Uncertainty-Aware Proxy Attribute Reasoning for Reliable Media Bias Detection

Chin-Po Chen[1], **Jeng-Lin Li**[2], **Ming-Ching Chang**[1]

Abstract

Detecting media bias is increasingly important in today's rapidly evolving digital landscape. While large language models (LLM) show promise for bias detection, their reasoning is often unreliable and difficult to interpret. To address this limitation, we propose an **uncertainty-aware proxy reasoning (UAPR)** method that integrates proxy attribute prediction with uncertainty estimation to assess and stratify reasoning quality. By explicitly modeling uncertainty, our method narrows down the intermediate reasoning space and identifies trustworthy bias indicators, improving transparency and interpretability for end users and downstream applications. We evaluate our method on the BASIL benchmark and compare it against strong LLM baselines and recent state-of-the-art approaches. Our experiments assess both decision accuracy and quality. Results show a 6.25% relative F1 improvement over competitive baselines, demonstrating that uncertainty-aware reasoning boosts performance and provides insight into decision making, thereby improving the reliability and interpretability of media bias detection.

1 Introduction

The presentation of information from different perspectives can be deliberately manipulated to influence both individual and collective thinking. When personal biases and preferences drive communication, they often lead to skewed or slanted language. Mass media, by its nature, provides fertile ground for such subjectivity and informational inconsistency to arise (Corner, 2013). To maintain content quality, media outlets rely on media bias assessments to identify and manage these distortions. Media bias has been categorized by how it manifests, including coverage bias, gatekeeping bias, and statement bias (Rodrigo-Ginés et al., 2024). Statement bias, also known as presentation bias, concerns how information is conveyed through the selective use of specific words or phrases. This form of bias is particularly common in contemporary news language and remains a significant challenge to detect and manage effectively.

Reliable decision-making involves two key dimensions: decision accuracy and decision quality (Vilela and Oluyemi, 2021). Media bias detection is a highly complex social decision-making problem, shaped by latent factors and subject to diverse interpretations (Morelli et al., 2022). This complexity often leads to controversy, as even human annotators struggle to maintain consistent reasoning when labeling bias. Consequently, prior studies on media bias detection have largely focused on prediction accuracy while neglecting decision quality. Although incorporating intermediate reasoning steps can improve accuracy, decision quality ultimately depends on effectively optimizing and converging across multiple reasoning pathways (Schutte et al., 2025). Therefore, integrating quality-aware intermediary reasoning into media bias detection is essential for enhancing both practical reliability and analytical depth. Figure 1 illustrates our bias decision-making framework, which employs proxy attributes to narrow the reasoning space and uses uncertainty estimation to reduce decision variability.

Automatic sentence-level media bias detection has attracted significant attention due to its potential to improve prediction accuracy (Lin et al., 2024; Lei and Huang, 2024; Wang et al., 2025). A line of NLP-based research focuses specifically on sentence-level bias, including aspects such as lexicon bias and information bias, as documented in expert-curated datasets (Fan et al., 2019). In this setting, sentences or claims are labeled as biased when they exhibit a preference toward specific entities, without relying on external knowledge beyond the given context (Van Den Berg and Markert, 2020). Large language models (LLMs) have recently emerged as a promising approach by lever-

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 40–63

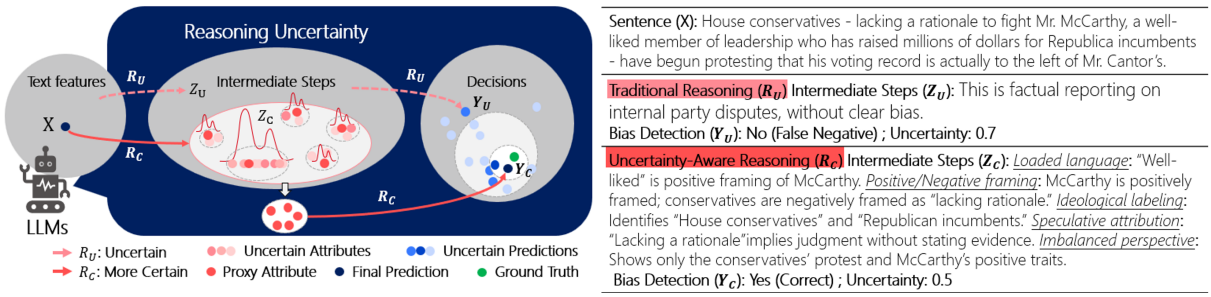


Figure 1: Proxy attributes (orange dots) and uncertainty-aware decisions help narrow the decision space and improve decision quality for media bias detection. Uncertainty estimation identifies reliable LLM intermediate steps (Z_C) that lead to more certain predictions (Y_C). See texts for explanation.

aging carefully designed prompts to expose bias tendencies in few-shot settings (Maab et al., 2024). These methods extend generative models to binary prediction tasks, opening new directions for media bias detection. However, hallucinations and inconsistent outputs from generative models continue to limit their reliability in real-world applications.

Quantifying uncertainty is a fundamental component of decision quality, yet it remains underexplored in media bias detection. This gap stems in part from the inherent complexity and nuance of media content. Effective bias detection often requires intermediate reasoning over subtle linguistic cues, selective presentation of facts, and varying editorial standards. Such complexity makes it difficult to model media bias directly from raw features, particularly when deeper reasoning is required. Uncertainty therefore serves as a valuable auxiliary dimension for revealing latent patterns of bias. By explicitly modeling uncertainty, systems can better capture subtle manifestations of bias and distinguish reliable indicators from ambiguous evidence. In this context, the objective extends beyond accurate feature prediction to informed decision-making supported by trustworthy bias indicators.

We propose an **uncertainty-aware proxy reasoning (UAPR)** framework for media bias detection, which introduces proxy attribute prediction alongside composite uncertainty (CU) estimation to stratify decision quality. Proxy attributes serve as informative intermediate steps that help narrow the decision space, while uncertainty quantifies prediction variance. To evaluate uncertainty scoring, we conduct a comprehensive comparison of methods using coverage errors, prediction set size, and a newly proposed **Coverage-Efficiency Harmonic Index (CEHI)**. CEHI balances the tradeoff between coverage errors and prediction set size, providing a more holistic assessment of uncertainty quality. Our proposed CU achieves 0.867 CEHI for uncertainty evaluation, and UAPR outperforms the GPT4.1-mini chain-of-thought (CoT) baseline by 6.25% F1 score on the BASIL dataset. Our experiments address several research questions: **RQ1:** What techniques effectively leverage uncertainty and proxy attributes to improve prediction accuracy and decision quality? **RQ2:** How can proxy attributes and CoT reasoning enhance decision quality through quantitative uncertainty assessment? **RQ3:** Are there entities that achieve high prediction accuracy yet exhibit high uncertainty? Our contributions are summarized as follows:

- We present the first study to investigate uncertainty estimation in media bias detection, enabling effective bias risk stratification.
- We propose an automatic uncertainty-aware bias detection framework that leverages proxy attributes to improve prediction accuracy.
- We provide comprehensive uncertainty measurement results specific to media bias detection.
- We conduct in-depth analyses of proxy parameters, revealing their potential to enhance decision quality in bias assessment.

2 Related Work

2.1 Automatic Media Bias Detection

Sentence-level media bias detection typically focuses on two main types of bias: lexicon bias and information bias (Fan et al., 2019). Some studies address these types separately due to their distinct manifestations (Maab et al., 2023a). Lexicon bias is closely tied to linguistic features, whereas information bias depends more on context—for example, highlighting some details while downplaying others. To address this, target-aware context augmentation (Maab et al., 2023b) leverages other statements

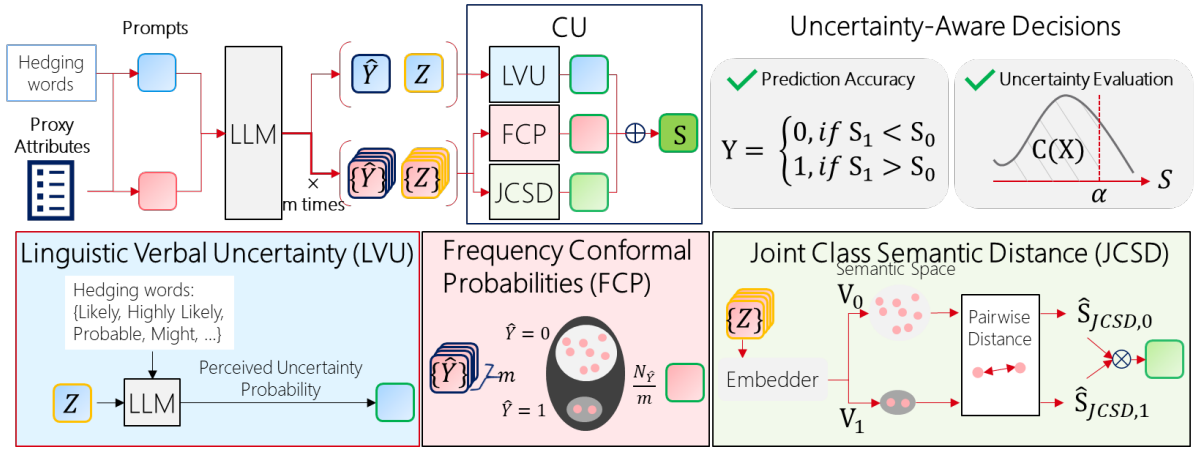


Figure 2: Overview of the proposed UAPR framework. Given prompt input X, the model produces an intermediate prediction $\hat{y}$ and intermediate reasoning outputs Z. Based on these outputs, an uncertainty score S is computed via compositional uncertainty (CU), which integrates linguistic variability uncertainty (LVU), prediction frequency consistency (FCP), and joint class semantic discrepancy (JCSD). S serves as critical value to generate conformal prediction set $C(X)$ or final prediction Y.

biased toward the same topic to strengthen the bias signal and stabilize model training. Constructing an event-aware graph attention network can further improve bias discrimination (Lei and Huang, 2024). These media bias detection tasks rely heavily on high-quality, human-annotated datasets such as BASIL (Fan et al., 2019), reflecting the inherent ambiguity and inconsistency of free-form labeling. However, strong dependence on supervised learning limits generalization: models often overfit to dataset-specific lexical or topical cues and fail under temporal or domain shifts. They also struggle with label scarcity and subjectivity common in political and media contexts. In contrast, leveraging LLMs provides a promising direction for addressing multiple bias types in a unified, training-free framework.

2.2 Bias Detection Using LLMs

Few-shot LLM inference can improve bias detection accuracy (Maab et al., 2024), although performance still varies depending on the chosen few-shot examples. Using LLM-generated annotations in a few-shot setting has empirically achieved results comparable to human-annotated data (Horych et al., 2025). LLMs can be guided to understand different bias cases through tailored prompts and then perform detection by comparing distances between latent embeddings (Lin et al., 2024).

