Explaining News Bias Detection: A Comparative SHAP Analysis of Transformer Model Decision Mechanisms

Anonymous ACL submission

Abstract

Automated bias detection in news text supports journalistic analysis and media accountability, but little is known about how these models arrive at decisions or why they fail. We present a comparative interpretability study of three transformer-based bias detection models—bias-detector, DA-RoBERTa-BABE-FT, and a BABE-fine-tuned DeBERTa-base spanning two encoder backbone families—using SHapley Additive exPlanations (SHAP), evaluated on BABE and BASIL, a stylistically distinct news-bias corpus. All three models attend to overwhelmingly similar lexical content (90–96% Emotional/Evaluative language across matched-pool comparisons on both corpora) and overlap heavily on shared errors (81% top-token Jaccard on shared false positives). On positives carrying both lexical and inferential bias cues, recall matches the prediction from treating the two cue families as independent detection signals to within 2.4 points, indicating that the shared mechanism comprises two largely independent sub-signals rather than a single lexical shortcut. The three models differ systematically in attribution magnitude rather than category: DA-RoBERTa-BABE-FT's attribution on correctly identified non-biased content is roughly half of bias-detector's on both corpora, and on shared false positives bias-detector assigns 33% higher magnitude than DA-RoBERTa-BABE-FT to the same words. DeBERTa-base sits at an intermediate calibration point between the two while attending to the same lexical content, extending this pattern across architecture families. Error overlap between the two RoBERTa-family models exceeds independence predictions by 7.2× for shared false positives and 1.4× for shared misses, further indicating same-mechanism detection with calibration-level differences. These findings reframe architectural and training choices in bias detection as primarily affecting calibration rather than the linguistic signals identified, with practical implications for deployment in journalistic contexts.

Keywords: news bias detection; model interpretability; SHAP analysis; transformer models; linguistic bias;

1 Introduction

Detecting bias in news text is crucial to maintaining journalistic standards and helping readers assess media credibility. Natural language processing models trained to identify biased language in news articles promise to support these goals, but understanding how they make decisions is essential for trust and effective deployment (Bolukbasi et al., 2016; Caliskan et al., 2017). As automated bias detection tools enter newsrooms and fact-checking workflows, transparency about their decision-making becomes critical (Blodgett et al., 2020).

Bias in news text is multi-faceted: selection bias in which stories or sources are covered, framing bias in how events are presented, lexical bias in evaluative word choice, and structural bias in placement and prominence (Hamborg et al., 2019). This work focuses on sentence-level bias signals—bias observable in word choice and within-sentence framing—as operationalized by our two evaluation corpora. BABE annotates sentences as biased when they contain loaded language or framing devices. BASIL further distinguishes lexical bias (loaded word choice, e.g., "lashed out") from informational bias (selection or framing within prose). Throughout this paper, "bias" refers to the union of these two sentence-level signals. We do not address bias arising from source selection, story prominence, or other extra-sentential editorial decisions, which require article- or corpus-level analysis.

Model interpretability helps validate bias detection systems and diagnose failures. While accuracy metrics show overall performance, interpretability reveals which linguistic signals drive predictions,

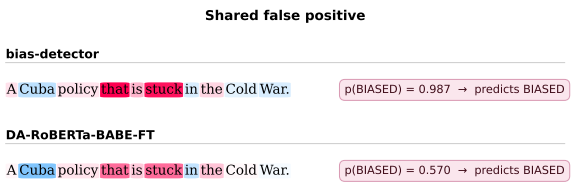

Figure 1: Shared false positive (gold label: non-biased). Both models attribute most strongly to the same tokens (*"that"*, *"stuck"*) but differ sharply in confidence ($p = 0.987$ vs 0.570). Same evidence, different calibration; quantified in Section 4.2.

enabling validation against journalistic principles and identification of problematic shortcuts (Lipton, 2017; Doshi-Velez and Kim, 2017). For journalism applications, interpretability is especially important: incorrect flagging can affect credibility, and stakeholders need to understand what constitutes bias from the model's perspective (Rudin, 2019).

We apply SHAP (SHapley Additive exPlanations) to compare how bias-detector (Ghosh et al., 2025), DA-RoBERTa-BABE-FT (Krieger et al., 2022), and a BABE-fine-tuned DeBERTa-base (He et al., 2021) interpret news text on BABE (Spinde et al., 2021) and on BASIL (Fan et al., 2019), a stylistically distinct news-bias corpus that differs from BABE in both distribution and annotation philosophy. By analyzing word-level attributions across all prediction categories on both corpora, we investigate: (1) how the models identify bias in news, (2) what linguistic patterns distinguish correct predictions from errors, and (3) whether architectural differences produce different detection mechanisms or different calibration.

Our analysis reveals: (1) a cross-corpus reversal in attribution strength for bias-detector, with false positives exceeding true positives in SHAP magnitude, indicating reliance on strong but misleading cues; (2) under matched-pool feature selection, all three models attend to overwhelmingly similar lexical content on both corpora and the two RoBERTa-family models overlap heavily on shared errors; (3) the three models differ systematically in attribution magnitude rather than category, with DA-RoBERTa-BABE-FT more conservative than bias-detector across categories (63% fewer false positives on BABE; AUPRC dominance across most BASIL strata) and DeBERTa-base at a statistically distinct intermediate calibration level, jointly supporting a same-mechanism-different-calibration interpretation; and (4) the shared mechanism comprises two largely independent detection signals—recall on positives carrying both lexical and inferential cues matches the prediction from treating them as independent to within 2.4 points.

This work contributes: (1) a cross-corpus, three-model comparative interpretability framework using SHAP for bias detection in news; (2) a methodological correction for vocabulary-pool-size confound in top-K feature selection; (3) identification of cross-corpus misalignment between attribution magnitude and prediction correctness in bias-detector; and (4) evidence across three encoder architectures that architectural differences in this model class primarily produce calibration differences rather than different detection mechanisms.[1]

2 Related Work

2.1 Transformer Models for NLP

Transformer architectures (Vaswani et al., 2017) and their derivatives such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021) form the backbone of modern bias detection systems, including the models analyzed in this work.

2.2 Bias Detection in News

Automated bias detection in news text has gained attention with the rise of computational journalism. Spinde et al. (2021) introduced BABE (Bias Analysis by Experts), a sentence-level dataset for bias detection in news articles using distant supervision from expert annotations. Ghosh et al. (2025) developed the bias-detector model, which we analyze here, and Krieger et al. (2022) proposed DA-RoBERTa-BABE-FT, a domain-adaptive pretraining approach for language bias detection in news.

2.3 Interpretability Methods in NLP

Interpretability methods for NLP span model-agnostic approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), and model-specific approaches based on attention or internal representations. Whether attention weights provide faithful explanations is contested (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019), and benchmarks such as ERASER (DeYoung et al., 2020) have been developed for evaluating rationalized models. We

[1] Code: https://github.com/NickDerWerner/SHAP-Analysis-Bias-Models

use SHAP for its theoretical grounding and model-agnostic applicability across architectures.

2.4 SHAP Applications in NLP

SHAP has been widely applied to explain NLP models due to its theoretical grounding and locally accurate attributions. Mosca et al. (2022) provide a comprehensive review of SHAP-based explanation methods for NLP interpretability across tasks.

2.5 Positioning

Prior interpretability work on bias detection has largely examined individual models in isolation, applying attribution methods to characterize a single system's failure modes or feature reliance. To our knowledge, no prior work has performed a comparative SHAP-based analysis across multiple deployed news-bias detectors, nor tested whether such interpretability findings generalize across stylistically distinct corpora. We address this gap by comparing three encoder-based bias detectors spanning RoBERTa and DeBERTa backbone families on BABE and BASIL, identifying a vocabulary-pool-size confound that affects prior top-K feature comparisons and characterizing whether observed differences reflect detection mechanism or calibration.

3 Methodology

3.1 Dataset and Models

We evaluate three bias detection models on the BABE (Bias Analysis by Experts) dataset[2] (Spinde et al., 2021), a sentence-level dataset for bias detection in news articles. BABE contains 4,121 expert-annotated sentences drawn from 15 U.S. news outlets, labeled at the sentence level as biased or non-biased. We evaluate on the 1,000-sentence test split (55.9% positive prevalence).

To assess cross-corpus generalization, we additionally evaluate all three models on BASIL (Fan et al., 2019), a corpus of 300 articles covering 100 news events with 7,984 sentences and a positive prevalence of 20%. BASIL provides span-level annotations categorized as lexical bias (loaded language), informational bias (framing or selection), or both, with each span further marked as appearing within a quotation or in unquoted prose. For sentence-level binary classification consistent with BABE, we treat any sentence containing at least one annotated span as biased and all others as non-biased; quoted spans are retained since their inclusion or framing can itself reflect editorial bias. The two datasets differ in annotation philosophy — BASIL annotates spans within potentially neutral surrounding text while BABE annotates whole sentences — and the cross-corpus evaluation accordingly reflects both distribution shift and this annotation difference.

Models: We analyze three transformer-based bias detection models: (1) bias-detector[3] (Ghosh et al., 2025), a model specifically designed for detecting bias in news text; (2) DA-RoBERTa-BABE-FT[4], a domain-adaptive RoBERTa model (Liu et al., 2019) fine-tuned on the BABE dataset (Krieger et al., 2022); and (3) DeBERTa-base (He et al., 2021), which we fine-tuned on BABE in this work (macro-F1 0.835, AUPRC 0.948) to test cross-architecture generalization. All three are transformer-based encoders (Vaswani et al., 2017; Devlin et al., 2019). We use the test split of BABE and the full BASIL corpus for evaluation.

Model selection rationale: bias-detector and DA-RoBERTa-BABE-FT are the established encoder-based state-of-the-art for sentence-level news bias detection, with public weights and journalism-adjacent deployment. We add a BABE-fine-tuned DeBERTa-base as a third encoder from a different backbone family to test whether our findings generalize beyond the RoBERTa lineage. Applying SHAP to decoder LLMs raises distinct methodological challenges (prediction-extraction strategy, prompt sensitivity, computational cost) that we leave to future work.

Preprocessing: For each model, we use the corresponding tokenizer from the HuggingFace Transformers library. Input sequences are tokenized with a maximum length of 256 tokens, following standard practices for sentence classification tasks. Labels are normalized to binary format (0 for non-biased, 1 for biased) to ensure consistent analysis across models.

3.2 SHAP-Based Interpretability Framework

We use SHAP (Lundberg and Lee, 2017), which assigns each feature an additive attribution based on Shapley values (Shapley, 1953). We apply the

[2] We use the `mediabiasgroup/BABE` release on HuggingFace (https://huggingface.co/datasets/mediabiasgroup/BABE), which contains 4,121 sentences and differs from the SG1/SG2 splits reported in Spinde et al. (2021).

[3] https://huggingface.co/himel7/bias-detector

[4] https://huggingface.co/mediabiasgroup/DA-RoBERTa-BABE-FT

SHAP library's TextExplainer with a Text masker, which handles tokenization and masking for transformer models and computes token-level attributions against a background dataset of test-set sentences. For each model, we construct a prediction function returning softmax probabilities and extract attributions for the positive (biased) class.

3.3 Word-Level Attribution Aggregation

Transformer-based models operate on subword units, while bias analysis is more interpretable at the word level. To bridge this gap, we aggregate token-level SHAP attributions into word-level explanations using a structured aggregation pipeline, illustrated in Figure 2.

Figure 2: Word-level SHAP aggregation pipeline. Token-level SHAP values produced by a transformer model are grouped into words based on tokenizer boundaries and text alignment, then aggregated to produce interpretable word-level attributions.

Given an input sentence, we first compute token-level SHAP values using the model-specific tokenizer and SHAP TextExplainer. Subword tokens are then grouped into words based on tokenizer boundary markers and original text alignment. For each word, we aggregate the SHAP values of all constituent tokens by summation, preserving the additive property of Shapley values. This produces a single attribution score per word, reflecting its overall contribution to the model's predicted probability of the biased class.

