. For example, the statement "LIVE | Sea, land, sky... Israel and Hamas clashed, Hamas fired 5 thousand rockets... Israel opened ammunition stockpile" suggests both parties are engaged in harmful actions, without favoring one side. Another example is "In the latest chapter of their endless and futile conflict, Israeli and Palestinian forces have once again engaged in senseless violence, showing a com plete disregard for peace or human life." This type of article scrutinizes the actions of both sides and expresses opinions that both sides' irresponsibility is the cause for the breakdown of peace. The media supports a narrative that suggests both parties are equally responsible for the violence and instability by portraying the dispute in this way.

Biased against others: A text has been labeled as "Biased against others" if it portrayed entities other than Israel or Palestine in a negative, distorted, or unfair manner by misrepresenting their roles or intentions in the Israel-Palestine conflict. For instance: "Iranian President Raisi meets Putin, accuses US of genocide in Gaza." This text criticizes the US with loaded language like "genocide" without providing balanced or factual evidence, leading to misrepresentation of its role. Similarly the text "France, Germany, England, Austria, the entire EU, the USA... are responsible for the bombing of a hospital in Gaza by the Israeli army!!.." he text unfairly blames multiple international actors for an Israeli military action, misrepresenting their in volvement in the conflict and promoting bias against them. Such texts are labeled "Biased against others" when they target groups or countries apart from Israel and Palestine with unfair or emotionally charged portrayals.

Not Applicable: Articles or headlines have been labeled as "Not Applicable" if the text didn't pertain to the Israel-Palestine conflict. These texts contained unrelated topics which didn't require any bias analysis and even if the texts were biased, they were irrelevant to the Israel Palestine conflict or didn't mention Israel, Palestine, or any entities involved in the conflict. For instance: "Karine Jean-Pierre briefs the press alongside John Kirby.", "In the United Kingdom, the BBC is under fire: red paint was even sprayed on the facade of the group."