2.3 Uncertainty Quantification for NLP

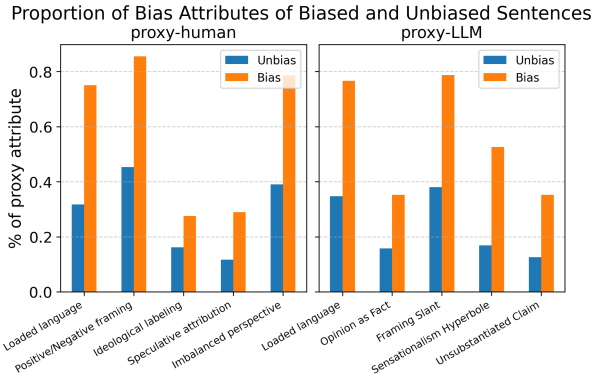
Generative models have consistently struggled with faithfulness and robustness issues in validated applications (Huang et al., 2025). Uncertainty measurement provides a critical way to evaluate model outputs beyond standard accuracy metrics. However, prior research on uncertainty has primarily focused on traditional classification models. Only few recent studies have explored conformal prediction sets for LLMs, using frequency, entropy, or text similarity calculation (Su et al., 2024; Qiu and Miikkulainen, 2024).

Most LLM studies still neglect uncertainty estimation on common benchmarks (Ye et al., 2024). Token probability-based uncertainty can be derived directly from softmax outputs for next-token prediction (Han et al., 2024), while numerical verbal uncertainty is obtained by prompting LLMs to self-report confidence (Xiong et al., 2024). Consistency-based uncertainty measures disagreement across multiple generated responses (Kuhn et al., 2023). Linguistic verbal uncertainty captures hedging level in LLM outputs (Belém et al., 2024), such as "certainly", "probably", or "might". Adaptive conformal prediction (Cherian et al., 2024) improves claim filtering by accounting for prompt characteristics. Calibrated uncertainty scores have been shown to closely reflect potential error rates, providing robust confidence estimates. Without uncertainty estimation, media bias detection remains largely limited to binary predictions, restricting its reliability and applicability in real-world settings.

3 Method

We introduce uncertainty estimation methods to evaluate reasoning quality and present the

uncertainty-aware proxy reasoning (UAPR) framework. A high-level overview is illustrated in Figure 2, external knowledge like hedging words or proxy attributes are combined to prompts X to generate intermediate reasoning Z with intermediate decision $\hat{y}$. Z and $\hat{y}$ are then to compute CU—our proposed method to estimate uncertainty S. S is then used to generate conformal prediction set $C(X)$ or final prediction Y. The full UAPR algorithm is provided in Appendix § B.2.



Human Knowledge-Based Proxy Attributes	
Loaded Language	Emotionally charged or judgmental words (e.g., "public menace", "rage")
Positive/Negative Framing	Portraying someone favorably or unfavorably via comparison
Ideological Labeling	Labeling with partisan or ideological terms
Speculative Attribution	Suggesting hidden motives without evidence
Imbalanced Perspective	Only presenting one side of the issue
LLM-Based Proxy Attributes	
Loaded Language	Use of emotionally charged words (e.g., "disastrous," "heroic")
Opinion as Fact	Presenting a subjective belief as an objective truth
Framing/Slant	Selecting details or words to favor one perspective
Sensationalism/Hyperbole	Exaggeration to provoke a strong reaction
Unsubstantiated Claim	Making a statement without providing evidence

Figure 3: Statistics and definitions of proxy attributes.

3.1 Preliminaries of Conformal Prediction

Conformal prediction (CP) is a model-agnostic, distribution-free approach for uncertainty estimation that generates prediction sets designed to contain the ground-truth labels with a specified error rate α. A commonly used CP method is *split CP*, which provides formal coverage guarantees. Given input features X and labels Y, and a target error rate $\alpha \in (0, 1)$, the conformal prediction set $C_{1-\alpha}(X_{test})$ satisfies:

$$P(X_{test} \in C_{1-\alpha}(X_{test})) \leq 1 - \alpha. \quad (1)$$

For uncertainty estimation in LLMs, logit-free CP estimation (Su et al., 2024) combines frequency, entropy, and semantic similarity to construct conformal sets with valid coverage guarantees while keeping prediction sizes small. Further derivation details are provided in Appendix § B.1.

3.2 Proxy Attribute Extraction

We identify intermediate reasoning steps within a conditional inference paradigm similar to chain-of-thought (CoT) reasoning. We refer to these intermediate steps as *proxy attributes* for media bias detection. We consider three types of proxy attributes, illustrated in Figure 3:

- *Knowledge-based proxy attributes*: derived from definitions and concepts established in prior social science research (Piskorski et al., 2023; Rodrigo-Ginés et al., 2024; Spinde et al., 2023).
- *LLM-based proxy attributes*: obtained by prompting GPT-4.1-mini for relevant bias cues. This setting excludes external expert input, serving as a baseline for automated proxy generation when domain-specific knowledge is unavailable.
- *Joint proxy attributes*: created by combining knowledge-based and LLM-based attributes using GPT-4 prompts to reduce redundancy.

We denote UAPR-* as the UAPR framework instantiated with a specific proxy attribute type, where the suffix indicates the attribute used (*e.g.*, UAPR-Joint uses Joint proxy attributes). In this framework, information gain from bias cues reduces prediction uncertainty by calibrating decision boundaries through conformal sets. Uncertainty estimation thus serves as a validation mechanism for our hypothesis, testing whether proxy attributes function as effective intermediate reasoning steps.

3.3 Composite Uncertainty (CU) Estimation

We introduce **composite uncertainty (CU)** to capture multi-aspect statistical cues using the conformal prediction framework. CU combines three components: **Linguistic-Verbal Uncertainty (LVU)**, **Frequency-based Conformal Probabilities (FCP)**, and **Joint Class Semantic Distance (JCSD)**. Given N input prompts X, an LLM generates intermediate reasoning outputs $Z = z_1, z_2, ..., z_N$ and final prediction $Y = y_1, ..., y_N$. To calibrate predictions in an uncertainty-aware manner, we denote $\hat{Y}$ as the intermediate prediction, distinct from the final prediction Y. Across m inference rounds, the multiple outputs are represented as $\{Z_i\}_{i=1}^m$ and $\{\hat{Y}_i\}_{i=1}^m$, which we simplify as $\{Z\}$ and $\{\hat{Y}\}$. The composite uncertainty score S is defined as weighted sum of the three measures:

$$S = \text{FCP}(\{\hat{Y}\}) + \text{LVU}(Z) + \lambda\ \text{JCSD}(\{Z\}), \quad (2)$$

where λ is a scaling hyperparameter accounting for $JCSD$ numeric ranges, with a default of 100, as results are largely insensitive to this choice (see Appendix § D.2). Higher CU indicates greater uncertainty about the question, and lower CU indicates more confident reasoning.

LVU estimates uncertainty by analyzing the LLM's use of hedging language in its reasoning. The model is prompted to generate an explanation (Z) for its prediction, prefaced with a hedging expression (e.g., "very likely," "might," or "not likely"). The LLM then evaluates this hedged explanation to produce a numeric uncertainty score in $[0, 1]$, reflecting the confidence implied by the hedging language; implementation details are provided in Appendix B.6.

FCP quantifies uncertainty using frequency calculations, a proven approach for representing conformal prediction results. For the k^{th} class, we compute the occurrence rate across m inference rounds, $\{\hat{Y}\}$, and define FCP as:

$$S_{FCP,k} = FCP(\{\hat{Y}\}, m) = 1 - \frac{N_k}{m}, \quad (3)$$

where N_k is the number of times $\hat{Y}$ is predicted as class k. FCP reflects the inverse of model confidence: higher prediction frequency for a class corresponds to lower uncertainty for that class.

JCSD estimates a multiplicative uncertainty value by capturing semantic variability across all classes. Larger JCSD values indicate higher semantic diversity within each class, while smaller values reflect greater consistency. Building on the idea of using semantic distance to measure uncertainty (Qiu and Miikkulainen, 2024), we proceed as follows. Let Z_k denote the set of intermediate reasoning outputs predicted as class k. Each sentence is encoded into a D-dimensional embeddings using a pretrained BERT model, forming $V_k \in \mathbb{R}^{N_k \times D}$. The average pairwise cosine distance within each class is computed as:

$$\hat{S}_{JCSD,k} = \frac{2}{|Z_k|(|Z_k|-1)} \sum_{i,j \in Z_k; i<j} d_{\cos}(v_i, v_j),$$

where $d_{\cos}$ denotes cosine distance. The final JCSD uncertainty aggregates the K class-wise distances by taking their product:

$$\text{JCSD}(\{Z\}) = \Pi_{k=1}^{K} \hat{S}_{JCSD,k}. \quad (4)$$

Uncertainty-Aware Decisions uses the composite uncertainty (CU) scores $S = \{S_k\}_{k=1}^{K}$ to adjust the original decision boundary, where S_k represents the CU score for samples predicted as class k in the intermediate prediction $\hat{Y}$. The final media bias prediction Y is then determined by selecting the class with the lowest uncertainty:

$$Y = \arg\min_k CU_k. \quad (5)$$

4 Experiments and Discussions

4.1 Experimental Setup

Experiments are conducted using 10-fold story-split cross-validation on the BASIL benchmark dataset (Fan et al., 2019), following prior work (Lin et al., 2024). The story-split scheme ensures a fair comparison that sentences from the same article do not appear in both training and test sets. The BASIL dataset contains 7,984 samples covering the same political events reported by three distinct outlets: New York Times, Fox News, and Huffington Post. Each sentence is annotated with bias labels based on consensus from three raters.

Metric: Following (Su et al., 2024), we adopt standard metrics for uncertainty evaluation:

- **Empirical Coverage Rate (ECR)**: measures whether the conformal prediction method achieves its expected theoretical coverage.
- **Size-Stratified Coverage (SSC)**: evaluates the lowest coverage rate among prediction sets of different sizes, reflecting worst-case reliability.
- **Average Prediction Set Size (APSS)**: assesses the efficiency of the conformal predictor; smaller APSS values indicate higher efficiency.
- **Coverage-Efficiency Harmonic Index (CEHI)**: computes the harmonic mean of ECR and APSS, providing a balanced measure of coverage and efficiency.

Training	**P**	**R**	**F1**	**Zero-Shot**	**P**	**R**	**F1**
Roberta	0.43	0.41	0.42	GPT4.1-mini	0.40	0.61	0.48
Bert	0.38	0.35	0.37	+CoT	0.43	0.53	0.48
Longformer	0.46	0.45	0.46	+CU	0.41	0.61	0.49
ERG	0.50	0.54	0.52	+Proxy+Few-Shot	0.42	0.57	0.49
Zero-Shot	**P**	**R**	**F1**	LoFreeCP	0.38	0.74	0.50
IndiVec	0.32	0.35	0.34	UAPR-CoT	0.45	0.47	0.46
LLaMA3-8B	0.27	0.78	0.40	UAPR-Human	0.40	0.69	**0.51**
Gpt4o	0.39	0.65	0.48	UAPR-LLM	0.46	0.36	0.40
				UAPR-Joint	0.41	0.59	0.48

Table 1: Comparison of BASIL media bias detection performance with SOTA methods on multiple scenarios.