To ensure linguistic consistency, we apply normalization steps that correct tokenization artifacts and align word boundaries with the original text. These steps prevent spurious attributions arising from subword fragmentation or punctuation handling. The resulting word-level attributions enable consistent comparison across instances and models while retaining faithfulness to the underlying token-level explanations.

3.4 Stratified Sampling Strategy

To ensure comprehensive analysis across different prediction types, we employ stratified sampling to select instances for SHAP analysis. For each model, we first evaluate all test instances to obtain predictions and ground truth labels. We then categorize instances into: (1) True Positives (TP): correctly predicted as biased, (2) False Positives (FP): incorrectly predicted as biased, (3) True Negatives (TN): correctly predicted as non-biased, and (4) False Negatives (FN): incorrectly predicted as non-biased.

Sampling Procedure: We sample up to 100 instances from each category (TP, FP, TN, FN) per model to ensure balanced representation. For bias-detector on BABE, this yields 337 instances (100 TP, 37 FP, 100 TN, 100 FN); for DA-RoBERTa-BABE-FT, 312 instances (100 TP, 12 FP, 100 TN, 100 FN). FP counts reflect all available test-set false positives for each model. FN sampling was not part of the original BABE analysis design, which focused on predicted-positive and predicted-negative-correct categories; we extend it here to enable cross-corpus comparison of attribution on missed positives (Section 4.2). For BASIL, we apply the same four-category stratification with 100 instances per cell. Sampling is performed with a fixed random seed to ensure reproducibility.

3.5 Analysis Framework

For each sampled instance we compute and store word-level SHAP attributions alongside gold and predicted labels and probabilities, aggregated into global word-importance metrics (mean |SHAP|, occurrence count, mean signed SHAP) and per-category statistics (magnitude distributions, top words, positive vs. negative SHAP balance) for TP, FP, TN, and FN.

Matched-Pool Feature Selection: For analyses comparing word category composition between models or prediction partitions (Section 4.4), we use threshold-based feature selection: words whose pooled mean |SHAP| exceeds the 75th percentile of the comparison set are retained. This controls for attribution-vocabulary-size differences that would otherwise confound top-K comparisons; sensitivity analysis is in Appendix A.

3.6 Statistical Significance Testing

We assess paired-prediction differences between models using McNemar's test (Dietterich, 1998) with continuity correction, and distributional differences in per-instance SHAP magnitudes using two-sided Mann-Whitney U tests with rank-biserial effect sizes.

3.7 Evaluation Metrics

We report standard classification metrics: precision, recall, F1 score, false positive rate (FPR), and area under the precision-recall curve (AUPRC). Unless otherwise stated, classification decisions are made at a probability threshold of 0.5. Additionally, we analyze SHAP-specific metrics: (1) mean absolute SHAP value per instance category (TP, FP, TN, FN), (2) word-level attribution distributions, and (3) word category composition under matched-pool feature selection. These metrics provide both performance and interpretability perspectives on model behavior.

4 Evaluation Results

4.1 Overall Model Performance

Table 1 summarizes performance metrics for the three models on BABE and BASIL. On BABE, bias-detector achieves higher F1 while DA-RoBERTa-BABE-FT shows lower FPR and higher precision; DeBERTa-base sits between them on F1 (0.836) and matches DA-RoBERTa's low FPR (0.061). On BASIL, all three show substantially lower F1 at threshold 0.5, partly reflecting annotation-philosophy differences (BABE annotates whole sentences; BASIL converts span annotations to a sentence-level binary task) alongside distribution shift. Threshold-independent AUPRC indicates all three retain discriminative ability (lift 1.79–2.03× over baseline), with DeBERTa's BASIL operating point at threshold 0.5 less favorable than its ranking ability would suggest—consistent with a calibration rather than mechanism difference. The performance gap motivates the SHAP-level analysis that follows. If the models truly differ in detection mechanism, we expect divergent attribution patterns; if they differ only in calibration, attribution should look similar in shape but differ systematically in magnitude.

	BABE			BASIL		
Metric	**bd**	**dr**	**deb**	**bd**	**dr**	**deb**
Precision	0.924	0.970	0.940	0.463	0.488	0.471
Recall	0.801	0.705	0.753	0.316	0.272	0.207
F1	0.858	0.817	0.836	0.376	0.349	0.288
FPR	0.156	0.057	0.061	0.093	0.073	0.059
AUPRC	0.953	0.957	0.948	0.391	0.412	0.364

Table 1: Model Performance on BABE and BASIL Test Sets. deb = DeBERTa-base fine-tuned on BABE in this work. F1 is binary F1 at threshold 0.5; AUPRC is threshold-independent.

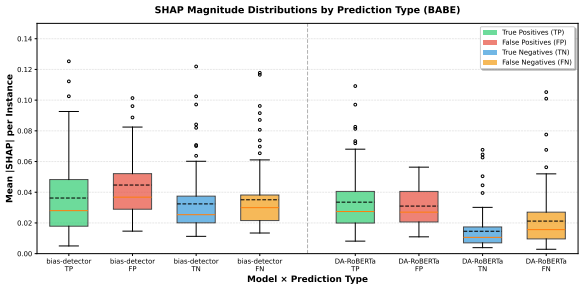

Figure 3: Per-instance mean |SHAP| by prediction category on BABE. Dashed line: mean; orange line: median. BABE FP samples reflect all available test-set false positives (bd: n=37, dr: n=12); other categories n=100. DeBERTa-base medians are reported in Table 2 and discussed in the text.

4.2 SHAP Magnitude Patterns

Word-level SHAP attribution magnitudes reveal differences in how models interpret linguistic signals. Fig. 3 presents per-instance mean |SHAP| by prediction category on BABE; Table 2 summarizes the medians across both corpora.

