Detecting and Reasoning about Bias in Multimodal Content

Shaina Raza[1,*], **Caesar Saleh**[1], **Azib Farooq**[2], **Emrul Hasan**[1], **Franklin Ogidi**[1], **Haad Zahid**[3], **Maximus Powers**[4], **Marcelo Lotif**[1], **Anam Zahid**[5], **Karanpal Sekhon**[1], **Veronica Chatrath**[1], **Vahid Reza Khazaie**[1]

[1]Vector Institute, Toronto, ON, Canada
[2]University of Cincinnati, Cincinnati, OH, USA
[3]LUMS, Lahore, Pakistan
[4]Arizona State University, Arizona, USA
[5]ITU, Lahore, Pakistan
[*]shaina.raza@torontomu.ca

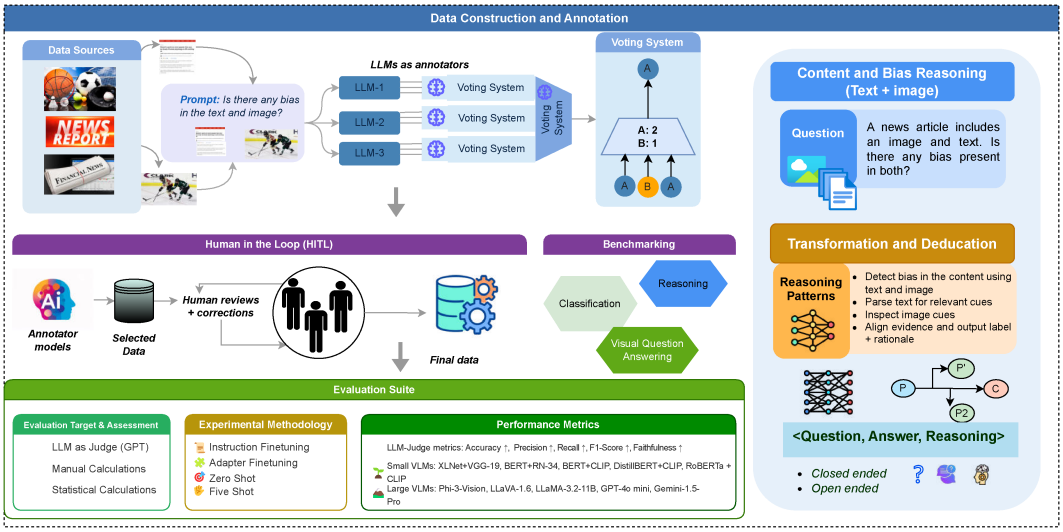

Figure 1: **ViLBias** Framework. The pipeline comprises three main stages: (1) Data Collection and Preprocessing, (2) Automated and Human-Annotated Labeling, and (3) Final Evaluation.

Abstract

Detecting bias in multimodal news requires models that reason over textimage pairs, not just classify text. In response, we present ViLBias, a VQA-style benchmark and framework for detecting and reasoning about bias in multimodal news. The dataset comprises 40,945 textimage pairs from diverse outlets, each annotated with a bias label and concise rationale using a two-stage LLM-as-annotator pipeline with hierarchical majority voting and human-in-the-loop validation. We evaluate Small Language Models (SLMs), Large Language Models (LLMs), and Vision Language Models (VLMs) across closed-ended classification and open-ended reasoning (oVQA), and compare parameter-efficient tuning strategies. Results show that incorporating images alongside text improves detection accuracy by 3-5%, and that LLMs/VLMs better capture subtle framing and text-image inconsistencies than SLMs. Parameter–efficient methods (LoRA/QLoRA/Adapters)

Article presented at the Second IASEAI Conference (IASEAI'26).

recover 97-99% of full fine–tuning performance with <5% trainable parameters. For oVQA, reasoning accuracy spans 52-79% and faithfulness 68-89%, both improved by instruction tuning; closed accuracy correlates strongly with reasoning (r = 0.91). ViLBias offers a scalable benchmark and strong baselines for multimodal bias detection and rationale quality. The data and code is available at [1].

1 Introduction

Recent advances in natural language processing (NLP) and multimodal learning, exemplified by the OpenAI GPT series [26] and the Meta LLaMA series [4], have substantially improved the analysis and interpretation of textual and visual content. However, despite their capacity to process large volumes of information, these systems can also inherit and amplify biases present in both modalities. In multimodal settings, bias extends beyond language to include visual strategies such as image selection, cropping, staging, and emotion-provoking imagery.

In this work, we define *bias* as directional framing in text, image, or their interaction that skews salience or interpretation (e.g., loaded wording, selective emphasis, image selection/cropping/staging, cross-modal mismatch). Our focus is news media, and we distinguish framing bias from misinformation. The topic of multimodal fake news detection has received substantial attention [2], systematic bias detection remains comparatively underexplored, partly due to its inherent subjectivity, which underscores the need for targeted research in this domain.

Existing benchmarks for news and social media bias remain limited, with most focusing on text-only signals. Resources such as MBIB [51], BABE [42], and related corpora [37] emphasize textual bias. Multimodal datasets (e.g., NewsBag [16], MMFakeBench [22], FakeNewsNet [40]) primarily address misinformation rather than *framing* bias. Current research shows three gaps: (i) most benchmarks target fake news/misinformation rather than framing bias; when bias is considered, resources are predominantly *text-only* and multimodal bias benchmarks are scarce; (ii) most resources are *label-only*, lacking rationales or evidence, which precludes evaluation of *reasoning quality*; and (iii) scalable, audited annotation protocols (e.g., LLM-as-annotator with human validation) are uncommon, limiting dataset size and consistency. These gaps motivate benchmarks that integrate textual and visual dimensions of bias and evaluate both *closed-ended* classification and *open-ended* reasoning.

We present VILBIAS, a visual question answering (VQA) style, multimodal benchmark for news media bias. Each textimage pair is combined with a question, a label (bias presence/type), and a concise reasoning (rationale), enabling evaluation of both prediction and reasoning. VILBIAS consists of 40k unique multimodal pairs curated from diverse real-world news sources. Given the scale of the data, annotations are produced through a hybrid strategy that combines LLM-based automated labeling with extensive human review to ensure quality and consistency. The dataset covers multiple bias categories, ranging from overt ideological framing to subtle visual manipulations such as cropping, staging, and emotive imagery. We evaluate a range of state-of-the-art (SoTA) small language models (SLMs)[2], large language models (LLMs) and vision languahe models (VLMs), assessing their ability to answer multimodal bias-related questions. An overview of the VILBIAS framework is shown in Figure 1, and its comparison with state-of-the-art (SOTA) bias-detection benchmarks is presented in Table 1.

Contributions Our work can be summarized into four main contributions:

1. We present VILBIAS, a VQA-style multimodal benchmark for news media bias.
2. We introduce BiasCorpus, a dataset of 40k textimage pairs from diverse news sources, each annotated with bias labels and rationales to frame bias detection as a multimodal task.
3. We design a hybrid annotation pipeline that stabilizes outputs through multi-LLM voting and integrates expert human review to ensure high-quality annotations.
4. We conduct a systematic evaluation of SLMs, LLMs, and VLMs, highlighting their respective strengths and limitations in multimodal media bias detection.

[1] ViLBias Link

[2] We define SLMs as encoder-only pre-ChatGPT architectures [48].

Table 1: Bias and related multimodal (MM) benchmarks. Legend: ✓=yes, ✗=no, ○=partial; Mod.=modality (T=Text, I=Image); Foc.=focus (B=Bias, MI=Misinformation); Ann.=annotation source (LLM, H=Human); Rat.=rationale; MM.=multimodal checks; Reason.=reasoning.

Benchmark	Mod.	Foc.	Ann.	Rat.	MM.	Reason.
BiasLab [41]	T	B	H	✓	—	○
BIASsist [25]	T	B	LLM	✓	—	✓
MDAM3 [53]	T+I	MI	AI+H	✗	○	✗
MMFakeBench [22]	T+I	MI	AI+H	✗	✓	○
IndiTag [20]	T	B	LLM	○	—	○
VERITE [28]	T+I	MI	H	✗	✓	✗
MBIB [51]	T	B	H	○	—	✗
BABE [42]	T	B	H	✓	—	✗
BASIL [11]	T	B	H	✓	—	✗
NewsBag [16]	T+I	MI	H	✗	○	✗
FakeNewsNet [40]	T+I	MI	AI+H	✗	○	✗
VILBIAS (ours)	**T+I**	**B**	**AI+H**	✓	✓	✓

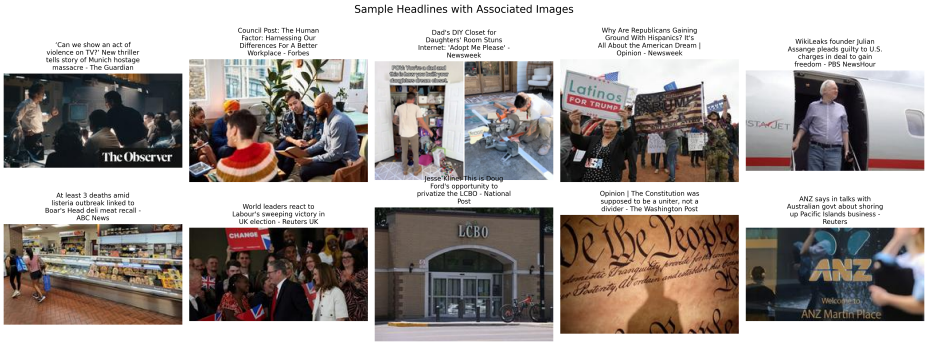

Figure 2: Sample headlines and their associated images. Each image corresponds to a specific news article, illustrating the multimodal structure of the dataset.

2 Related Work

The study of fake news can be categorized into disinformation, misinformation [39], and bias news [32]. Biased news typically presents information with a particular slant, emphasizing certain viewpoints while downplaying others [33], whereas fake news or misinformation consists of entirely fabricated content intended to mislead readers. This work primarily focuses on bias detection in news. Research on bias detection typically uses SLMs for classification and LLMs for open-ended reasoning. Prior work indicates that SLMs can be effective on overt lexical biases when fine-tuned for specific tasks [56], but they struggle with subtler forms such as framing and ideological slant, where LLMs demonstrate stronger generalization and reasoning abilities [29, 52].

Annotation Process Multimodal bias detection suffers from limited labeled data. Early efforts relied on crowd workers [46] or experts [50], with some attempts at AI-generated labels [12]. Manual labeling is costly, slow, and error-prone, prompting interest in AI-assisted approaches [9]. Recent studies use GPT models for annotation [43], leading to hybrid frameworks that combine AI-generated labels with human verification [13], while others rely solely on AI annotations [27].

Multimodal Bias Detection. Bias in news often emerges from textimage interactions, requiring models that capture cross-modal framing [1, 7]. Approaches using CLIP-style alignment or contrastive objectives improve detection but are largely aimed at misinformation rather than bias [31, 8]. Benchmarks reveal pitfalls such as unimodal shortcuts, where text- or image-only models perform well on multimodal tasks [28]. Recent efforts like VLBiasBench expose demographic biases in VLMs, highlighting the need for reasoning-oriented benchmarks [49]. Empirically, visual cues provide modest gains (approx. 35%) beyond text, underscoring the value of multimodal supervision.

In this work, we address these gaps by introducing VILBIAS, a multimodal benchmark with binary labels and reasonings, annotated via an LLM-as-annotator plus human validation pipeline, and we establish baselines for both closed-ended classification and open-ended (VQA-style) models.

3 VILBIAS Methodology

Our proposed framework, **ViLBias**, is structured into three core stages, as illustrated in Figure 1. First, we introduce *BiasCorpus*, which is a systematic collection and preparation of multimodal data relevant to news bias detection. Second, we employ a *Human-In-The-Loop (HITL) BiasAnnotation* process to incorporate two step expert-driven annotations, enhancing the reliability of the dataset. Finally, the curated and annotated data are integrated into a comprehensive benchmarking suite designed to evaluate and analyze model performance under diverse bias scenarios. Together, these components provide a robust and transparent pipeline for studying bias in multiple models.

3.1 Problem Definition

The main goal of this work is to assess the ability of language models, specifically VLMs, to detect and reason about media bias. We introduce BiasCorpus, a dataset usable in both classification (close-ended) and VQA-style (open-ended) settings. In the latter, each textimage pair is paired with a query such as "Does this content exhibit bias? Please explain your reasoning". Formally, given a textual snippet $\mathcal{T}$, a corresponding visual element $\mathcal{V}$, and a bias-related query $\mathcal{Q}$, the task is defined as

$$f:\ (\mathcal{T},\mathcal{V},\mathcal{Q})\ \mapsto\ (\mathcal{A}_{\text{cls}},\mathcal{A}_{\text{reasoning}}),$$

where $\mathcal{A}_{\text{cls}} \in \{\text{biased}, \text{not biased}\}$ is the categorical prediction and $\mathcal{A}_{\text{reasoning}}$ is an open-ended explanation that justifies the decision by pointing to textual framing (e.g., emotive wording, selective omission) and/or visual cues (e.g., cropping, staging, image sentiment). This joint formulation allows evaluation of both predictive accuracy and reasoning quality.

3.2 BiasCorpus Curation

We constructed BIASCORPUS by ingesting articles from Google News RSS and official outlet feeds (May 2023September 2024), de-duplicating by canonical URLs and filtering non-article formats. The sources, spanning CNN, BBC, The New York Times, Al Jazeera, and others as seen in Table 2, were selected to cover varied geographies, editorial orientations, and audience trust profiles [3, 55]. To ensure quality, we removed duplicates, discarded short articles (fewer than twenty sentences), and filtered out multimedia-only items (e.g., videos, slideshows). Each retained article was paired with its first encountered image to establish a consistent multimodal reference. This study introduces a distinct labeling protocol based on multi-LLM annotation with human review for robustness. Sample headlines and images are shown in Figure 3.

3.3 HITL BiasAnnotation

To efficiently annotate a large, complex dataset, we integrated automated annotators, using LLM-as-annotators for bulk labeling of textimage pairs. The use of LLMs for annotation has recently been shown to be an effective and scalable strategy [14]. A distinct aspect of our annotation pipeline is a *two-step voting mechanism* designed to maximize reliability while reducing stochastic variability in LLM outputs. The process, as shown in Figure 3 operates as follows:
Step 1 Stabilizing each LLM: For each model, the same item is queried three times. