BiasExpert: Applying the Less-Is-More Principle to Media Bias Detection in News Articles

Wagner Costa Santos[1,2], **Elin Törnquist**[2],
Arnaldo Candido Junior[1], **Robert Alexander Caulk**[2]

[1] Institute of Biosciences, Humanities and Exact Sciences
São Paulo State University (UNESP)
São José do Rio Preto, SP – Brazil

[2]Emergent Methods – Arvada, CO – USA

{wagner, elin, rob}@emergentmethods.ai, arnaldo.candido@unesp.br

Abstract. *This research addresses the critical challenge of media bias detection through the development of a specialized reasoning-enhanced Large Language Model (LLM). Leveraging the advanced analytical capabilities of modern LLMs, we propose a novel approach that combines reasoning mechanisms with bias detection frameworks to create more transparent and objective news content analysis. Our methodology employs a model consensus strategy using multiple reasoning-capable LLMs (Claude 3.7, DeepSeek-R1, o3-mini, and Gemini 2.5) to generate a curated dataset derived from the MN-DS news corpus. This consensus-driven approach ensures robust bias identification across various news categories while maintaining balanced representation. We fine-tune the Qwen3 4B model using Parameter-Efficient Fine-Tuning (PEFT) with QLoRA techniques, demonstrating that smaller models can effectively acquire reasoning and bias detection capabilities when trained on high-quality examples. Our findings support the "Less-Is-More" hypothesis for reasoning, suggesting that sophisticated bias analysis can emerge without reinforcement learning when models are exposed to well-structured demonstrations. This work contributes to the advancement of more ethical journalism by providing a transparent framework for bias detection in news articles.*

1. Introduction

1.1. Problem Definition

The proliferation of digital news media has exponentially increased information consumption, highlighting the critical need for transparent and unbiased journalism. News articles inherently contain various forms of bias that can significantly influence public opinion and decision-making [Spinde et al. 2021]. Despite the journalistic ideal of objectivity, complete neutrality remains elusive due to the inherent subjectivity in language, framing, and topic selection [Hamborg et al. 2019].

Media bias manifests in multiple forms, including framing bias, where specific aspects of a story are emphasized; linguistic bias, evidenced through loaded language and subjective adjectives; and selection bias, which determines which stories receive coverage [Spinde et al. 2021]. The boundary between legitimate opinions and problematic bias remains challenging to define precisely, creating a complex landscape for automated

detection systems [Hamborg et al. 2019]. This ambiguity necessitates sophisticated analytical approaches that can navigate the nuanced spectrum of media content.

The detection and mitigation of bias in news content serve several crucial functions: (1) enhancing readers' awareness of potential slants in reporting, (2) supporting journalists in producing more balanced content, and (3) promoting a more informed democratic discourse by ensuring access to less skewed information [Raza et al. 2022]. However, existing approaches to bias detection often rely on simplistic keyword analysis or require extensive labeled datasets that are costly to produce and may themselves contain biases.

1.2. Proposed Approach

Our research leverages recent advancements in Large Language Models (LLMs), which have demonstrated remarkable capabilities in text analysis and classification tasks through few-shot learning paradigms [Brown et al. 2020]. Particularly promising are the enhanced reasoning abilities of modern LLMs, which enable them to engage in step-by-step analytical processes before producing outputs [OpenAI et al. 2024, DeepSeek-AI et al. 2025].

We propose a novel approach that leverages the reasoning capabilities of LLMs for sophisticated text analysis in the context of bias detection. Our methodology employs a three-stage process:

First, we develop a comprehensive taxonomy of media bias by integrating classifications from AllSides [Mastrine et al. 2019], the multi-dimensional framework by [Rodrigo-Ginés et al. 2024], and other established sources [Spinde et al. 2021, Raza et al. 2022]. This integrated taxonomy enables standardized identification of 18 diverse bias types across news content (see Appendix A for the complete taxonomy).

Second, we implement a model consensus strategy to generate a high-quality dataset. Starting with the MN-DS multilabeled news dataset [Petukhova and Fachada 2023], we extract a balanced subset of 1,220 articles across various categories and subcategories. Four reasoning-capable LLMs, Claude 3.7 [Anthropic 2025], DeepSeek-R1 [DeepSeek-AI et al. 2025], o3-mini [OpenAI 2025], and Gemini 2.5 [Google 2025]—analyze these articles using identical prompts to identify and classify biases at four distinct levels of granularity. This structured output is captured in JSON format, allowing for systematic comparison and validation. We employ a distance-based consensus mechanism with Claude 3.7 as our baseline, establishing a threshold that ensures only high-agreement data points are included in our final dataset.