On BABE, bias-detector's false positives exhibit higher attribution magnitude than true positives—a reversal in attribution strength between correct and incorrect positive predictions, driven by more numerous moderate-magnitude signals rather than stronger peaks (Appendix B). DA-RoBERTa-BABE-FT shows comparable TP and FP magnitudes. The two models also differ on true negatives: DA-RoBERTa-BABE-FT's median TN attribution is roughly 42% of bias-detector's. False-negative attribution sits at TP level for bias-detector, consistent with its broader over-flagging character; DA-RoBERTa-BABE-FT FN sits clearly between TN and TP, suggesting sub-threshold detection on missed sentences.

To assess whether these patterns generalize, we evaluated SHAP attribution on BASIL, a stylistically and topically distinct news-bias corpus (Fig. 4).

The bias-detector reversal persists on BASIL, as does the TN attribution gap (DA-RoBERTa-BABE-FT TN ratio 0.51 vs. 0.42 on BABE). DA-RoBERTa-BABE-FT's BABE-specific TP ≈ FP pattern, however, does not generalize: on BASIL it also shows FP > TP. The original "alignment" interpretation does not reflect a robust calibration property of the model and should be revised; the cross-corpus calibration signature is instead the TN gap.

False-negative attribution generalizes asymmet-

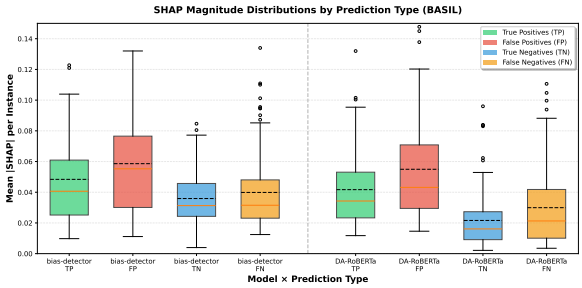

Figure 4: Per-instance mean |SHAP| by prediction category on BASIL. Dashed line: mean; orange line: median. All categories sampled at n=100. DeBERTa-base medians are reported in Table 2 and discussed in the text.

	BABE			BASIL		
	bd	dr	deb	bd	dr	deb
TP	0.028	0.027	0.028	0.041	0.034	0.037
FP	0.037	0.027	0.034	0.055	0.043	0.058
TN	0.025	0.011	**0.016**	0.031	0.016	**0.023**
FN	0.030	0.016	0.023	0.032	0.021	0.025

Table 2: Median per-instance mean |SHAP| by prediction category across three models. DeBERTa-base sits between bd and dr on the TN calibration axis on both corpora (BABE: 0.025 / 0.016 / 0.011; BASIL: 0.031 / 0.023 / 0.016), forming a smooth three-model calibration spectrum.

rically. DA-RoBERTa-BABE-FT maintains FN above the TN baseline on both corpora—consistent sub-threshold detection across distributions. Bias-detector shows FN > TN on BABE but FN ≈ TN on BASIL: its missed-positive attribution is corpus-dependent in a way DA-RoBERTa-BABE-FT's is not.

Figure 1 illustrates this pattern: on a shared false positive (*"A Cuba policy that is stuck in the Cold War"*), both models attribute most strongly to the same tokens (*"that"*, *"stuck"*), yet bias-detector predicts BIASED at $p = 0.987$ while DA-RoBERTa-BABE-FT predicts BIASED at $p = 0.570$—barely past threshold. Same evidence, different confidence.

Paired analyses on sentences where both models commit the same error reinforce this interpretation. On 100 shared false-positives in BASIL, the two models share 81% of their top-20 attributed tokens, yet bias-detector assigns 33% higher magnitude than DA-RoBERTa-BABE-FT to those shared words. On 100 shared false-negatives, Jaccard is 0.72 and bias-detector's magnitude is double DA-RoBERTa-BABE-FT's. The two models attend to substantially the same linguistic signals but attribute systematically different magnitudes—same-mechanism detection with different calibration. The TN calibration signature extends to a smooth three-model spectrum when DeBERTa-base is included: median per-instance mean |SHAP| on TNs is 0.025 (bd) > 0.016 (DeBERTa) > 0.011 (dr) on BABE, and 0.031 > 0.023 > 0.016 on BASIL—consistent ordering across both corpora. FN attribution shows the same three-model ordering on both corpora. Pairwise Mann-Whitney U tests confirm DeBERTa's TN distribution differs significantly from both bd and dr on both corpora ($p < 10^{-4}$ in all four cross-model TN comparisons; Section 4.5). Three architecturally distinct encoders, spanning RoBERTa and DeBERTa backbone families, thus occupy three calibration levels on the TN axis with comparable discriminative ability (Table 1). Under matched-pool feature selection (Section 4.4), they attend to the same lexical categories. The calibration-vs-mechanism interpretation therefore extends across architecture families.

4.3 Global Bias Indicators

We aggregated word-level SHAP attributions across all instances to identify global bias indicators. Fig. 5 presents the top 10 words with highest mean absolute SHAP values for each model.

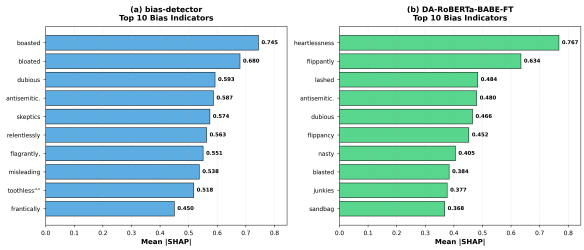

Figure 5: Top 10 bias indicators (highest mean |SHAP|) for bd and dr on BABE; DeBERTa-base indicators discussed in Section 4.3.

For bias-detector, top indicators include emotional and evaluative language: "boasted", "bloated", "dubious", "antisemitic", and "skeptics". Many top words are negative descriptors ("bloated", "dubious") or explicit bias terms ("antisemitic"), suggesting the model relies heavily on loaded language and explicit bias markers. For DA-RoBERTa-BABE-FT, top indicators emphasize dismissive and moral language: "heartlessness", "flippantly", "lashed", "antisemitic", and "dubious". This model shows a stronger focus on dismissive language patterns ("flippantly", "heart-

Model	Pred.	n	Em/Ev	Expl.	Fr/Rep	Fu/Tm
bd	TP	399	96.0	0.8	1.0	2.3
	FP	231	90.5	0.4	2.2	6.9
dr	TP	330	96.7	0.6	2.4	0.3
	FP	39	87.2	0.0	7.7	5.1
deb	TP	371	97.6	0.0	1.3	1.1
	FP	134	91.0	0.7	0.7	7.5

Table 3: Matched-pool TP-vs-FP category composition (percentages of threshold-selected words) on BABE. All three models' TPs are dominated by Emotional/Evaluative content (96.0–97.6%); FPs show modest, model-symmetric shifts toward Function/Temporal words.

	Cell	n	Em/Ev	Expl.	Fr/Rep	Fu/Tm
BABE	Shared (bd)	5	92.9	0.0	0.0	7.1
	Shared (dr)	5	89.5	0.0	0.0	10.5
	bd-only	32	89.8	0.6	2.8	6.8
	dr-only	7	75.0	0.0	15.0	10.0
BASIL	Shared (bd)	100	94.4	0.3	2.5	2.8
	Shared (dr)	100	96.2	0.4	1.7	1.7
	bd-only	100	94.4	0.0	1.9	3.7
	dr-only	100	91.3	0.4	2.6	5.7

Table 4: Three-way FP partition under matched-pool feature selection. Values are percentages of threshold-selected words per category. Emotional/Evaluative dominates every cell on both corpora.

lessness") and emotional descriptors compared to bias-detector. Some overlap exists ("dubious", "antisemitic", "lashed"), suggesting consensus on certain linguistic bias markers.

4.4 Word-Level Patterns

Table 3 reports matched-pool category composition of true positives and false positives for all three models on BABE. Evaluative and emotionally charged language dominates the highest-attribution features under matched-pool selection (threshold $\geq$ p75 of pooled mean |SHAP|) for all three models. Bias-detector's TP and FP compositions are both dominated by Emotional/Evaluative language (96.0% vs. 90.5%), with a modest rise in Function/Temporal words on FPs (2.3% vs. 6.9%). The model relies on similar lexical cues regardless of whether its prediction is correct, supporting the over-flagging interpretation from Section 4.2.

DA-RoBERTa-BABE-FT shows the same direction and roughly equal magnitude of TP-to-FP shift toward Function/Temporal words (0.3% vs. 5.1%, compared to bias-detector's +4.7 pt rise). Under matched feature pools, Emotional/Evaluative still dominates DA-RoBERTa-BABE-FT FPs (87.2%). The earlier finding that DA-RoBERTa-BABE-FT FPs shift substantially toward function and temporal words reflected a vocabulary-pool-size confound (Appendix A): under fixed top-K selection, DA-RoBERTa-BABE-FT's smaller FP vocabulary (249 unique types vs. bias-detector's 660) was exhausted at lower attribution magnitudes, inflating the apparent function/temporal proportion.

DeBERTa-base shows the same qualitative pattern: TPs are 97.6% Emotional/Evaluative and FPs 91.0%, with the largest TP-to-FP shift again going to Function/Temporal words (+6.4 pt). The TP cells across all three models are essentially identical in composition (94.4–97.6% Emotional/Evaluative), indicating that what the models attend to when correctly flagging biased content does not vary meaningfully with architecture. The three-model matched-pool comparison thus confirms that the calibration spectrum documented in Section 4.2 operates over the same lexical surface across all three encoders.

To examine whether the two RoBERTa-family models attend to similar content when they fail in the same way, we partition FPs into three cells: shared (both models commit FPs on the same sentence), bd-only, and dr-only.

Emotional/Evaluative dominates all partitions on both corpora (75.0–92.9% on BABE; 91.3–96.2% on BASIL); BASIL cells (n=100 each) confirm the dominance pattern robustly. The bd-only partition is largest on both corpora, consistent with bias-detector's over-flagging generating the bulk of its FPs on sentences DA-RoBERTa-BABE-FT correctly identifies as non-biased. Across both corpora, the two models attend to the same lexical categories whether correct or incorrect; they differ primarily in how readily they exceed the decision threshold—a calibration difference rather than a difference in reasoning over the linguistic surface.

4.5 Statistical Significance Testing

We applied McNemar's test on the BABE test set to assess whether the observed difference in error patterns between the two models is statistically significant. The test yielded $\chi^2 = 5.723$ with $p = 0.0167$, indicating the difference in false positive rates (15.6% vs. 5.7%) is unlikely to be due to chance.

We additionally test whether the SHAP magnitude differences underlying the Section 4.2 reversal claim are statistically significant. Using two-

sided Mann-Whitney U tests on per-instance mean |SHAP| distributions, bias-detector's false positives exhibit significantly higher magnitude than its true positives on both corpora (BABE: $U = 1335$, $p = 0.013$; BASIL: $U = 4079$, $p = 0.025$; rank-biserial $|r| = 0.28$ and 0.18 respectively). DA-RoBERTa-BABE-FT shows no significant TP-vs-FP difference on BABE ($U = 627$, $p = 0.80$; the BABE FP set has $n = 12$, limiting power) but a significant FP-greater-than-TP pattern on BASIL ($U = 3897$, $p = 0.007$, $|r| = 0.22$). The cross-corpus TN attribution gap that anchors the calibration interpretation is also significant with large effect sizes: bias-detector's TN magnitudes exceed DA-RoBERTa-BABE-FT's on both BABE ($U = 8627$, $p < 10^{-18}$, $|r| = 0.73$) and BASIL ($U = 7719$, $p < 10^{-10}$, $|r| = 0.54$). DeBERTa-base's TN distribution also differs significantly from both bd and dr on both corpora ($p < 10^{-4}$ in all four cross-model TN comparisons), placing it at a statistically distinct intermediate calibration level rather than collapsing onto either of the RoBERTa-family models. These direct tests confirm that the magnitude patterns visible in Figures 3 and 4 reflect statistically significant distributional differences, not artifacts of summary medians.

On BASIL's 6,361 negative examples, the two RoBERTa-family models share 314 false positives at threshold 0.5—7.3× more than if their errors were independent (53.0% shared rate among bias-detector's FPs vs. 7.3% predicted under independence). Including DeBERTa-base, 197 FPs are shared by all three models, an 8.2× excess over the three-way independence prediction. Among the 405 lex-only positives, the two RoBERTa-family models share 221 misses (54.6% vs. 38.6% under independence; 1.4× excess). All three models tend to err on the same content.

At threshold 0.3 on BASIL's both-types stratum (n=44 positives carrying both lexical and inferential bias cues), recall is 0.64 / 0.66 for bias-detector and DA-RoBERTa-BABE-FT. Treating the two cue pathways as independent and using the per-stratum lex-only and inf-only recalls (0.47 / 0.34 for bias-detector, 0.49 / 0.37 for DA-RoBERTa-BABE-FT) predicts 0.65 / 0.68 via $P_{\text{both}} = 1 - (1 - P_{\text{lex}})(1 - P_{\text{inf}})$. Observed matches predicted within 2.4 points for both models. For DeBERTa-base on the same stratum, observed recall is 0.432 vs. 0.480 predicted under independence—a modest sub-additive gap of -0.048 points, larger than the bd/dr deviations but in the same direction. The three-model picture is consistent with two approximately-independent detection pathways, with mild and model-dependent interaction effects rather than a single shared lexical shortcut.

5 Conclusion

We conducted a comparative interpretability analysis of three transformer-based bias detectors using SHAP, evaluated on BABE and on BASIL for cross-corpus generalization. Bias-detector assigns stronger internal evidence to false positives than to correct detections, a reversal that persists across both corpora and contributes to systematic over-flagging of neutral journalistic content—a concern for deployment in editorial settings.

DA-RoBERTa-BABE-FT and DeBERTa-base produce fewer false positives than bias-detector on BABE and show lower attribution magnitude across most prediction categories on both corpora, with DeBERTa-base occupying a statistically distinct intermediate position on the TN calibration axis. Under matched-pool feature selection, all three models attend to overwhelmingly similar lexical content (90–98% Emotional/Evaluative) and the two RoBERTa-family models overlap heavily on shared errors. The three models thus differ in how confidently they convert linguistic signals into predictions—a calibration difference spanning two encoder backbone families—rather than in the linguistic signals themselves.

Our results highlight the importance of interpretability-aware evaluation for journalistic applications. The calibration framing suggests that improving deployed bias detectors may benefit more from calibration-aware training than from architectural changes aimed at altering attribution patterns. Future work should examine whether this distinction generalizes to decoder LLMs, which require methodological adaptations for SHAP-style attribution on generative models.

6 Limitations

This study has several limitations. First, our analysis is restricted to two English-language news-bias corpora (BABE and BASIL). Both reflect Western journalistic norms and specific annotation guidelines—BABE annotates whole sentences while BASIL annotates spans within sentences—and the cross-corpus performance gap conflates distribution shift with these annotation-philosophy dif-

ferences. Generalization to other languages, journalistic traditions, or cultural definitions of bias requires further evaluation. We also note that one originally reported pattern—DA-RoBERTa-BABE-FT's near-equal TP and FP attribution magnitudes on BABE—did not generalize to BASIL and is accordingly revised in our cross-corpus analysis (Section 4.2).

Second, our analysis focuses on three encoder-based bias detectors spanning RoBERTa and DeBERTa backbone families. Whether the calibration-vs-mechanism distinction extends to decoder LLMs—which require methodological adaptations for SHAP-style attribution—is an open question for future work.

Third, our interpretability analysis relies on SHAP as a post-hoc explanation method. SHAP approximates feature contributions under independence assumptions that may not fully capture complex interactions in transformer architectures; word-level attributions should be interpreted as explanatory signals rather than definitive causal explanations. Our matched-pool feature selection (p75 threshold) provides one specific control for vocabulary-size confound in cross-model word-category comparisons; alternative selection rules are explored in Appendix A.

Finally, while we compare attribution patterns across correct and incorrect predictions, we do not evaluate alternative explanation methods or formally assess explanation faithfulness beyond consistency with model behavior. Future work could extend this analysis by incorporating multiple interpretability techniques, calibration metrics, or human evaluation of explanations to further validate reliability and usability.

7 Ethical Considerations

This work analyzes bias detection models and their explanations, but does not propose automated systems as substitutes for human judgment. Bias in journalism is inherently contextual and normatively defined, and model predictions along with their explanations may reflect annotation guidelines, cultural assumptions, or training data biases present in the BABE and BASIL datasets. As such, outputs from these models should not be treated as objective or authoritative assessments of bias.

Our analysis highlights the risk of false positives, which may incorrectly label neutral reporting as biased and potentially undermine journalistic credibility if deployed without oversight. We therefore emphasize that bias detection models should be used as assistive tools within human-in-the-loop workflows, where explanations can support critical inspection rather than automated decision-making.

Finally, while interpretability techniques such as SHAP improve transparency, they do not guarantee correctness or fairness. Explanations should be understood as diagnostic aids for model behavior, not as justifications for deploying models in high-stakes or adversarial settings without careful evaluation and governance.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. *Preprint*, arXiv:1911.03429.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of

factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Himel Ghosh, Ahmed Mosharafa, and Georg Groh. 2025. To bias or not to bias: Detecting bias in news with bias-detector. *Preprint*, arXiv:2505.13010.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Preprint*, arXiv:1902.10186.

Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22. ACM.

Zachary C. Lipton. 2017. The mythos of model interpretability. *Preprint*, arXiv:1606.03490.

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. *Preprint*, arXiv:1907.11692.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and George Louis Groh. 2022. Shap-based explanation methods: A review for nlp interpretability. In *International Conference on Computational Linguistics*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Preprint*, arXiv:1811.10154.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? *Preprint*, arXiv:1906.03731.

Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Timo Spinde, Hendrik Westermann, and Georg Rehm. 2021. Neural media bias detection using distant supervision with babe. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3518–3530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Preprint*, arXiv:1908.04626.