	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LLaMA3-8B	0.816	1.654	0.469	0.486	0.466	1.000	0.466	0.636	0.466	1.000	0.466	0.636	0.466	1.000	0.466	0.636
Gpt4o	0.672	1.000	0.672	0.804	0.672	1.000	0.672	0.804	0.672	1.000	0.672	0.804	0.672	1.000	0.672	0.804
GPT4.1-mini	0.808	1.169	0.768	0.819	0.756	1.049	0.744	0.843	0.738	1.000	0.738	0.849	0.723	1.000	0.738	0.849
+CoT	0.802	1.076	0.785	0.858	0.766	1.000	0.766	0.867	0.766	1.000	0.766	0.867	0.766	1.000	0.766	0.867
+CU	0.802	1.132	0.771	0.834	0.758	1.033	0.749	0.849	0.742	1.000	0.742	0.852	0.742	1.000	0.742	0.852
UAPR-CoT	0.830	1.109	0.809	0.859	0.808	1.054	0.797	0.872	0.786	1.000	0.786	**0.880**	0.786	1.000	0.786	**0.880**
UAPR-Human	0.810	1.186	0.766	0.812	0.782	1.114	0.754	0.831	0.744	1.032	0.736	0.841	0.731	1.000	0.731	0.845
UAPR-LLM	0.806	1.061	0.793	**0.867**	0.788	1.011	0.786	**0.877**	0.788	1.010	0.786	0.877	0.784	1.001	0.784	0.879
UAPR-Joint	0.820	1.171	0.783	0.825	0.796	1.106	0.771	0.842	0.764	1.033	0.756	0.854	0.746	1.000	0.746	0.855

Table 2: BASIL dataset benchmark results for uncertainty measurement in media bias detection. The uncertainty method shown is selected based on the ablation analysis. Complete results for $\alpha > 0.35$ are provided in Appendix Section D.3.

Implementation details are in Appendix § B.3. These experiments address a key research gap, as uncertainty assessment has not yet been explored in media bias detection tasks.

4.2 Main Results

We next report results on the BASIL dataset; results on other datasets are provided in Appendix § D.1.

Media bias detection results: Refer to Table 1. Previous work fine-tunes various transformer models (Lei and Huang, 2024; Van Den Berg and Markert, 2020; Lei et al., 2022) to form strong training-based baselines. Among them, ERG (Lei and Huang, 2024) reaches 0.52 F1 by exploiting causal event structure. However, these methods carry a significant risk of overfitting—a concern given the label sparsity and annotation subjectivity. Although zero-shot LLMs offer a promising path for improved generalization, their accuracy remains limited. Gpt4o and GPT4.1-mini achieve a 0.48 F1 score, trailing ERG by 0.04 points. LLaMA3-8B underperforms further, and IndiVec (Lin et al., 2024), which relies on embedding distance comparisons, does not improve accuracy. These results highlight that the inherent complexity of the task poses significant challenges for zero-shot settings.

UAPR-Human, UAPR-LLM, and UAPR-Joint are the variance of our UAPR framework using different proxy attribute sets introduced in the method section. UAPR-CoT replaces the proxy attribute reasoning step with standard CoT without pre-specified knowledge. +CU represents uncertainty-aware detection without a proxy. +Proxy+Few-Shot represents using a proxy with few-shot examples. The result shows that **UAPR-Human** achieves 0.51 F1 score, representing a 6.25% relative improvement over the GPT4.1-mini baseline with p-value < 0.001. However, the other UAPR with other proxy attributes fall below even the GPT4.1-mini baseline. Only UAPR-Human surpasses other previous works, such as LoFreeCP (Su et al., 2024) or GPT4.1-mini with few-shot. The result address **RQ1**, confirming that knowledge-based proxy attributes, combined with uncertainty-aware decision making, enable the LLM's reasoning process to better align with human judgment. Consequently, we observe accuracy comparable to training-based prior work.

Measuring uncertainty for media bias detection: Uncertainty is a crucial evaluation dimension complementing detection accuracy. Table 2 (and detailed results in Appendix D.3) shows the uncertainty measurement results for GPT4.1-mini. UAPR-CoT, +CoT, and UAPR-LLM—approaches that exploit LLM reasoning or proxy interence—achieve higher CEHI across various α conditions, For example, UAPR-LLM outperforms the baseline (GPT4.1-mini) by 4.8% at $\alpha = 0.2$ and 3.4% at $\alpha = 0.25$; UAPR-CoT achieves 0.88 CEHI at $\alpha = 0.3$. The results also address **RQ2** that with quantitative uncertainty assessment, we found UAPR-LLM or UAPR-CoT reasoning maintain good coverage rate with a relatively smaller conformal prediction set, demonstrating good prediction efficiency. However, according to Table 1, UAPR-LLM and UAPR-CoT do not outperform the baseline, suggesting that even with lower uncertainty in intermediate steps, anchoring bias (Nguyen, 2024) in the initial reasoning can still lead the LLM's outputs to deviate from human labels.

Proxy attribute uncertainty: Table 3 reports the uncertainty assessment of the proxy attributes, reflecting their individual contribution to decision

Human Knowledge-Based	**SD(E^{-4})**	**FCP**	**N_p**
Positive/Negative framing	7.04	0.114	4218
Imbalanced perspective	6.77	0.172	3717
Loaded language	3.60	0.108	3201
Ideological labeling	8.10	0.093	1458
Speculative attribution	8.46	0.081	1200
LLM-Based	**SD(E^{-4})**	**FCP**	**N_p**
Framing Slant	5.13	0.134	3675
Loaded language	1.34	0.111	3428
Sensationalism Hyperbole	3.48	0.098	1911
Opinion as Fact	2.49	0.178	1560
Unsubstantiated Claim	2.81	0.144	1361
Joint proxy attribute	**SD(E^{-4})**	**FCP**	**N_p**
Imbalanced perspective	6.32	0.183	4422
Positive/Negative framing	3.43	0.114	4286
Loaded language	1.72	0.110	3306
Sensationalism Hyperbole	5.86	0.101	2172
Ideological labeling	5.29	0.096	1571
Opinion as Fact	5.57	0.166	1477
Speculative attribution	5.82	0.086	1459
Unsubstantiated Claim	2.49	0.123	702

Table 3: Uncertainty evaluation results for proxy attributes on BASIL dataset, each with N_p predicted samples. SD and FCP are semantic distance and frequency-based uncertainty measures.

quality when used with LLMs. The consistently low semantic distance (SD) indicates high consistency in wording and descriptive patterns when responding to these attributes. This is expected, as these attributes are expert-defined to simplify the complex nature of media bias. Most attributes also show an FCP value of around 0.1, indicating moderate frequency uncertainty. We also report N_p, the number of samples predicted to exhibit each proxy attribute, as a reference for their prevalence.

For common human-based proxy attributes ($N_p > 3000$), the *imbalanced perspective* achieves relatively high FCP, while *loaded language* and *positive/negative framing* yield more consistent predictions. Notably, these two attributes also appear in the LLM-generated proxy set. Although some highly frequent attributes overlap between the human-based and LLM-generated sets, their prediction outcomes differ because differences in less frequent attributes further amplify conceptual divergence, making their concept fundamentally different. from Tables 1 and 2, we observe that human-based proxy attributes improve decision accuracy, while LLM-based proxies contribute more to enhancing decision quality. In brief, this experiment highlights how proxy definitions shape both prediction behavior and uncertainty evaluation.

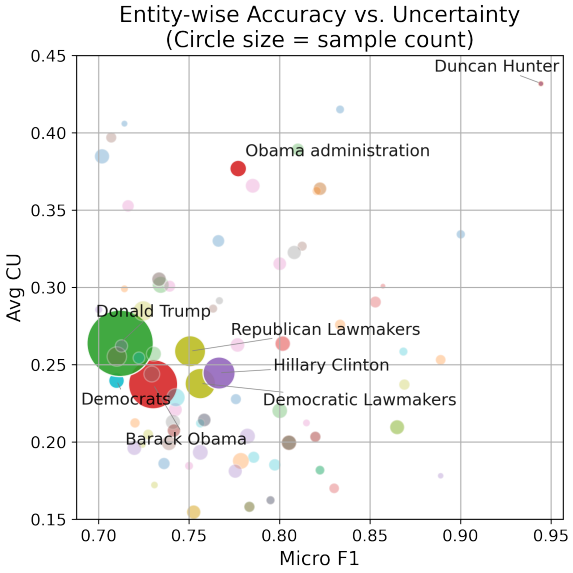


Figure 4: F1 score *vs.* CU across media entities. Each circle represents a media entity, with its size indicating the number of associated sentences.

The Joint proxy attribute rows in Table 3 shows results when these proxy sets are combined. Common attributes tend to produce larger prediction sets, while less frequent attributes often yield smaller sets. Further investigation of proxy attributes is essential for achieving consistent and robust bias measurement.

4.3 Ablation Study of Uncertainty Scoring Approaches

We conduct an uncertainty assessment ablation study on the proposed CU estimation using a target error rate ($\alpha = 0.2$). We also compared with other uncertainty estimation methods in (Tao et al., 2025), including numerical verbal uncertainty (Tian et al., 2023) and LoFreeCP (Su et al., 2024) in Appendix Section D.3. Table 4 presents the performance of different uncertainty scoring approaches on BASIL. Using *CU* consistently yields higher CEHI scores across various proxy settings, with the *LVU+FCP* combination contributing the most to overall uncertainty estimation quality. Incorporating JCSD emphasizes samples with high uncertainty in both biased and unbiased classes, which helps explain the strong performance of the *LVU+JCSD* configuration for UAPR-LLM.

Method	+CU				UAPR-Human				UAPR-LLM				UAPR-Joint			
Metric	ECR	APSS	SSC	CEHI	ECR	APSS	SSC	CEHI	ECR	APSS	SSC	CEHI	ECR	APSS	SSC	CEHI
FCP	0.808	1.169	0.768	0.819	0.805	1.218	0.751	0.793	0.803	1.120	0.776	0.840	0.804	1.163	0.766	0.820
JCSD	0.947	1.801	0.734	0.329	0.969	1.799	0.848	0.332	0.967	1.800	0.836	0.332	0.970	1.800	0.850	0.331
LVU	0.860	1.270	0.808	0.790	0.865	1.477	0.743	0.652	0.816	1.098	0.796	0.857	0.857	1.439	0.745	0.678
FCP + JCSD	0.802	1.155	0.766	0.823	0.800	1.204	0.749	0.798	0.801	1.115	0.775	0.841	0.803	1.159	0.765	0.821
LVU + FCP	0.805	1.138	0.773	0.832	0.815	1.204	0.768	0.805	0.821	1.090	0.803	0.863	0.823	1.180	0.784	0.821
LVU + JCSD	0.846	1.239	0.798	0.802	0.858	1.454	0.739	0.667	0.809	1.067	0.795	**0.867**	0.851	1.419	0.744	0.691
CU	0.802	1.132	0.771	**0.834**	0.810	1.186	0.766	**0.812**	0.806	1.061	0.793	**0.867**	0.820	1.171	0.783	**0.825**

Table 4: Comparison of different uncertainty scoring approaches for media bias detection on BASIL dataset. Our proposed CU estimation method (LVU + FCP + JCSD) outperforms other uncertainty measurement techniques.

Entity	R_{CU}	G_{CU}		
		R_{LVU}	R_{FCP}	R_{JCSD}
Duncan Hunter	44%	12%	87%	25%
Obama Administration	54%	40%	65%	23%

Table 5: Case study of uncertainty distribution for high uncertainty samples. For each entity, R_{CU} is the proportion of samples in the top-ranked CU group G_{CU}. R_M denotes the proportion of samples in G_{CU} that also belong to the top-ranked M group, where $M \in \{LVU, FCP, JCSP\}$.

4.4 Uncertainty Difference of media Entities

Figure 4 shows the F1–CU score relationship across media entities in BASIL. Each circle represents an entity, with its radius indicating the number of associated sentences. We observe outliers—entities like Duncan Hunter and Obama Administration—that exhibit high uncertainty despite strong F1 scores, indicating both high accuracy and uncertainty. We further examine these cases by analyzing the distribution of their uncertainty measure in Table 5. Specifically, we group the top 25% of samples in the dataset for each uncertainty measure as top-M groups, where M can be LVU, FCP, JCSD, or CU. For an entity with N_e samples, we define $R_{CU} = \frac{N_s}{N_e}$ as the proportion of samples N_s that belong to the top-CU group. To understand which uncertainty types contribute the most to these top-CU samples, we calculate the proportion $R_M = \frac{N_M}{N_s}$, where N_M is the number of top-CU samples that also in the top-M group.

As shown in Table 5, the top-CU samples account for 44% for *Duncan Hunter*, while 54% for *Obama Administration*. Among *Duncan Hunter*'s top-CU samples, 87% are also in the top-FCP group, indicating that frequency-based uncertainty is the dominant factor of uncertainty. In contrast, the R_{LVU} for *Duncan Hunter* is only 12%, showing relatively low linguistic uncertainty compared to *Obama Administration*.

We observe that the low R_{LVU} of *Duncan Hunter* stems from the sentences where the entity appears within quotations. These quotations often act as strong linguistic cues for confident statements, reducing linguistic uncertainty. However, the multi-round reasoning using proxy attributes reveals diverse reasoning paths in a high R_{FCP}. In contrast, the higher R_{LVU} of *Obama Administration* aligns with the media's tendency to mirror political tone and Obama's rhetorical style. The *Obama Administration*'s deliberate control over its public messaging contributed to a lower R_{FCP} and more consistent reporting across mass media outlets. Overall, LVU captures uncertainty from linguistic cues, while FCP reflects prediction consistency across reasoning steps. Since media bias can appear through multiple pathways, incorporating both dimensions is essential for a more comprehensive uncertainty assessment.

5 Conclusion

This study is the first to improve media bias decision quality by integrating uncertainty estimation and proxy attributes into a training-free LLM framework. Our findings on various uncertainty scores highlight the limitations of prior work that focused solely on detection accuracy while overlooking decision quality. The analysis of proxy attributes demonstrates how intermediate reasoning steps can enhance prediction reliability and transparency. Future work includes real-world user studies to validate the quality and usability of proxy attributes. We also aim to enhance graph-based reasoning and context augmentation within our uncertainty-aware framework to improve explainable bias detection.

6 Limitations

Although UAPR improves decision quality, it increases both cost and latency because estimating the uncertainty S requires m+2 prompts per question (m for FCP, 2 for LVU). These costs can be reduced by using smaller models, such as GPT-*-mini, or by training a custom LLM for media bias detection. To our knowledge, the current open source LLMs

do not have knowledge for generating proper reasoning for media bias detection. In this context, developing a custom model via reinforcement learning would require a minimum of an NVIDIA A100 GPU and datasets like BASIL and BABE.

In this paper we did not use separate methods to tackle lexical and information bias. The two bias types are handled by LLM by evaluating on our designed proxies. In addition, we do not discuss the our result with respect to these two bias types.

7 Ethical Considerations

This study does not involve the collection of new user data or interaction with human subjects. All experiments are conducted solely on the publicly available BASIL, BABE, and BIASEDSENT benchmark datasets for media bias detection. No personal, sensitive, or identifiable information is collected, processed, or stored, and no additional ethical concerns arise beyond those already addressed by the dataset's original data collection and release. However, there is a potential ethical issue that detection tools could be misused for selective censorship or political manipulation. A prudent decision on using powerful tools needs to be considered for the good of society.

References

Catarina Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8467–8502.

Chin-Po Chen and Jeng-Lin Li. 2024. Profiling patient transcript using large language model reasoning augmentation for alzheimer's disease detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4.

John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842.

John Corner. 2013. Theorising media: Power, form and subjectivity. In *Theorising Media*. Manchester University Press.

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.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.

Po-Hsuan Huang, Jeng-Lin Li, Chin-Po Chen, Ming-Ching Chang, and Wei-Chao Chen. 2025. Who brings the frisbee: Probing hidden hallucination factors in large vision-language model via causality analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6125–6135. IEEE.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5225–5238.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 10040–10050.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam Fai Wong. 2024. Indivec: An exploration of

leveraging large language models for media bias detection with fine-grained bias indicators. In *18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024-Findings of EACL 2024*, pages 1038–1050. Association for Computational Linguistics (ACL).

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023a. An effective approach for informational and lexical bias detection. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 66–77.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023b. Target-aware contextual political bias detection in news. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 782–792.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. Media bias detection across families of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4083–4098.

Matteo Morelli, Maria Casagrande, and Giuseppe Forte. 2022. Decision making: A theoretical review. *Integrative Psychological and Behavioral Science*, 56(3):609–629.

Jeremy K Nguyen. 2024. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533.

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.

Noah Schutte, Grigorii Veviurko, Krzysztof Postek, and Neil Yorke-Smith. 2025. Sufficient decision proxies for decision-focused learning. *arXiv preprint arXiv:2505.03953*.

Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

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.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. 2025. Revisiting uncertainty estimation and calibration of large language models. *arXiv preprint arXiv:2505.23854*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Esther Van Den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326.

Martin J Vilela and Gbenga F Oluyemi. 2021. Decision-making: Concepts, principles, and uncertainty. In *Value of Information and Flexibility: Making Decisions Under Uncertainties*, pages 1–20. Springer.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–27.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

A Concept Formulation for Bias Reasoning

Distinct problem-solving strategies yield different levels of cognitive loads, which also vary decision error rates (Sweller, 1988). That is, reasoning paths for bias detection can prominently impact the detection accuracy. This work introduces uncertainty estimation alongside proxy attributes to support intermediate reasoning. The approach is analogous to the use of diagnostic scales in the medical field for assessing mental health conditions. These scales provide structured evidence that helps clinicians interpret complex mental states based on behavioral assessments. The reliability of each item on the scale depends on its consistency across diverse clinical scenarios, highlighting the critical role of uncertainty estimation in this context. A relevant example involves using LLMs to aid in Alzheimer's disease detection (Chen and Li, 2024). This case illustrates how designing intermediate reasoning steps can enhance the overall decision-making process in complex diagnostic tasks. We consolidate these ideas to formulate our proposed media bias detection framework for enhanced decision accuracy and decision quality.

Media bias arises because the same event can be selectively framed. For example, in BASIL, identical topics are reported by different outlets with divergent emphases; partial, deliberate, or inadvertent information selection produces potentially biased statements. Prior social science work (Rodrigo-Ginés et al., 2024) shows such statements can be flagged via diagnostic attributes (e.g., loaded language, framing, selective omission). These attributes act as intermediate reasoning variables (Z) in our pipeline for the input prompt X and output Y. Hence, given the LLM f, the reasoning procedure is formalized in a probability form:

$$P(Z, Y|X) = P(Y|Z)P(Z|X). \tag{6}$$

Considering the uncertainty U, we can rewrite the reasoning procedure with the Bayesian theorem:

$$P(Z, Y|U, X) = \frac{P(U|X, Z, Y)P(Z, Y|X)}{P(U|X)}. \tag{7}$$

We aim to maximize $P(Z, Y|U, X)$, which depends on $P(U|X, Z, Y)$ and $P(Z, Y|X)$. These two terms provide a clue to delve into the uncertainty and intermediate steps. The minimal values for the denominator term $P(U|X)$ denote the lower likelihood to estimate the uncertainty directly from X. First, $P(U|X, Z, Y)$, is positively correlated with different terms, $P(U|X, Z)$ and $P(U|X, Y)$. Here, the uncertainty estimation approaches in our proposed composite uncertainty (CU) method correspond to these mathematical forms, revealing the roles of robust uncertainty estimation.

- Linguistic-Verbal Uncertainty (LVU): estimates the uncertainty by verbalizing Z with the dedicated hedging prompts. LVU corresponds to $P(U|X, Z)$, focusing on the linguistic cues expressed in Z.
- Frequency Conformal Probabilities (FCP) Uncertainty: calculates the frequency of bias prediction candidates, which corresponds to $P(U|X, Y)$.
- Joint Class Semantic Distance (JCSD) Uncertainty: designs to combine the effects of Y and Z. The conditional setting formulates the uncertainty estimation problem with separate terms $P(U|X, Z, Y = 0)$ and $P(U|X, Z, Y = 1)$.

Second, the term $P(Z, Y|X)$ is decomposed via the standard CoT approach in equation (6). We introduce proxy attributes A as a latent factor of Z that can increase $P(Z(A), Y|X)$. The human knowledge-based proxy attributes are derived based on empirical evidence with high $P(Y|Z(A))$. In this case, the proxy attributes will be effective if $P(Z(A)|X)$ is large, which has been manually examined for the attribute prediction accuracy. Our main results directly evaluate the final accuracy to show the effectiveness of introducing proxy attributes. This study remains future work to validate the alignment of attribute prediction with humans' understanding due to current resource limitations in large-scale human annotators.

B Implementation Details

B.1 Conformal Set Derivation

First, measure the nonconformity score (uncertainty measure in this paper) $S(x, y) \in \mathbb{R}$ to indicate the sample-wise uncertainty in the training data. Second, we specify the quantile of the nonconformity score $\hat{q}$ as $\frac{(n+1)(1-\alpha)}{n}$, where n denotes the number of samples in the training data. Finally, the prediction set is restricted to the $\hat{q}$ quantile to include sufficiently confident samples in the set:

$C(X_{test}) = \{Y : S(X_{test}, Y) \leq \hat{q}\}$. The varied size of this prediction set can reflect the uncertainty of these predictions.

B.1.1 Proof Sketch of Uncertainty Coverage Guarantee

Assuming that the calibration set $(X_i, Y_i)_{i=1,\ldots,n}$ and (X_{test}, Y_{test}) are $i.i.d$, the non-conformity score S can be subsequently calculated. The desired error rates α_1 and α_2, where $\alpha_1 > \alpha_2$ will lead to $\hat{q}_1 \leq \hat{q}_2$. The conformity set $C(X_{test}) = \{Y : S(X_{test}, Y) \leq \hat{q}\}$, thus we can derive $C_{1-\alpha_1}(X) \subseteq C_{1-\alpha_2}(X)$.

Algorithm 1 Uncertainty-aware proxy reasoning

Require: Input prompt X, LLM $\mathcal{M}$, rounds m, weight λ, encoder $\mathcal{E}$, nonconformity score quantile $\hat{q}$
Ensure: $Y, C(X)$

// Linguistic-Verbal Uncertainty (LVU)
LVU $\leftarrow \mathcal{M}(Z)$, where $Z \leftarrow \mathcal{M}(X)$
// Multi-round sampling
Initialize $\{\hat{Y}\} \leftarrow \emptyset$, $\{Z\} \leftarrow \emptyset$
for $i = 1$ to m **do**
    $z_i, \hat{y}_i \leftarrow \mathcal{M}(X)$
    $\{Z\} \leftarrow \{Z\} \cup \{z_i\}$, $\{\hat{Y}\} \leftarrow \{\hat{Y}\} \cup \{\hat{y}_i\}$
end for
Identify candidate classes K from $\{\hat{Y}\}$
// Frequency Conformal Probabilities (FCP)
for each class $k \in K$ **do**
    $\text{FCP}_k \leftarrow \frac{|\{\hat{y}_i = k\}|}{m}$
end for
// Joint Class Semantic Distance (JCSD)
for each class $k \in K$ **do**
    $Z_k \leftarrow \{z_i \mid \hat{y}_i = k\}$
    $V_k \leftarrow [\mathcal{E}(z) \text{ for } z \in Z_k]$
    $\hat{S}_{\text{JCSD},k} \leftarrow \frac{2}{|Z_k|(|Z_k|-1)} \sum_{i<j} d_{\cos}(v_i^k, v_j^k)$
end for
JCSD $\leftarrow \prod_{k=1}^{K} \hat{S}_{\text{JCSD},k}$
Initialize $C(X) \leftarrow \emptyset$
for each class $k \in K$ **do**
    $S_k \leftarrow \text{FCP}_k + \text{LVU} + \text{JCSD} \cdot \lambda$
    if $S_k \leq \hat{q}$ **then**
        $C(X) \leftarrow C(X) \cup \{k\}$
    end if
end for
$Y \leftarrow \arg\min_k S_k$
return $Y, C(X)$

B.2 The UAPR Algorithm

Algorithm 1 presents the proposed method, Uncertainty-Aware Proxy Reasoning (UAPR), as discussed in Section 3. UAPR takes a prompt X as input, computes and aggregates three types of uncertainty scores (LVU, FCP, JCSD) for each decision, aggregates them, and outputs the final decision Y. Using a user-specified error rate α and nonconformity scores from the training data, it then derives the quantile $\hat{q}$ to construct the conformal prediction set $C(X)$.

B.3 Implementation details of uncertainty measurement

Empirical Coverage Rate (ECR): measures whether the conformal prediction method achieves its expected theoretical coverage. It is calculated by whether the ground truth is contained in the conformal prediction set $C_{1-\alpha}(X_{test})$. The equation is defined as:

$$ECR = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y_i \in C_{1-\alpha}(x_i)\} \tag{8}$$

, where $x_i \in X_{test}$ and $y_i \in Y_{test}$, and N is the number of testing data

Size-Stratified Coverage (SSC): evaluates the lowest coverage rate among prediction sets of different sizes to show the worst-case reliability. It is implemented by first grouping prediction sets by their size and computing the coverage rate for each group. Finally, we report the worse coverage rate among all prediction set groups.

$$SSC = \min_s \frac{1}{N_s} \sum_{\substack{i: \\ |C_{1-\alpha}(x_i)|=s}} \mathbf{1}\{y_i \in C_{1-\alpha}(x_i)\} \tag{9}$$

, where N_s is the number of prediction sets of size s.

Average Prediction Set Size (APSS): assesses the efficiency of the conformal predictor; a smaller APSS indicates higher efficiency. APSS is computed by the average size of all the conformal prediction sets.

$$APSS = \frac{1}{N} \sum_{i=1}^{N} |C_{1-\alpha}(x_i)| \tag{10}$$

, where N represents the number of samples in the testing set. **Coverage-Efficiency Harmonic Index (CEHI)**: computes the harmonic mean of ECR and APSS to provide a balanced evaluation

of coverage and efficiency. It is defined as the F-measure combining Empirical Coverage Rate and the inverse of Average Prediction Set Size:

$$CEHI = 2 * \frac{ECR * (1/APSS)}{ECR + (1/APSS)} \quad (11)$$

B.4 GPUs

We utilize one NVIDIA A40 graphics cards to support experiments of LLaMA3-8B, and for the needs to load a transformer-based model like text encoder.

B.5 Implementation details proxy attributes extraction

We extracted 5 common attributes: Positive/Negative framing, Imbalanced perspective, Loaded language, Ideological labeling, Speculative attribution, inspired by past research (Spinde et al., 2023). As for LM-based proxy attributes, we input the prompt "List 5 key indicators to check for when detecting media bias in news writing. Please keep the list concise and focus on the most common types of bias." and adopt the suggestion from GPT-4.1-mini. Hence we get the intermediate proxies: Framing Slant, Loaded Language, Sensationalism / Hyperbole, Opinion as Fact, Unsubstantiated Claims from GPT-4.1-mini's response. We can see there are a few proxy overlaps between the prior work and LLM-suggested prompts, e.g., Loaded Language, Framing Slant. We then prompt again to GPT-4 to remove the semantically overlapped prompts and form the Joint proxy attributes.

B.6 Implementation details of Linguistic-Verbal Uncertainty and Frequency-based Conformal Probabilities

The Composite Uncertainty comprises LVU, FCP, and JCSD. To implement LVU, we use the hedging prompt in Figure 10 to elicit intermediate output Z. Z presents as intermediate reasoning output incorporated with a hedging word. Then we use the prompt shown in Figure 11 to turn to Linguistic-Verbal-Confidence score denote as LVC in the range [1,100] and then normalized to [0,1]. Finally, we map the confidence score to the uncertainty score by $LVU = 1 - LVC$. To implement FCP, we use the prompt mentioned in Figure 5 with varied intermediate proxy, instruction, and formatting examples depending on the experiment. Given a decided prompt, we run 20 rounds in order to compute FCP; that is, we set $m = 20$.

The structure of prompt design

You are a media analyst detecting bias in news writing. A sentence may be biased if it shows any of the following:

[intermediate proxy or CoT reasoning]

[instruction]

[formating example]

Sentences to evaluate:

[input sentences] ← (presented as a numbered list: 0. ..., 1. ..., 2. ...)

Answer (analyze step by step, then finalize list):

Figure 5: Prompt structure in this study

Prompt for Gpt4o, Gpt4.1-mini, CU, and LoFreeCP

You are a media analyst detecting bias in news writing. A biased sentence with political bias refers to content—whether supportive, critical, or suggestive—that sways readers' opinions toward an ideological direction.

For each sentence:
* Briefly explain why you think its biased.
* Decide whether the sentence is biased (Yes/No).
* The output should be formated as the following example:
```
31. Sentence: "Democratic campaign officials accused Republicans...
- Explanation: it's a straightforward report
- Decision: No
32. Sentence: "Trump...
- Explanation: because it highlights Trump's change in
- Decision: Yes
```
At the end, output only the list of indices of biased sentences in this format:
[[index1, index2, index3]]

Sentences to evaluate:
[input sentences] ← (presented as a numbered list: 0. ..., 1. ..., 2. ...)

Answer :

Figure 6: Prompt for Gpt4o, Gpt4.1-mini, +CU, and LoFreeCP in this study

C Prompt used in this study

The prompt used in this study shares general information like Bias detection introduction "You are a media analyst detecting bias in news writing. A sentence may be biased if it shows any of the following:.". The intermediate proxy or CoT reasoning, instruction, formating example varies for methods: UAPR-Human, UAPR-LLM, UAPR-Joint, and UAPR-CoT. The details of these three template are elaborated below.

C.1 Fundemental prompt example

We denote fundemental prompt as baseline prompt with no intermediate reasoning step, shown in Figure 6. The instruction for this prompt is simply stated as

```
For each sentence:
* Briefly explain why you think it's biased.
* Decide whether the sentence is biased (Yes/No).
```

In addition, to make it easier to extract the intermediate responses (Z,$\hat{Y}$) we also prompt the LLM to follow the output format:

```
* The output should be formatted as the following example:

31. Sentence: "Democratic campaign officials accused Republicans..."
- Explanation: it's a straightforward report
- Decision: No
32. Sentence: "Trump..."
- Explanation: because it highlights Trump's change in
- Decision: Yes
```

The fundamental prompt from is used in the methods: Gpt4o, Gpt4.1-mini, +CU, and LoFreeCP. The prompts with proxy attributes or chain-of-thought are extended from the fundamental prompt. Figure 5 shows the prompt structure.

C.2 Prompts for UAPR-Human, UAPR-LLM, UAPR-Joint

The prompts for UAPR-Human, UAPR-LLM, UAPR-Joint extends from the fundemental prompt by including intermediate proxy, instruction, and formatting examples. Figure 7, 8, and 9 shows the intermediate proxy and formatting examples of

Intermediate proxy or CoT reasoning for UAPR-Human

1. **Loaded language** – emotionally charged or judgmental words (e.g., "public menace", "rage")
2. **Positive/Negative framing** – portraying someone favorably or unfavorably via comparison
3. **Ideological labeling** – labeling with partisan or ideological terms
4. **Speculative attribution** – suggesting hidden motives without evidence
5. **Imbalanced perspective** – only presenting one side of the issue

Formatting example for UAPR-Human

* The output should be formated as the following example:

```
31. Sentence: "Democratic campaign officials accused Republicans...
- Loaded language: Words like "accused" and "exploiting" carry a ...
- Positive/Negative framing: The sentence presents Democrats' ...
- Ideological labeling: The sentence explicitly identifies ...
- Speculative attribution: The claim that Republicans are ...
- Imbalanced perspective: The sentence only presents ...
```

Figure 7: intermediate proxy and formatting examples of UAPR-Human

intermediate proxy or CoT reasoning for UAPR-LLM

1. **Loaded Language**: Use of emotionally charged words (e.g., "disastrous," "heroic").
2. **Opinion as Fact**: Presenting a subjective belief as an objective truth.
3. **Framing/Slant**: Selecting details or words to favor one perspective.
4. **Sensationalism/Hyperbole**: Exaggeration to provoke a strong reaction.
5. **Unsubstantiated Claim**: Making a statement without providing evidence.

Formatting example for UAPR-LLM

* The output should be formated as the following example:

```
31. Sentence: "Democratic campaign officials accused Republicans...
- Loaded language: Words like "accused" and "exploiting" carry a ...
- Opinion as Fact: The sentence presents Democrats' ...
- Framing/Slant: The sentence explicitly identifies ...
- Sensationalism/Hyperbole: The claim that Republicans are ...
- Unsubstantiated Claim: The sentence only presents ...
```

Figure 8: intermediate proxy and formatting examples of UAPR-LLM

intermediate proxy or CoT reasoning for UAPR-Joint

1. **Loaded Language**: Use of emotionally charged words (e.g., "disastrous," "heroic").
2. **Positive/Negative framing** – portraying someone favorably or unfavorably via comparison
3. **Ideological labeling** – labeling with partisan or ideological terms
4. **Speculative attribution** – suggesting hidden motives without evidence
5. **Imbalanced perspective** – only presenting one side of the issue
6. **Opinion as Fact**: Presenting a subjective belief as an objective truth.
7. **Sensationalism/Hyperbole**: Exaggeration to provoke a strong reaction.
8. **Unsubstantiated Claim**: Making a statement without providing evidence.

Formatting example for UAPR-Joint

* The output should be formated as the following example:

```
31. Sentence: "Democratic campaign officials accused Republicans...
- Loaded language: Words like "accused" and "exploiting" carry a ...
- Positive/Negative framing: The sentence presents Democrats' ...
- Ideological labeling: The sentence explicitly identifies ...
- Speculative attribution: The claim that Republicans are ...
- Imbalanced perspective: The sentence only presents ...
- Opinion as Fact: The sentence presents Democrats' ...
- Sensationalism/Hyperbole: The claim that Republicans are ...
- Unsubstantiated Claim: The sentence only presents ...
```
```

Figure 9: intermediate proxy and formatting examples of UAPR-joint
```

You are a media analyst detecting bias in news writing. A sentence may be biased if it shows any of the following:

[intermediate proxy or CoT reasoning]

Instructions:
For each sentence:
1. Copy the full sentence exactly as shown.
2. Repeat the sentence index and sentence text in the output.
3. Classify the sentence as biased: "Yes" or "No"
5. Start your explanation with a **hedging phrase** (choose one from this exact list):

"almost certain", "highly likely", "very likely", "probable", "somewhat likely", "possible", "uncertain", "somewhat unlikely", "unlikely", "not likely", "doubtful", "very unlikely", "highly unlikely"

Each output block **must strictly follow this format**:

Sentence {{i}}: {{original sentence}}
Bias Decision: [Yes or No]
Explanation: [Start with one hedging word, then brief justification]

Repeat this block for each sentence, with no missing fields and no extra commentary.

Sentences to evaluate:

[input sentences] ← (presented as a numbered list: 0. ..., 1. ..., 2. ...)

Answer (analyze step by step, then finalize list):

Figure 10: The prompt template for uncertainty description elicitation

Instruction for estimating LVU

Evaluate the certainty/probability for each of the following LLM bias decisions. For each numbered statement, provide only the probability score (0-100) that the statement is true, considering the use of hedging words and confidence indicators in the explanations.

{statements}

Respond with only the probability scores in the following format:
0: [score]
1: [score]
2: [score]
...
{max_index}: [score]

Figure 11: The prompt template for uncertainty score evaluation

UAPR-Human, UAPR-LLM, UAPR-Joint.

C.3 Prompts for estimating LVU

Prompts for estimating LVU involves a prompt to elicit uncerainty description and a prompt to quantify the uncertainty description to an uncertainty score. The prompt to elicit uncerainty description extends from the fundamental prompt, with instructions to use a hedging phrase to express the uncertainty about the answer. Figure 10. The LLM responses then provide responses like:
`Very likely unbiased; it states facts about groups...`
Then the linguistic-verbal evaluation prompt in Figure 11 was used to turn the uncertainty description into scores.

Model	BASIL					BABE					BIASEDSENT				
	P	R	F1	$F1_{Macro}$	$F1_{Micro}$	P	R	F1	$F1_{Macro}$	$F1_{Micro}$	P	R	F1	$F1_{Macro}$	$F1_{Micro}$
Gpt4.1-mini	0.40	0.61	0.48	0.65	0.74	0.65	0.84	0.74	0.74	0.74	0.39	0.35	0.37	0.53	0.58
+CoT	0.43	0.53	0.48	0.66	0.77	0.73	0.80	**0.77**	0.78	0.78	0.42	0.24	0.31	0.52	0.59
+CU	0.41	0.61	0.49	0.66	0.74	0.69	0.84	0.76	0.76	0.76	0.38	0.41	**0.40**	0.53	0.59
UAPR-Human	0.40	0.69	**0.51**	0.66	0.73	0.70	0.86	**0.77**	0.77	0.77	0.38	0.41	**0.40**	0.55	0.60
UAPR-LLM	0.46	0.36	0.40	0.64	0.78	0.73	0.83	**0.77**	0.78	0.78	0.39	0.27	0.32	0.52	0.60
UAPR-joint	0.41	0.59	0.48	0.66	0.75	0.68	0.87	0.76	0.75	0.75	0.39	0.37	0.38	0.53	0.58
UAPR-CoT	0.45	0.47	0.46	0.66	0.78	0.73	0.80	0.76	0.76	0.76	0.40	0.23	0.29	0.51	0.61

Table 6: Performance comparison of different models across BASIL, BABE, and BIASEDSENT datasets. Metrics shown are Precision (P), Recall (R), F1 score, F1 Macro, and F1 Micro.

Model	BASIL				BABE				BIASEDSENT			
	ECR ↑	APSS ↓	SSC ↑	CEHI ↑	ECR ↑	APSS ↓	SSC ↑	CEHI ↑	ECR ↑	APSS ↓	SSC ↑	CEHI ↑
Gpt4.1-mini		0.808	1.169	0.768	0.819	✗	✗	✗	✗	✗	✗	✗
✗												
+CoT	0.889	1.257	0.851	0.810	0.847	1.191	0.811	0.828	✗	✗	✗	✗
+CU	0.802	1.132	0.771	0.834	0.820	1.140	0.790	0.840	0.840	1.570	0.630	0.570
UAPR-Human	0.810	1.186	0.766	0.812	0.810	1.160	0.770	0.820	0.800	1.480	0.610	**0.630**
UAPR-LLM	0.806	1.061	0.793	**0.867**	0.810	1.100	0.790	**0.850**	0.810	1.520	0.570	0.600
UAPR-Joint	0.820	1.171	0.783	0.825	0.810	1.220	0.750	0.790	0.810	1.560	0.570	0.570
UAPR-CoT	0.830	1.109	0.809	0.859	0.820	1.210	0.760	0.800	0.830	1.590	0.590	0.550

Table 7: Comprehensive evaluation of models for α=0.2. Panel (a) shows standard performance metrics (Precision, Recall, F1). Panel (b) shows uncertainty quantification metrics (ECR, APSS, SSC, CEHI). The cross symbol 7 denotes that the method fails to produce the set with the desired error rate.

D AdditionalResults

D.1 Generalization to other datasets

This section presents the results of decision accuracy and quality of our UAPR-* methods on three media bias detection datasets: BASIL (Fan et al., 2019), BABE (Spinde et al., 2021), and BIASEDSENT (Lim et al., 2020).

D.1.1 prediction accuracy of UAPR

Table 6 presents the detailed decision accuracy of UAPR across three datasets. The results indicate that using human-knowledge-based proxies consistently achieves the highest F1 score among the three datasets. This finding implies that knowledge-based proxies provide validity, interpretability, and alignment with established domain literature.

D.1.2 Decision quality of UAPR

Table 7 presents the decision quality across three datasets. Observations from the BASIL dataset suggest that LLM-based proxies contribute more to enhancing decision quality, a trend that also applies to the BABE and BIASEDSENT datasets. Notably, in BIASEDSENT, the human-knowledge proxy set achieves the highest CEHI score. This result implies that UAPR-human demonstrates superior decision accuracy and quality. Perhaps the crowd-sourced media bias annotations in BIASEDSENT contain more bias cues captured by human-knowledge-defined proxy attributes, leading to improved prediction quality.

D.2 Sensitivity of hyperparameter λ

Table 12 is ablation study of hyperparameter λ in equation 2. We can see both decision accuracy and quality are not sensitive to the hyperparameter λ.

D.3 Full result of uncertainty measurement for BASIL dataset

Table 8, 9, 10, 11, and 12 show the full uncertainty result benchmark, an extended version of Table. 4, for BASIL dataset. In this experiment, we demonstrate various uncertainty scoring approaches to estimate uncertainty in different statistical aspects:

- NCU: numerical verbal uncertainty (Tian et al., 2023).
- LoFreeCP (FCP+EN+SD): method from (Su et al., 2024).

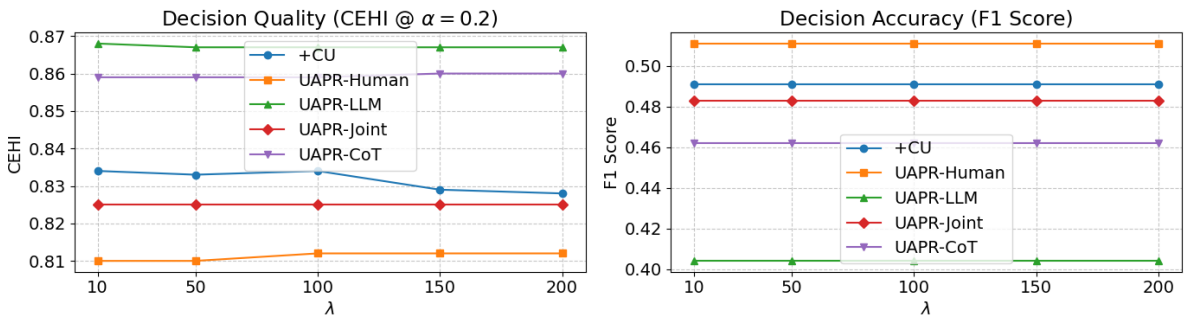


Figure 12: The ablation study of hyperparameter λ in equation 2. In this figure we report CEFI for decision quality and F1 score for Decision accuracy. The target error rate α is 0.2 for the decision quality metric

- LVU: Linguistic-Verbal Uncertainty (Tao et al., 2025).
- FCP: Frequency-based Conformal Probabilities from (Su et al., 2024).
- JCSD: Joint Class Semantic Distance.

, where these uncertainty approaches are tried under multiple settings: +CU, UAPR-CoT, UAPR-Human, UAPR-LLM, UAPR-Joint.