Third, we fine-tune the Qwen3 4B model [Yang et al. 2025] using Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA (Low-Rank Adaptation) [Han et al. 2024, Hu et al. 2021, Dettmers et al. 2023]. This approach enables us to efficiently transfer bias detection capabilities to a smaller model while maintaining high performance. Our methodology builds on the "Less-Is-More" reasoning hypothesis proposed by [Ye et al. 2025], which suggests that sophisticated reasoning capabilities can emerge from minimal but well-structured demonstrations without reinforcement learning, challenging the conventional approach of using RL for enhancing reasoning in LLMs. Our model is designed to output the complete reasoning process, providing transparency into how it arrives at bias classifications. By explicitly documenting each step of anal-

ysis, from identifying linguistic patterns to evaluating framing choices, the model offers users insight into the decision-making process rather than merely presenting conclusions. This transparency serves multiple purposes: It allows users to understand the specific elements that contribute to bias detection, enables verification of the model's reasoning, and provides educational value by demonstrating systematic bias analysis. The explicit reasoning output also helps mitigate the "black-box" problem common in AI systems, fostering greater trust in the model's assessments while encouraging users to improve their own critical evaluation skills when consuming news media.

This research contributes to the fields of computational linguistics and automated media analysis by demonstrating how reasoning capabilities can be effectively applied to the complex task of bias detection, potentially transforming how we evaluate and consume news media. By creating a model that provides structured and transparent analysis of bias in news articles, our goal is to enhance journalistic integrity and readers' critical awareness of media bias in an increasingly complex information landscape.

2. Methodology

Our methodology comprises two main phases: dataset engineering with consensus-based validation and model fine-tuning using Parameter-Efficient Fine-Tuning techniques, as illustrated in Figure 1. This approach ensures high-quality training data while maintaining computational efficiency in the model development process.

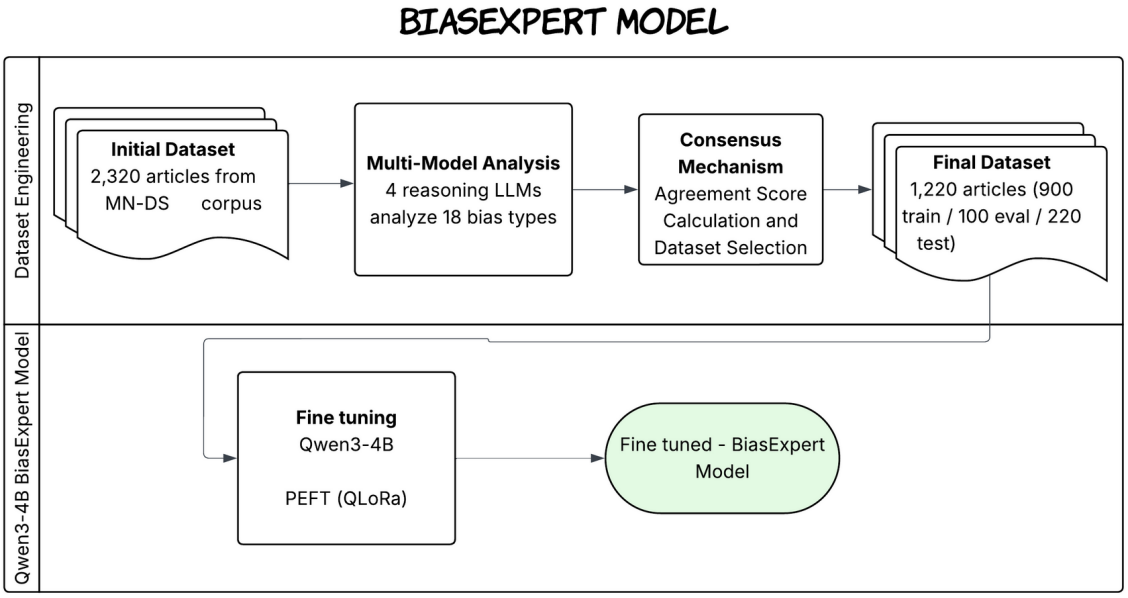

Figure 1. BiasExpert methodology overview showing the two main phases: dataset engineering with consensus-based validation and model fine-tuning using Parameter-Efficient Fine-Tuning techniques.

2.1. Dataset Engineering and Consensus Mechanism

We initiated our data collection process with 2,320 articles from the MN-DS multilabeled news dataset [Petukhova and Fachada 2023], obtained through two collection rounds: an initial batch of 1,500 articles followed by an additional 820 articles to reach our target dataset size. Our sampling strategy prioritized balanced representation across news

categories by grouping articles according to their category and subcategory fields, then sampling proportionally from each group to maintain categorical diversity. Articles were filtered to include only those with more than 400 words to ensure sufficient content for bias analysis.

Four reasoning-capable Large Language Models were employed to analyze each article: Claude 3.7 [Anthropic 2025], DeepSeek-R1 [DeepSeek-AI et al. 2025], o3-mini [OpenAI 2025], and Gemini 2.5 [Google 2025]. Each model received identical prompts to identify and classify 18 distinct bias types across four granularity levels: None (0), Low (1), Moderate (2), and High (3).

2.1.1. Agreement Score Calculation

To ensure dataset quality, we implemented a distance-based agreement scoring mechanism. First, we mapped qualitative bias levels to numerical values as shown in Table 1.