A Vocabulary-Pool-Size Confound in Top-K Feature Selection

The original analysis presented in Section 4.4 identified an apparent shift toward function and temporal words in DA-RoBERTa-BABE-FT false positives (13.0% Function/Temporal at K=100 vs. bias-detector's 4.0%). This appendix documents the sensitivity analysis showing that the gap collapses or inverts under controls for vocabulary pool size, indicating the original finding was artifactual.

Vocabulary statistics. The two models produce FP sets of markedly different vocabulary sizes on BABE (Table 5). DA-RoBERTa-BABE-FT's FP vocabulary is roughly one-third of bias-detector's. Under fixed top-K selection, the smaller pool is exhausted at lower attribution magnitudes, inflating the apparent proportion of marginal categories in the selected words.

	bias-detector FP	DA-RoBERTa FP
Unique word types	660	249
n samples	37	12
Vocab ratio (dr/bd)	—	0.38

Table 5: FP vocabulary statistics on BABE.

Sensitivity sweep. Table 6 reports DA-RoBERTa-BABE-FT FP Function/Temporal proportion under varying selection rules. Three patterns emerge: (i) the 13.0% figure at K=100 does not generalize to other K values; (ii) under proportional-K selection (at fixed percentages of each model's vocabulary), DA-RoBERTa-BABE-FT's proportion is comparable to or below bias-detector's; (iii) under threshold-based selection (p75 of pooled mean |SHAP|), DA-RoBERTa-BABE-FT's proportion is 5.1% vs. bias-detector's 6.9%.

Selection rule	F/T (%)	n_{selected}
K=25	4.0	25
K=50	8.0	50
K=100 (original)	**13.0**	100
K=200	15.0	200
K=500	13.7	249 (exhausted)
Proportional-K, top 5%	0.0	12
Proportional-K, top 10%	4.0	25
Proportional-K, top 20%	8.0	50
Proportional-K, top 30%	10.7	75
Threshold $\geq$ p75 pooled	**5.1**	39

Table 6: Function/Temporal proportion in DA-RoBERTa-BABE-FT FPs under varying selection rules. For reference, bias-detector's proportion under threshold-based selection is 6.9%.

Corpus	Model	TP	FP	TN	FN
BABE	bd	0.172	0.180	0.103	0.133
	dr	0.181	0.161	0.044	0.085
BASIL	bd	0.220	0.242	0.106	0.135
	dr	0.194	0.231	0.058	0.094

Table 7: Median per-instance max |SHAP| by prediction category across BABE and BASIL.

Corpus	Model	FP/TP (mean)	FP/TP (max)
BABE	bd	1.31	1.04
	dr	0.99	0.89
BASIL	bd	1.34	1.10
	dr	1.26	1.19

Table 8: FP-vs-TP ratio of median |SHAP| at the per-instance mean and max levels. The max ratio is smaller than the mean ratio in all four model$\times$corpus cells, indicating that the FP-vs-TP magnitude gap reflects breadth of moderate-magnitude attribution rather than stronger peak signals.

Bootstrap analysis. A bootstrap analysis at K=100 over 1,000 resamples of DA-RoBERTa-BABE-FT's FP set yields a 95% confidence interval of [11.0%, 24.0%] for the Function/Temporal proportion (mean 17.5%). While this interval excludes the bias-detector baseline of 4.0% under matched fixed-K selection, the gap is driven entirely by the vocabulary-pool-size effect rather than by a genuine difference in attribution behavior.

Implication. The original finding that DA-RoBERTa-BABE-FT errors shift toward function and temporal words relative to bias-detector reflects a vocabulary-pool-size confound rather than a difference in linguistic categorization. The corrected analysis in Section 4.4 uses threshold-based feature selection to remove the confound and demonstrates that both models attend to similar content categories under matched pools, with comparable modest TP-to-FP shifts in both models.

B Max |SHAP| Analysis

Section 4.2 characterizes the attribution-magnitude reversal in bias-detector using mean |SHAP| per instance. Whether this reversal reflects a few stronger trigger words on false positives or more moderate signals spread across the sentence can be assessed by comparing the per-instance mean to the per-instance maximum.

Interpretively, when a false positive and a true positive both carry one strong-ish trigger word, the per-instance max attribution equalizes them. The mean is higher on FPs because the trigger word sits in a sentence where more surrounding words also carry moderate attribution. This is the calibration-not-mechanism reading made quantitative: bias-detector spreads moderate attribution more widely across over-flagged sentences without amplifying any single signal.

Two caveats. First, DA-RoBERTa-BABE-FT's BABE FP set contains only $n = 12$ instances, making the BABE max FP/TP ratio of 0.89 a high-variance estimate; the BASIL counterpart ($n = 100$, max ratio 1.19) provides the more robust cross-corpus reading. Second, BABE max ratios compress more aggressively than BASIL's (bias-detector: 1.31 mean $\rightarrow$ 1.04 max on BABE vs. 1.34 $\rightarrow$ 1.10 max on BASIL), consistent with BABE's shorter sentences and tighter SHAP magnitude distributions. The cross-corpus pattern is consistent in direction: max attribution always equalizes more than mean attribution.

The TN attribution gap that anchors the calibration interpretation in the main text persists in max attribution: DA-RoBERTa-BABE-FT's median max |SHAP| on TNs is 43% of bias-detector's on BABE (0.044 vs. 0.103) and 55% on BASIL (0.058 vs. 0.106), closely matching the mean ratios (42% and 51%, respectively). The cross-corpus calibration signature is robust across both peak and mean attribution metrics. False-negative attribution sits above the TN baseline for both models on both corpora in max as in mean, consistent with sub-threshold detection on missed positives.

C AUPRC by Bias-Type Stratum on BASIL

Table 9 reports area under the precision-recall curve (AUPRC) for both models on BASIL across the full evaluation set and five bias-type or quote-handling strata. Baseline equals positive-class prevalence within the stratum (equivalent to the AUPRC of a random classifier); lift equals AUPRC / baseline.

Both models exhibit above-baseline discrimination on every stratum. The lex-only and inf-only lifts (3.3–3.5× and 1.9–2.0×) support the bias-type independence interpretation in Section 4.5: each cue family is detectable on its own, and the two detection pathways are approximately statistically independent rather than reflecting a single shared lexical shortcut. The high lift on the both-types stratum (8.5–14.8×) reflects the stratum's small baseline prevalence (0.69%) rather than uniformly stronger performance; absolute AUPRC there is the lowest of all strata (0.058 / 0.101), consistent with the small positive sample size ($n_{\text{both-types positive}} = 44$). Quote-excluded and quote-only strata show similar lifts (2.1–2.8×), indicating that quote handling does not materially alter the cross-corpus discrimination picture.

D Token-Level SHAP Examples

True Positives. Refer to Fig. 6 below.

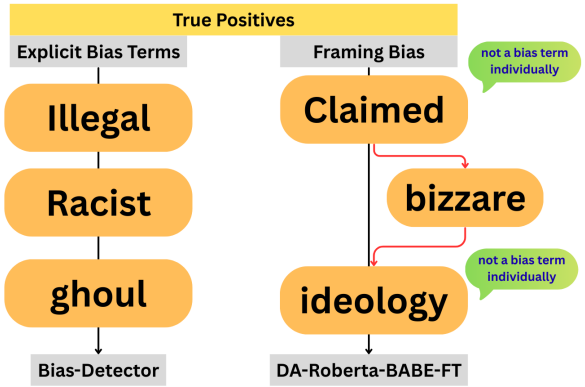

Figure 6: Examples of Biased Terms detected by both the models for the True Positives.

False Positives. **Bias-detector:** On its 37 false positives (mean confidence 0.80), recurring high-attribution tokens include *pandemic*, *illegal*, *not*, *Tuesday*, and *could*. The word *illegal* illustrates the TP/FP invariance discussed in Section 4.4: the model attributes to it both when correctly flagging biased content (mean |SHAP| = 0.349 across true positives) and when incorrectly flagging neutral content (mean |SHAP| = 0.140 across false positives). The same lexical cue drives both correct and incorrect predictions; surrounding context distinguishes the gold labels, but the model under-weights it. Several recurring FP tokens (*said*, *Tuesday*, *could*, *not*) carry negative SHAP, indicating they should reduce predicted bias probability, yet the model still predicts biased with high confidence—consistent with the mean-vs-max-SHAP analysis in Appendix B, where bias-detector's FP magnitude excess reflects breadth of moderate-magnitude positive attribution rather than stronger peaks.

DA-RoBERTa-BABE-FT: On its 12 false positives (mean confidence 0.67), top contributors include *gruesome*, *unbelievable*, and the *claims/claimed* stem. The stem appears in both true positives (*claimed*, *claiming*, *claim*; mean |SHAP| = 0.125) and false positives (mean |SHAP| = 0.206), with notably higher attribution on incorrect predictions—a parallel to bias-detector's *illegal* pattern, operating at lower overall confidence consistent with the calibration gap characterized in Section 4.2.

True Negatives. Refer to Fig. 7 below.

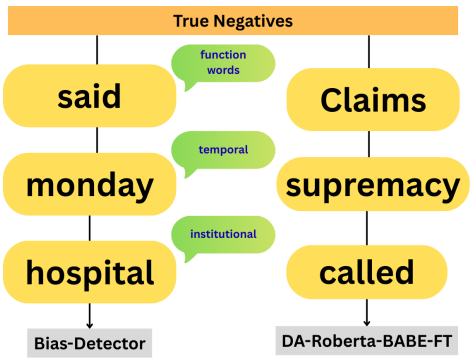

Figure 7: Examples of Biased Terms detected by both the models for the True Negatives.

E Word-Level Attribution Aggregation in Detail

Transformer models operate on subword tokens (e.g., BPE tokens in RoBERTa), but interpretability for bias detection benefits from word-level analysis. We aggregate token-level SHAP values to word-level attributions through a multi-step process.

Token-to-Word Mapping: We identify word boundaries using tokenization markers. RoBERTa-style tokenizers use special markers (e.g., Ġ or _)

Stratum	n	BD AUPRC	DR AUPRC	Baseline	BD lift	DR lift
Full	7,984	0.392	0.412	0.203	1.93×	2.03×
Lex-only	6,766	0.195	0.209	0.060	3.25×	3.50×
Inf-only	7,535	0.290	0.313	0.156	1.86×	2.01×
Both-types	6,405	0.058	0.101	0.007	8.49×	14.75×
Quote-excluded	7,297	0.275	0.274	0.128	2.14×	2.14×
Quote-only	7,048	0.226	0.270	0.098	2.32×	2.77×

Table 9: AUPRC by bias-type and quote-handling stratum on BASIL. Each stratum's n is its positive-class count plus all 6,361 negatives in the BASIL evaluation pool.

to indicate tokens that begin new words. We detect these markers and group consecutive subword tokens that belong to the same word. For tokens without explicit markers, we use whitespace information from the original text to determine word boundaries.

Word Boundary Detection: To handle cases where tokens may be incorrectly merged (e.g., "tuesdayquoting" or "nationalismfueled"), we implement heuristic word boundary detection. We split tokens when: (1) a lowercase letter is followed by an uppercase letter (e.g., "tuesdayQuoting" → ["tuesday", "Quoting"]), (2) common word endings (e.g., "day", "ing", "ed", "ly") are followed by lowercase letters that could start new words, or (3) punctuation or hyphens indicate word separation.

Attribution Aggregation: For each word, we sum the SHAP values of all tokens that comprise that word. If a word is split across multiple tokens, the word-level attribution is the sum of all constituent token attributions. This aggregation preserves the additive property of SHAP values while providing interpretable word-level explanations.

Normalization: We apply normalization to handle tokenization artifacts. Merged words (e.g., "dmnboasted" → "boasted") are split, and punctuation artifacts are stripped from word boundaries. This ensures that our word-level analysis reflects actual linguistic patterns rather than tokenization idiosyncrasies.