Table 8: Uncertainty result benchmark for media bias detection of +CU

Bias Type	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.81	1.17	0.77	0.82	0.76	1.05	0.74	0.84	0.73	1.0	0.73	0.85	0.73	1.0	0.73	0.85
FCP	0.81	1.17	0.77	0.82	0.76	1.05	0.74	0.84	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
JCSD	0.95	1.80	0.74	0.33	0.94	1.76	0.73	0.38	0.94	1.76	0.73	0.38	0.94	1.76	0.73	0.38
NCU	0.89	1.22	0.86	0.83	0.84	1.08	0.83	0.88	0.84	1.08	0.83	0.88	0.84	1.08	0.83	0.88
LVU	0.86	1.27	0.81	0.79	0.82	1.16	0.78	0.83	0.81	1.13	0.78	0.84	0.81	1.13	0.78	0.84
FCP + JCSD	0.80	1.16	0.77	0.82	0.75	1.04	0.74	0.84	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
LVU + FCP	0.81	1.14	0.78	0.83	0.76	1.04	0.75	0.85	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
LVU + JCSD	0.85	1.24	0.80	0.80	0.82	1.16	0.78	0.83	0.80	1.12	0.77	0.84	0.79	1.10	0.77	0.84
CU	0.80	1.13	0.77	0.83	0.76	1.03	0.75	0.85	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85

Bias Type	$\alpha = 0.4$				$\alpha = 0.45$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.73	1.0	0.73	0.85	0.73	1.0	0.73	0.85
FCP	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
JCSD	0.94	1.76	0.73	0.38	0.94	1.76	0.73	0.38
NCU	0.82	1.02	0.81	0.89	0.82	1.02	0.81	0.89
LVU	0.80	1.12	0.78	0.84	0.78	1.05	0.77	0.86
FCP + JCSD	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
LVU + FCP	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85
LVU + JCSD	0.79	1.10	0.77	0.84	0.78	1.04	0.77	0.86
CU	0.74	1.00	0.74	0.85	0.74	1.00	0.74	0.85

Table 9: Uncertainty result benchmark for media bias detection of UAPR-CoT

Bias Type	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.8	1.17	0.76	0.82	0.75	1.06	0.74	0.84	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
FCP	0.8	1.08	0.79	0.86	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87
JCSD	0.94	1.8	0.71	0.33	0.93	1.75	0.7	0.39	0.91	1.7	0.7	0.45	0.9	1.67	0.69	0.48
NCU	0.84	1.3	0.77	0.76	0.84	1.29	0.77	0.77	0.84	1.29	0.77	0.77	0.84	1.29	0.77	0.77
LVU	0.85	1.37	0.76	0.72	0.84	1.35	0.75	0.73	0.82	1.29	0.74	0.76	0.81	1.28	0.74	0.76
FCP + JCSD	0.8	1.07	0.78	0.86	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87
LVU + FCP	0.83	1.13	0.81	0.85	0.8	1.06	0.79	0.87	0.78	1.0	0.78	0.87	0.78	1.0	0.78	0.87
LVU + JCSD	0.85	1.36	0.76	0.73	0.84	1.34	0.75	0.74	0.81	1.28	0.74	0.76	0.8	1.25	0.74	0.77
CU	0.83	1.13	0.81	0.85	0.8	1.06	0.79	0.87	0.78	1.0	0.78	0.87	0.78	1.0	0.78	0.87

Bias Type	$\alpha = 0.4$				$\alpha = 0.45$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
FCP	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87
JCSD	0.9	1.67	0.69	0.48	0.9	1.67	0.69	0.48
NCU	0.8	1.14	0.76	0.83	0.8	1.14	0.76	0.83
LVU	0.79	1.21	0.73	0.79	0.79	1.21	0.73	0.79
FCP + JCSD	0.77	1.0	0.77	0.87	0.77	1.0	0.77	0.87
LVU + FCP	0.78	1.0	0.78	0.87	0.78	1.0	0.78	0.87
LVU + JCSD	0.79	1.21	0.73	0.79	0.78	1.19	0.73	0.8
CU	0.78	1.0	0.78	0.87	0.78	1.0	0.78	0.87

Table 10: Uncertainty result benchmark for media bias detection of UAPR-Human

Bias Type	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.81	1.22	0.75	0.79	0.75	1.1	0.73	0.82	0.71	1.01	0.71	0.83	0.71	1.0	0.71	0.83
FCP	0.8	1.21	0.75	0.79	0.75	1.1	0.73	0.82	0.71	1.01	0.71	0.83	0.71	1.0	0.71	0.83
JCSD	0.97	1.8	0.85	0.33	0.96	1.75	0.85	0.4	0.95	1.7	0.84	0.46	0.94	1.65	0.83	0.51
NCU	0.87	1.51	0.73	0.63	0.87	1.51	0.73	0.63	0.87	1.51	0.73	0.63	0.84	1.41	0.74	0.7
LVU	0.87	1.48	0.74	0.65	0.87	1.48	0.74	0.65	0.82	1.37	0.72	0.71	0.82	1.37	0.72	0.71
FCP + JCSD	0.8	1.2	0.75	0.8	0.75	1.09	0.73	0.82	0.71	1.0	0.71	0.83	0.71	1.0	0.71	0.83
LVU + FCP	0.82	1.2	0.77	0.81	0.79	1.13	0.76	0.83	0.75	1.04	0.74	0.84	0.73	1.0	0.73	0.84
LVU + JCSD	0.86	1.45	0.74	0.67	0.84	1.4	0.73	0.7	0.82	1.35	0.72	0.72	0.81	1.32	0.72	0.74
CU	0.81	1.19	0.77	0.81	0.78	1.12	0.76	0.83	0.75	1.04	0.74	0.84	0.73	1.0	0.73	0.84

Bias Type	$\alpha = 0.4$				$\alpha = 0.45$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.71	1.0	0.71	0.83	0.71	1.0	0.71	0.83
FCP	0.71	1.0	0.71	0.83	0.71	1.0	0.71	0.83
JCSD	0.93	1.6	0.83	0.56	0.92	1.55	0.82	0.6
NCU	0.82	1.31	0.74	0.75	0.82	1.31	0.74	0.75
LVU	0.82	1.37	0.72	0.71	0.82	1.37	0.72	0.71
FCP + JCSD	0.71	1.0	0.71	0.83	0.71	1.0	0.71	0.83
LVU + FCP	0.73	1.0	0.73	0.84	0.73	1.0	0.73	0.84
LVU + JCSD	0.8	1.28	0.71	0.75	0.79	1.25	0.71	0.77
CU	0.73	1.0	0.73	0.84	0.73	1.0	0.73	0.84

Table 11: Uncertainty result benchmark for media bias detection of UAPR-LLM

Bias Type	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.8	1.11	0.77	0.84	0.76	1.02	0.75	0.85	0.75	1.0	0.75	0.85	0.75	1.0	0.75	0.85
FCP	0.8	1.12	0.78	0.84	0.75	1.01	0.75	0.85	0.75	1.0	0.75	0.86	0.75	1.0	0.75	0.86
JCSD	0.97	1.8	0.84	0.33	0.96	1.75	0.83	0.4	0.95	1.7	0.82	0.46	0.93	1.65	0.81	0.51
NCU	0.82	1.1	0.8	0.86	0.81	1.05	0.8	0.87	0.8	1.04	0.8	0.87	0.8	1.04	0.8	0.87
LVU	0.82	1.1	0.8	0.86	0.8	1.05	0.8	0.87	0.8	1.04	0.79	0.87	0.8	1.04	0.79	0.87
FCP + JCSD	0.8	1.11	0.77	0.84	0.75	1.01	0.75	0.85	0.75	1.0	0.75	0.86	0.75	1.0	0.75	0.86
LVU + FCP	0.82	1.09	0.8	0.86	0.79	1.01	0.79	0.88	0.79	1.01	0.79	0.88	0.78	1.0	0.78	0.88
LVU + JCSD	0.81	1.07	0.8	0.87	0.8	1.04	0.8	0.87	0.8	1.04	0.79	0.87	0.8	1.03	0.79	0.88
CU	0.81	1.06	0.79	0.87	0.79	1.01	0.79	0.88	0.79	1.01	0.79	0.88	0.78	1.0	0.78	0.88

Bias Type	$\alpha = 0.4$				$\alpha = 0.45$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.75	1.0	0.75	0.85	0.75	1.0	0.75	0.85
FCP	0.75	1.0	0.75	0.86	0.75	1.0	0.75	0.86
JCSD	0.92	1.6	0.81	0.56	0.91	1.55	0.8	0.6
NCU	0.8	1.04	0.8	0.87	0.8	1.02	0.79	0.88
LVU	0.8	1.03	0.79	0.88	0.8	1.03	0.79	0.88
FCP + JCSD	0.75	1.0	0.75	0.86	0.75	1.0	0.75	0.86
LVU + FCP	0.78	1.0	0.78	0.88	0.78	1.0	0.78	0.88
LVU + JCSD	0.8	1.03	0.79	0.88	0.79	1.02	0.79	0.88
CU	0.78	1.0	0.78	0.88	0.78	1.0	0.78	0.88

Table 12: Uncertainty result benchmark for media bias detection of UAPR-Joint

Bias Type	$\alpha = 0.2$				$\alpha = 0.25$				$\alpha = 0.3$				$\alpha = 0.35$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.8	1.17	0.76	0.82	0.75	1.06	0.74	0.84	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
FCP	0.8	1.16	0.77	0.82	0.75	1.06	0.74	0.84	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
JCSD	0.97	1.8	0.85	0.33	0.96	1.75	0.85	0.4	0.95	1.7	0.84	0.46	0.94	1.65	0.83	0.51
NCU	0.86	1.43	0.75	0.68	0.86	1.43	0.75	0.68	0.86	1.43	0.75	0.68	0.83	1.33	0.75	0.74
LVU	0.86	1.44	0.75	0.68	0.84	1.39	0.74	0.71	0.82	1.35	0.73	0.73	0.81	1.32	0.73	0.74
FCP + JCSD	0.8	1.16	0.76	0.82	0.75	1.06	0.74	0.84	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
LVU + FCP	0.82	1.18	0.78	0.82	0.8	1.11	0.77	0.84	0.76	1.03	0.76	0.85	0.75	1.0	0.75	0.85
LVU + JCSD	0.85	1.42	0.74	0.69	0.83	1.37	0.73	0.72	0.81	1.32	0.73	0.74	0.8	1.29	0.72	0.75
CU	0.82	1.17	0.78	0.82	0.8	1.1	0.77	0.84	0.76	1.03	0.76	0.85	0.75	1.0	0.75	0.85

Bias Type	$\alpha = 0.4$				$\alpha = 0.45$			
	ECR↑	APSS↓	SSC↑	CEHI↑	ECR↑	APSS↓	SSC↑	CEHI↑
LoFreeCP	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
FCP	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
JCSD	0.93	1.6	0.83	0.56	0.92	1.55	0.82	0.6
NCU	0.81	1.26	0.75	0.77	0.81	1.26	0.75	0.77
LVU	0.81	1.32	0.73	0.74	0.81	1.32	0.73	0.74
FCP + JCSD	0.72	1.0	0.72	0.84	0.72	1.0	0.72	0.84
LVU + FCP	0.75	1.0	0.75	0.85	0.75	1.0	0.75	0.85
LVU + JCSD	0.79	1.26	0.72	0.76	0.78	1.23	0.71	0.77
CU	0.75	1.0	0.75	0.85	0.75	1.0	0.75	0.85

E Case studies

This section aligns with section 4.4 that discusses the entities with high CU scores. Since Duncan Hunter has high CU samples mostly come from high FCP; Obama Administration has a moderate proportion of high CU samples with high LVU, we provide case studies for further investigation. We first investigate the high CU samples caused by LVU. Figure 13, 14, 15 show sample in top-ranked CU group in Obama Administration. Hedging words like "probable" and "Somewhat" usually appears when the LLM is judging on sentences relate to the Obama Administration. Despite most of the time the FCP is low, the model expresses its uncertainty by generating uncertain linguistic-verbal expression.

Figures 16, 17, 18 show the detail of LLM's intermediate reasons when making decision on the high FCP samples. We found prediction outcomes are not consistent in multi-round prompting. Take case 3 in Figure 18 as an example, LLM predicts 10 times bias and 10 times unbias. Moreover, the LLM gives different reasoning traces. When it predicts bias, it says the sentence frames the speaker positively, while it thinks the framing attributes to the entity (Duncan Hunter) instead of the journalist.

Case 1

Media Outlet: fox
Entity: Obama administration
sentence: In addition to identifying the presence of Al Qaeda and Ansar al-Sharia in Benghazi, the groups ultimately blamed for the assault, a key section of the cable, which remains classified, seemed to foreshadow the attack itself.

LLM's reasoning through LVU: Probable bias due to speculative attribution and ideological labeling. The phrase "seemed to foreshadow" suggests a speculative link without definitive proof. Also, naming groups as blameworthy incorporates ideological labeling that can influence perception.

score: [LVU: 0.2500, FCP: 0.2500, JCSD: 0.0007, CU: 0.5011]
Model prediction: bias, label: bias

Figure 13: case number 1 for sample in top-ranked CU group in Obama Administration.

Case 2

Media Outlet: fox
Entity: Obama administration
sentence: Clinton said she had no knowledge of the warning which former diplomatic security agents have told Fox News amounted to a "cry for help."

LLM's reasoning through LVU: Somewhat likely biased due to loaded language and imbalanced perspective. The phrase "cry for help" is emotionally charged, and the sentence presents only the critical perspective of former agents without alternative views.

score: [LVU: 0.4000, FCP: 0.0000, JCSD: 0.0074, CU: 0.4011]
Model prediction: bias, label: bias

Figure 14: case number 2 for sample in top-ranked CU group in Obama Administration.

Case 3

Media Outlet: fox
Entity: Obama administration
sentence: Graham said: "She said she didn't. Are you stunned that she didn't?"

LLM's reasoning through LVU: Somewhat likely biased due to loaded language ("Are you stunned") that implies criticism and negative framing of Clinton's knowledge, suggesting surprise or incompetence.

score: [LVU: 0.4000, FCP: 0.0000, JCSD: 0.0003, CU: 0.4000]
Model prediction: bias, label: unbias

Figure 15: case number 3 for sample in top-ranked CU group in Obama Administration.

Case1	
Media Outlet: hpo **Entity:** Duncan Hunter **sentence:** Hunter, 41, is a five-term congressman representing San Diego County who served for three tours in Afghanistan and Iraq as a Marine. **score:** [LVU: 0.0500, FCP: 0.8000, JCSD: 0.0075, CU: 0.8507] **Label:** unbias	
Reasons that predict **unbias** Frequency: 4/20	Reasons that predict **bias** Frequency: 16/20
- Loaded_language_description: Neutral, positive background info. - Positive/Negative_framing_description: Presents Hunter's military service positively. - Ideological_labeling_description: No. - Speculative_attribution_description: No. - Imbalanced_perspective_description: Provides positive background but balanced by indictment info elsewhere.	- Loaded_language_description: None; factual description. - Positive/Negative_framing_description: Positive framing by highlighting military service. - Ideological_labeling_description: None. - Speculative_attribution_description: None. - Imbalanced_perspective_description: Positive attributes are presented without negative context here.

Figure 16: case number 1 for sample in top-ranked CU group in Duncan Hunter.

Case2	
Media Outlet: hpo **Entity:** Duncan Hunter **sentence:** In a speech earlier this month, Hunter echoed President Donald Trump and castigated the Justice Department, accusing the agency of being "corrupt, answerable to no one" and said it "uses the law to extort the American people and effect political change." **score:** [LVU: 0.1500, FCP: 0.7500, JCSD: 0.0045, CU: 0.9003] **Label:** bias	
Reasons that predict **unbias** Frequency: 15/20	Reasons that predict **bias** Frequency: 5/20
- Loaded_language_description: The phrase "pretty well-liked" is mildly positive but is a direct quote. - Positive/Negative_framing_description: Presents Republicans in a positive light, but as this is a quote from one person, not editorializing. - Ideological_labeling_description: Identifies Republicans by party, but that is factual, not ideological labeling. - Speculative_attribution_description: No attribution of hidden motives. - Imbalanced_perspective_description: Only this positive view is presented here; however, it is a direct quote.	- Loaded_language_description: Words like "castigated," "corrupt," "extort" are strong negative language. - Positive/Negative_framing_description: Negative framing of DOJ. - Ideological_labeling_description: Links Hunter to Trump, possibly ideological. - Speculative_attribution_description: Hunter's accusations are stated but no evidence provided. - Imbalanced_perspective_description: Presents only Hunter's accusations without counter.

Figure 17: case number 2 for sample in top-ranked CU group in Duncan Hunter.

Experiment	3 Trials Cost	20 Trials Cost	20 Trials Seq Time	20 Trials Par Time
GPT-4.1-mini	$2.44	$2.44	5.3 h	5.3 h
+CU	$5.21	$34.75	70.0 h	43.9 h
UAPR-CoT	$5.54	$36.91	78.9 h	49.6 h
UAPR-Human	$6.91	$46.08	91.7 h	54.8 h

Table 13: Comparison of experiment costs and runtimes.

F Inference time and cost of UAPR method and baseline comparison

Table 13

G The use of Ai in this paper

We use Ai for python code generation, and to assist revising this paper. Using Ai generates more robust code avoiding false result due to coding error.

Case3	
Media Outlet: hpo Entity: Duncan Hunter sentence: He said polling had him safely above his opponent — “We’re looking good” — and he indicated he wasn’t really afraid of losing. score: [LVU: 0.1500, FCP: 0.5000, JCSD: 0.0010, CU: 0.6500] Label: unbias	
Reasons that predict unbias Frequency: 10/20	Reasons that predict bias Frequency: 10/20
- Loaded_language_description: The phrase "We’re looking good" is a positive sentiment but is a direct quote reflecting confidence, not necessarily loaded language. - Positive/Negative_framing_description: The sentence frames the subject as confident and ahead in the race, which could be seen as positive framing, but it is attributed directly to him. - Ideological_labeling_description: None. - Speculative_attribution_description: No suggestion of hidden motives. - Imbalanced_perspective_description: Only the subject’s optimistic view is presented; no opposing viewpoint on the poll or race is given.	- Loaded_language_description: Phrases like “safely above” and “We’re looking good” are positive but come from the subject’s own words. - Positive/Negative_framing_description:The sentence frames the speaker positively by highlighting confidence and a lead in the polls. - Ideological_labeling_description: None. - Speculative_attribution_description: No claims about hidden motives; it reports confidence expressed by the subject. - Imbalanced_perspective_description: Only the subject’s optimistic perspective is presented; no counterpoint is given.

Figure 18: case number 3 for sample in top-ranked CU group in Duncan Hunter.

H Responses to Reviewer Concerns

This section addresses key methodological questions raised during peer review.

Q1: LVU is class-agnostic while FCP and JCSD are class-specific. Why combine fundamentally different uncertainty types via simple weighted summation, and how are scale incompatibilities handled?

This characterization is a slight misreading of our formulation. LVU, FCP, and JCSD are all designed as *sentence-level* uncertainty signals, not class-specific ones. FCP is computed by sampling m inference rounds and measuring how frequently each class is selected; JCSD measures the semantic consistency of the intermediate reasoning *language* across those same m samples. Neither is a per-class score that is then aggregated—both are scalar summaries of prediction variability across rounds.

Regarding scale incompatibility: we explicitly acknowledge that the three components operate on different numeric ranges. We address this by introducing the hyperparameter λ in Equation 2, which aligns the magnitude of JCSD with those of FCP and LVU. A grid search over λ (Appendix D.2) shows that performance is largely insensitive to the exact value around $\lambda = 100$, confirming that the choice is principled rather than arbitrary.

Q2: The JCSD multiplicative product is poorly motivated. If each class has low intra-class variance but predictions are uniformly distributed (high inter-class uncertainty), JCSD would be misleadingly small.

The intuition behind JCSD is rooted in the causal relationship between a decision and its underlying reasoning. If an LLM arrives at a specific decision with zero reasoning variance, while the alternative decision's reasoning is extremely varied and inconsistent, the former decision should be prioritized because its rationale is definite. Our multiplicative formulation handles this naturally: if any class achieves near-zero reasoning variance, the joint product reflects a state of low uncertainty for that specific reasoning path. In such cases, the final decision (determined by Eq. 5) would lean toward the class with certain reasoning, and the overall JCSD would correctly reflect this.

We acknowledge that the current JCSD formulation has limitations and that future work exploring frequency-weighted aggregation—where each class is weighted by its prediction frequency before multiplication—could better align JCSD with inter-class uncertainty.

Q3: The best result (UAPR-Human) requires human-curated proxies, which undercuts the "automatic" framing of the method.

We agree that the term "automatic" warrants clarification. LLMs excel at automating many NLP tasks, but in specialized, subjective domains such as media bias, *alignment to human judgment remains an unsolved problem*. LLM-generated knowledge reflects the model's pre-training distribution, which may not correspond to the nuanced criteria used by human annotators in a given domain. Providing a small human-authored guideline—a list of domain-specific proxy attributes—requires minimal effort while substantially improving alignment.

We view this as analogous to prompt engineering or instruction tuning: the *inference pipeline* is fully automatic given the proxies, but the proxies themselves encode domain expertise. The contribution is not that all human input is eliminated, but that a structured, transparent reasoning scaffold (proxy attributes) combined with uncertainty estimation improves both accuracy and decision quality over unguided LLM inference. Future work on learning proxies from domain-specific feedback could further reduce the human effort required.

Q4: The conformal prediction exchangeability assumption may be violated by the story-split evaluation scheme.

This is a valid theoretical concern. Standard split conformal prediction requires that calibration and test samples are exchangeable—i.e., drawn from the same underlying distribution. Story-split cross-validation partitions data by article, meaning different folds cover different entities, topics, and language patterns, which could introduce distributional shift between calibration and test sets.

We address this concern from three angles. *Empirically*, our ECR values across all methods and α settings are consistently close to or above the theoretical target $1-\alpha$ (Tables 2 and the full results in Appendix D.3), indicating that the coverage guarantees hold in practice despite the potential shift.

We adopt story-split following established protocol in media bias detection (Lin et al., 2024) to ensure fair comparison with prior work. We acknowledge the exchangeability limitation and note that adaptive conformal prediction methods (Cherian

et al., 2024) designed for distribution shift settings represent a promising direction for future work.

Q5: Ethical considerations are minimal. Media bias detection systems carry societal implications beyond data privacy.

We thank the reviewer for raising this important point. Beyond the absence of new data collection, we identify the following ethical risks associated with deploying media bias detection systems:

Potential for misuse. Automated bias detection tools could be misused for *selective censorship or political manipulation*—for example, systematically flagging content from specific outlets or viewpoints under the guise of bias removal while leaving ideologically aligned content unexamined. The asymmetric application of such tools by governments, platforms, or corporations poses a genuine democratic risk.