Table 1. Bias level mapping to numerical values

Bias Level	Numerical Value
None	0
Low	1
Moderate	2
High	3

The agreement score A between two models i and j for a specific bias type is calculated using the following formula:

$$A(i,j) = \begin{cases} 1.0 & \text{if } L_i = L_j \\ 0.75 & \text{if } L_i > 0, L_j > 0 \text{ and } |L_i - L_j| = 1 \\ 0.5 & \text{if } L_i > 0, L_j > 0 \text{ and } |L_i - L_j| = 2 \\ 0.0 & \text{if } (L_i = 0 \text{ and } L_j \neq 0) \text{ or } (L_i \neq 0 \text{ and } L_j = 0) \end{cases} \tag{1}$$

where L_i and L_j represent the numerical bias levels assigned by models i and j respectively. This relationship is also illustrated in the agreement matrix shown in Table 2.

Table 2. Agreement score matrix between bias level assessments

Model 1 \Model 2	None	Low	Moderate	High
None	1.00	0.00	0.00	0.00
Low	0.00	1.00	0.75	0.50
Moderate	0.00	0.75	1.00	0.75
High	0.00	0.50	0.75	1.00

This scoring system assigns perfect agreement (1.00) for identical classifications, high agreement (0.75) for adjacent bias levels within the bias spectrum, moderate agree-

ment (0.50) for bias levels differing by two steps, and zero agreement (0.00) when one model detects no bias while another detects any level of bias.

The overall agreement score for each article is computed by averaging all pairwise agreement scores across models and bias types:

$$S_{overall} = \frac{1}{18} \sum_{b=1}^{18} \frac{1}{6} \sum_{\text{all pairs}} A(M_i, M_j)_b \tag{2}$$

where b represents each of the 18 bias types, the sum covers all six possible pairwise combinations of the four models, and $A(M_i, M_j)_b$ is the agreement score between models i and j for bias type b.

2.1.2. Dataset Selection Process

Our dataset selection procedure involved two sequential filtering steps:

Step 1 - Statistical Outlier Removal: We removed articles with agreement scores below -2 standard deviations from the mean.

Step 2 - Bias Confirmation Requirement: We implemented a confirmation mechanism using Claude 3.7 as our baseline model (selection rationale detailed in 3.1.1). For each bias type present in an article, we required that at least one additional model must confirm the baseline model's bias classification to ensure reliability.

The confirmation rule was defined as:

- If Claude 3.7 classifies a bias as present (level > 0), at least one of the three remaining models must also classify it as present
- If Claude 3.7 classifies a bias as absent (level $= 0$), at least one of the three remaining models must also classify it as absent
- Articles where Claude 3.7's classification cannot be confirmed by any other model for any bias type are excluded from the final dataset

This baseline-confirmation approach ensures that bias detection is not based on single-model predictions, thereby reducing hallucinations and improving dataset reliability. The filtering process resulted in a utilization rate of 1,220 out of 2,320 articles (52.6%), significantly improving dataset quality by ensuring multi-model consensus on bias classifications.

2.1.3. Final Dataset Composition

The cleaned dataset was partitioned as follows:

- Training set: 900 articles (73.8%)
- Validation set: 100 articles (8.2%)
- Test set: 220 articles (18.0%)

2.2. Model Fine-Tuning

We fine-tuned the Qwen3 4B model [Yang et al. 2025] using Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically employing QLoRA (Quantized Low-Rank Adaptation) [Dettmers et al. 2023]. The training process utilized the 900-article training set, with the 100-article validation set employed for checkpoint selection. The process employed LoRA with a rank of 32 and alpha scaling factor of 64 ($2\times$ rank), targeting all linear layers in the model architecture. Training was conducted over 4 epochs with a learning rate of 2×10^{-4} using the paged AdamW optimizer. The LoRA dropout rate was set to 0.05 to prevent overfitting, while gradient accumulation steps of 4 enabled effective batch processing within memory constraints.

Evaluation Metrics: Model performance was evaluated based on agreement scores with Claude 3.7 and JSON output validity. JSON validity is crucial as our models are required to produce structured outputs following a specific schema for bias classification. Invalid JSON outputs indicate parsing failures that render the analysis unusable, making this a critical reliability metric alongside agreement scores.

The fine-tuning process was designed to transfer the reasoning and bias detection capabilities demonstrated by the larger models to the more computationally efficient Qwen3 4B architecture, following the "Less-Is-More" principle for reasoning [Ye et al. 2025]. This approach enables the model to generate transparent, step-by-step reasoning processes while maintaining high performance in bias classification tasks.

3. Results

Our results are presented in two main parts: dataset engineering outcomes demonstrating the effectiveness of our multi-model consensus approach, and fine-tuning results showing the performance of the BiasExpert model.

3.1. Dataset Engineering Results

The dataset engineering phase involved comprehensive analysis of model agreement patterns and validation of our consensus-based filtering approach. We present both the multi-model agreement analysis and the effectiveness of our data selection procedures.

3.1.1. Multi-Model Agreement Analysis

We conducted a comprehensive agreement analysis across all four reasoning-capable LLMs using the complete dataset of 2,320 articles before applying our cleaning procedures. The pairwise agreement scores between models are presented in Table 3.

The analysis reveals that Claude and Gemini 2.5 achieved the highest pairwise agreement score of 0.827, indicating strong consensus in their bias assessments. The second highest agreement was observed between Gemini 2.5 and o3-mini (0.808), further reinforcing Gemini 2.5's position as the most consistently aligned model across different reasoning approaches. To evaluate overall model coherence, we calculated each model's average agreement with all other models, as shown in Table 4.

While Gemini 2.5 demonstrated the highest overall coherence score of 0.813, we selected Claude 3.7 as our baseline model for the consensus mechanism due to a critical requirement: our methodology requires explicit reasoning text to train the Qwen3

Table 3. Pairwise agreement scores between LLM models. The highest scores are indicated in bold.

Model Pair	Claude	DeepSeek-R1	Gemini 2.5	o3-mini
Claude	-	0.803	**0.827**	0.791
DeepSeek-R1	0.803	-	0.804	0.784
Gemini 2.5	**0.827**	0.804	-	0.808
o3-mini	0.791	0.784	0.808	-

Table 4. Model coherence scores (average agreement with other models). Selected baseline model is highlighted.

Model	Coherence Score
Claude 3.7	**0.807**
DeepSeek-R1	0.797
Gemini 2.5	0.813
o3-mini	0.794

4B model to generate transparent, step-by-step bias analysis. At the time of our experiments, only Claude 3.7 and DeepSeek-R1 provided detailed reasoning tokens through their API responses, while Gemini 2.5 and o3-mini did not expose their internal reasoning processes. Claude 3.7's strong coherence score of 0.807, combined with its essential reasoning outputs, made it the optimal choice for training the BiasExpert model to articulate bias detection decisions with the transparency and interpretability that are core objectives of our approach.

3.1.2. Dataset Cleaning Effectiveness

The two-step filtering process successfully reduced the dataset from 2,320 to 1,220 articles (52.6% utilization rate) while significantly improving data quality. The statistical outlier removal (Step 1) eliminated articles with agreement scores below the -2 standard deviation threshold. The confirmation requirement (Step 2) removed articles where at least one bias type lacked multi-model confirmation, ensuring that all retained data points represent genuine consensus among the reasoning models.

3.2. Fine-Tuning Results

We evaluated our fine-tuned model, denoted **Qwen3/4B BiasExpert**, against baseline Qwen3 models across multiple dimensions. Table 5 presents a comprehensive comparison of model performance on the test set of 220 articles.

To ensure fair comparison, we conducted an additional evaluation using only the 207 articles where both models produced valid JSON outputs. Table 6 shows the head-to-head performance between our fine-tuned model and the larger Qwen3/32B baseline.

This controlled comparison confirms that our fine-tuned 4B model consistently outperforms the 32B baseline by 6.3% (0.8459 vs 0.7961), demonstrating that the "Less-Is-More" principle combined with high-quality reasoning examples can enable smaller

Table 5. Comprehensive model performance comparison. The highest scores are indicated in bold.

Model	Avg. Claude Agreement	Invalid JSON (%)	Thinking Length (words)	Claude Thinking Length (words)
Qwen3/4B BiasExpert	**0.8459**	13 (5.91%)	8009 ± 1750	9344 ± 2575
Qwen3/4B	0.7505	49 (22.27%)	5478 ± 1247	9344 ± 2575
Qwen3/32B	0.8004	0 (0%)	5008 ± 1048	9344 ± 2575

Table 6. Direct comparison on articles with valid JSON outputs (207 out of 220 test articles). The highest scores are indicated in bold.

Model	Avg. Claude Agreement
Qwen3/4B BiasExpert	**0.8459**
Qwen3/32B	0.7961

models to achieve superior performance compared to larger, non-specialized models.

3.3. Bias Pattern Analysis

Figure 2 provides a detailed visualization of pairwise agreement patterns across all 18 bias categories, including our fine-tuned model. The radar chart reveals that models achieve consistently high agreement across most bias types, with particularly strong consensus on demographic biases (political, gender, ethnic/cultural) and structural biases (statement bias, opinion-as-fact). The chart demonstrates that our **Qwen3/4B BiasExpert** model achieves agreement patterns very similar to Claude across all bias categories, indicating successful knowledge transfer. Some variation is observed in more subjective categories such as slant and source selection bias, reflecting the inherent complexity of these bias types.

The comprehensive agreement analysis across bias categories is further illustrated in Figure 3, which shows both pairwise and one-vs-others agreement statistics for all models including the fine-tuned versions. The analysis reveals that demographic bias categories achieve the highest agreement scores, with maximum pairwise agreement consistently above 0.95. More complex linguistic biases show greater variability, reflecting their nuanced nature. The inclusion of our fine-tuned model maintains the overall agreement patterns observed in the original dataset, with the **Qwen3/4B BiasExpert** model contributing positively to the consensus.

To understand model behavior patterns, we analyzed the sensitivity of each model to different bias categories, as shown in Figure 4. This analysis reveals distinct detection patterns: Claude and Gemini 2.5 show higher sensitivity to political and opinion-based biases, while DeepSeek-R1 and o3-mini demonstrate more conservative detection patterns. Our **Qwen3/4B BiasExpert** model exhibits sensitivity patterns closely aligned with Claude, further confirming successful knowledge transfer.

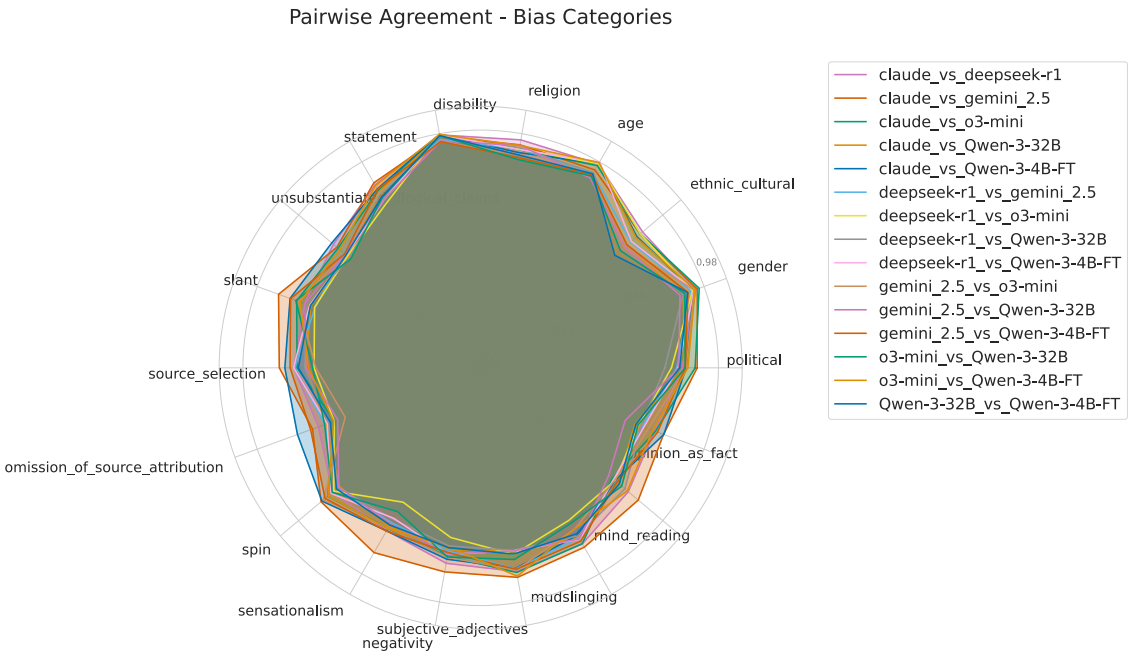

Figure 2. Pairwise agreement patterns across bias categories for the four reasoning-capable LLMs. The radar chart shows agreement scores (0-1 scale) for each of the 18 bias types, with different colored lines representing pairwise comparisons between Claude, DeepSeek-R1, Gemini 2.5, and o3-mini. Also showing successful knowledge transfer to fine tuned Qwen3-4B model.

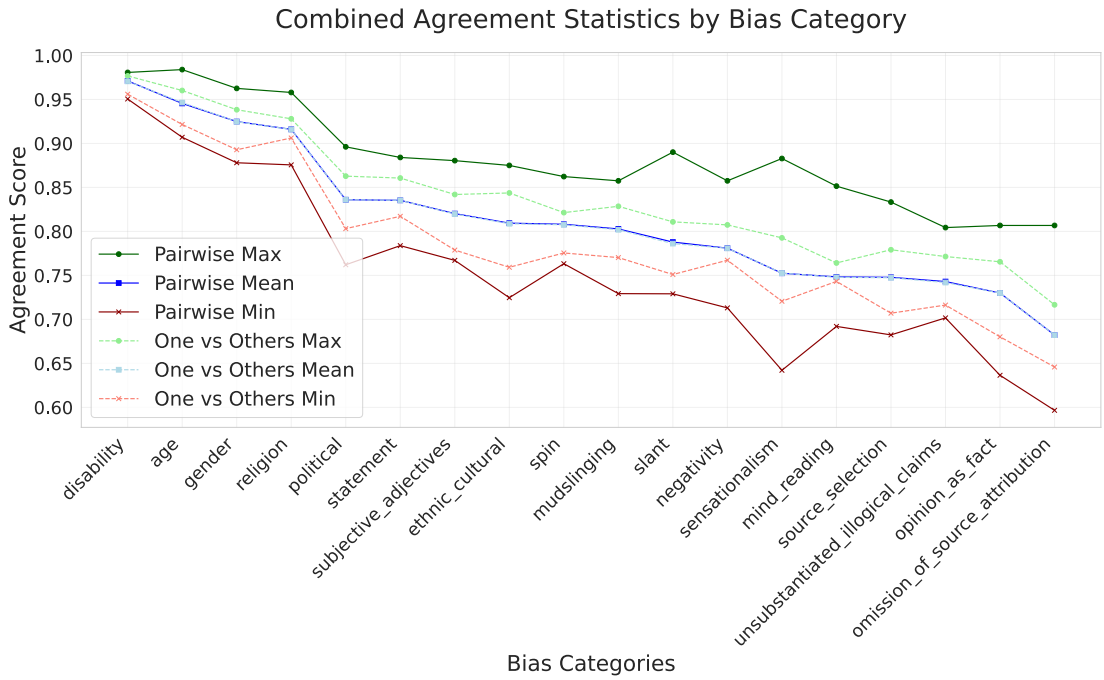

Figure 3. Combined agreement statistics by biaWrite something... s category for the original four LLMs showing pairwise and one-vs-others agreement patterns.

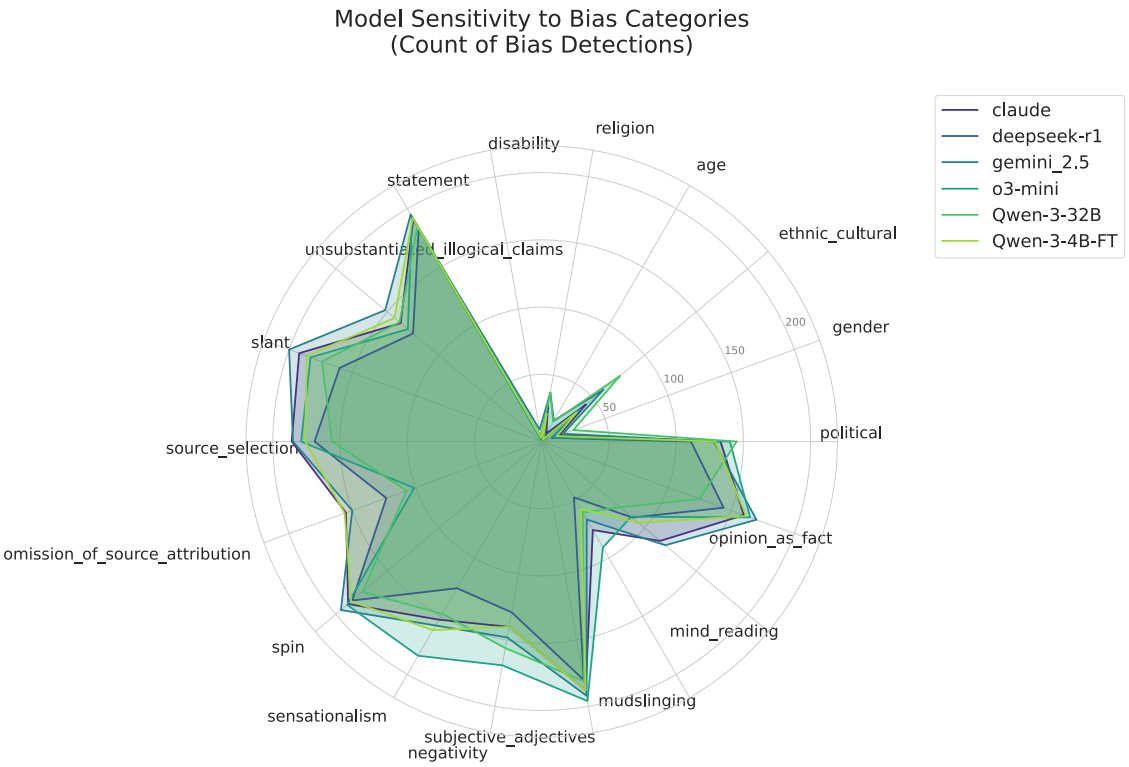

Figure 4. Model sensitivity to bias categories measured by detection frequency, including fine-tuned Qwen model, revealing distinct patterns across models.

3.4. Key Achievements

The fine-tuning results validate several core aspects of our approach. Most notably, our model achieved the highest agreement score (0.8459) with Claude among all tested models, indicating successful knowledge transfer from the reasoning examples. The approach demonstrated remarkable computational efficiency, showing that a 4B parameter model can outperform a 32B parameter model through targeted fine-tuning—representing an 8x reduction in model size while maintaining superior performance.

The model generates comprehensive reasoning outputs (8009 ± 1750 words) that approach the depth and detail of Claude's original reasoning, enabling transparency and interpretability. Additionally, it achieved significantly improved output reliability with a low invalid JSON rate (5.91%) compared to the baseline 4B model (22.27%), indicating enhanced structural consistency. The high agreement with Claude confirms that the model successfully learned to replicate the reasoning patterns and bias detection capabilities demonstrated in the training examples.

These results demonstrate that our consensus-based fine-tuning approach successfully creates a compact, efficient model that maintains the analytical depth and transparency of larger reasoning models while requiring significantly fewer computational resources.

4. Conclusion

This research successfully demonstrates that the "Less-Is-More" principle can be effectively applied to bias detection in news articles through a novel consensus-based fine-

tuning approach. Our work addresses the critical challenge of creating transparent, efficient, and reliable bias detection systems by combining multi-model consensus validation with targeted knowledge distillation.

4.1. Key Contributions

Our study makes several significant contributions to the field of automated bias detection:

Multi-Model Consensus Framework: We developed a robust methodology for generating high-quality training data through consensus among four state-of-the-art reasoning models (Claude 3.7, DeepSeek-R1, Gemini 2.5, and o3-mini). This approach achieved strong inter-model agreement scores, with the highest pairwise agreement of 0.827 between Claude and Gemini 2.5, validating the reliability of our consensus mechanism.

Efficient Model Architecture: Our fine-tuned Qwen3 4B model (BiasExpert) demonstrates that smaller, specialized models can outperform significantly larger general-purpose models. The BiasExpert model achieved a Claude agreement score of 0.8459, surpassing both the baseline Qwen3/4B (0.7505) and the 8x larger Qwen3/32B model (0.8004), representing computational efficiency gains while maintaining superior performance.

Reasoning Transparency: By leveraging Claude's explicit reasoning outputs as training examples, our model generates comprehensive analytical explanations (8009 ± 1750 words) that approach the depth of the original reasoning model (9344 ± 2575 words). This transparency enables users to understand and validate bias detection decisions, addressing a critical limitation of black-box approaches.

Robust Dataset Engineering: Our two-step filtering process successfully curated a high-quality dataset from 2,320 to 1,220 articles (52.6% utilization rate), removing statistical outliers and ensuring multi-model confirmation for all bias classifications. This methodology significantly improved data quality while maintaining dataset diversity across 18 bias types and four intensity levels.

4.2. Practical Implications

The results have important implications for real-world bias detection applications:

Scalability: The 4B parameter model requires significantly fewer computational resources than traditional large-scale approaches, making bias detection accessible to organizations with limited computational budgets while maintaining high accuracy.

Interpretability: The model's ability to provide detailed reasoning explanations addresses the critical need for transparency in bias detection systems, particularly important for journalistic applications where understanding the rationale behind bias classifications is essential.

Reliability: The low invalid JSON rate (5.91%) and high agreement scores demonstrate the model's reliability for production deployment, with consistent structured outputs suitable for automated processing pipelines.

4.3. Limitations and Future Work

While our approach demonstrates significant promise, several limitations warrant consideration. Our evaluation focused primarily on English-language news articles, and future work should explore the generalizability of this approach across different languages, domains, and cultural contexts to ensure broader applicability. Additionally, bias patterns and societal contexts evolve over time, necessitating long-term studies to assess the model's performance stability and determine the frequency of required retraining to maintain accuracy.

Our approach relies on the availability of reasoning-capable models with API access to reasoning tokens, which may limit its reproducibility and long-term sustainability. Future work could explore methods for generating synthetic reasoning examples or developing reasoning capabilities in open-source models to reduce this dependency. Furthermore, while our model addresses 18 bias types, the rapidly evolving media landscape may introduce new forms of bias that require continuous model updates and consensus validation to maintain comprehensive coverage.

4.4. Final Remarks

This research demonstrates that the combination of multi-model consensus and targeted fine-tuning can create efficient, transparent, and reliable bias detection systems. The BiasExpert model represents a significant step toward democratizing access to sophisticated bias analysis tools while maintaining the interpretability essential for responsible AI deployment in media analysis.

Our findings suggest that the "Less-Is-More" principle, when combined with high-quality consensus-validated training data, can enable smaller models to achieve performance levels that rival or exceed much larger systems. This approach opens new avenues for developing specialized AI systems that balance efficiency, transparency, and performance across various natural language processing tasks.

The methodology presented here provides a foundation for future research in consensus-based model training and offers practical solutions for organizations seeking to implement reliable, interpretable bias detection systems in their media analysis workflows.

References

[Anthropic 2025] Anthropic (2025). Claude 3.7 Sonnet and Claude Code.

[Brown et al. 2020] Brown et al. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[DeepSeek-AI et al. 2025] DeepSeek-AI et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs].

[Dettmers et al. 2023] Dettmers et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

[Google 2025] Google (2025). Gemini 2.5: Our most intelligent AI model.

[Hamborg et al. 2019] Hamborg et al. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

[Han et al. 2024] Han et al. (2024). Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv:2403.14608 [cs].

[Hu et al. 2021] Hu et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs].

[Mastrine et al. 2019] Mastrine et al. (2019). How to Spot 16 Types of Media Bias.

[OpenAI 2025] OpenAI (2025). OpenAI o3-mini.

[OpenAI et al. 2024] OpenAI et al. (2024). OpenAI o1 System Card. arXiv:2412.16720 [cs].

[Petukhova and Fachada 2023] Petukhova and Fachada (2023). MN-DS: A Multilabeled News Dataset for News Articles Hierarchical Classification. *Data*, 8(5):74. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[Raza et al. 2022] Raza et al. (2022). Dbias: Detecting biases and ensuring Fairness in news articles. arXiv:2208.05777 [cs].

[Rodrigo-Ginés et al. 2024] Rodrigo-Ginés et al. (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.

[Spinde et al. 2021] Spinde et al. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

[Yang et al. 2025] Yang et al. (2025). Qwen3 Technical Report. arXiv:2505.09388 [cs].

[Ye et al. 2025] Ye et al. (2025). LIMO: Less is More for Reasoning. arXiv:2502.03387 [cs].

A. Comprehensive Bias Type Taxonomy

This appendix provides a complete reference of the 18 bias types used in our BiasExpert model. The taxonomy integrates classifications from AllSides [Mastrine et al. 2019], Rodrigo-Ginés et al. [Rodrigo-Ginés et al. 2024], Raza et al. [Raza et al. 2022], and other established sources [Spinde et al. 2021, Hamborg et al. 2019]. Each bias type is defined with examples and source attributions to ensure comprehensive coverage of media bias manifestations in news content.

Table 7: Complete taxonomy of 18 bias types with definitions and examples

ID	Bias Type	Description
1	Political	Favors or criticizes a specific political viewpoint. *Example:* "The radical left continues to sabotage the economy."
2	Gender	Reinforces stereotypes or prejudices based on gender. *Example:* "The female engineer surprisingly solved the problem."
3	Cultural / Ethnicity	Unfairly portrays or generalizes ethnic or cultural groups. *Example:* "Immigrants are taking away local jobs."
4	Age	Unfairly stereotypes or discriminates based on age. *Example:* "Older employees rarely adapt to new technology."
5	Religion	Unfairly stereotypes or discriminates based on religion. *Example:* "Muslim neighborhoods are often hotspots of radicalism."
6	Disability	Portrays individuals with disabilities or mental health conditions in a negative, stereotypical, or dehumanizing way. Often includes outdated, offensive language or implies that disability or mental illness is shameful, dangerous, or abnormal. *Example:* "This facility is for retarded individuals."
7	Statement Bias (labelling and word choice)	Also called presentation bias, refers to how articles choose to inform about certain entities/concepts through loaded language or presenting one side as the only side. Labelling uses specific words to convey particular opinions. *Example:* Words like "gender-affirming care" vs. "sex reassignment procedure" or "racial justice protest" vs. "riot" reveal different perspectives on the same events.
8	Unsubstantiated or Illogical Claims	Occurs when journalists make claims without supporting evidence or use flawed reasoning to reach unjustified conclusions. Includes both unsubstantiated claims and logical fallacies. *Examples:* "The senator's absence clearly shows he doesn't care about the crisis" (no source + unjustified inference); "This political change caused an increase in crime" (false cause fallacy).
9	Slant (Bias by Omission)	Highlights or plays up one particular angle while ignoring other perspectives. Through cherry-picking information, slant prevents readers from getting the full story and narrows understanding scope.

Continued on next page

Table 7 – continued from previous page

ID	Bias Type	Description
10	Source Selection Bias	The tendency to choose sources that support the story rather than sources that provide accurate accounts. *Example:* Covering an environmental disaster by only interviewing company representatives without giving voice to affected community members or independent experts.
11	Omission of Source Attribution	Occurs when journalists don't back up claims with sources, or sources are diffuse or unspecific. Sometimes intentional to protect source anonymity. *Examples:* Phrases like "according to a source," "critics say," or "experts believe."
12	Spin	Occurs when journalists try to create a "memorable story" using loaded or emotional language, exaggeration, or selective fact presentation to make content more interesting. Includes "clickbait" headlines and drama-focused stories.
13	Sensationalism	Information is exaggerated to create emotional reactions, targeting and provoking readers' emotions. Often involves selective information that supports certain views while omitting contradictory information. *Example:* "Bloodbath at the debate stage last night!"
14	Negativity Bias	Emphasizes bad or negative news, or frames events negatively. Follows the media adage "If it bleeds, it leads." *Example:* "The country is collapsing under the weight of failed leadership."
15	Subjective Adjectives	Uses qualifying adjectives that characterize or attribute specific properties to nouns, suggesting how readers should interpret issues rather than presenting facts objectively. *Example:* "The disturbing trend in education continues" or "The politician made a serious allegation."
16	Ad Hominem / Mud-slinging	Makes unfair or insulting accusations to damage someone's reputation, or attacks a person's character instead of addressing their arguments or ideas. *Example:* "He's a clown with no experience or credibility."
17	Mind Reading	Assumes knowledge of what another person thinks, interpreting internal thoughts or emotions of individuals who haven't explicitly expressed such thoughts or feelings. *Example:* "She clearly intended to undermine the election."

Continued on next page

Table 7 – continued from previous page

ID	Bias Type	Description
18	Opinion-as-Fact	Uses subjective language or statements under the guise of objective reporting. Presents subjective statements as factual information in supposedly objective news pieces. *Example:* "This policy is proof that the government doesn't care about citizens."

This comprehensive taxonomy serves as the foundation for our multi-model consensus approach, enabling systematic identification and classification of bias across diverse news content. Each bias type is evaluated at four granularity levels (None, Low, Moderate, High) to provide nuanced analysis of bias intensity and manifestation patterns in news articles.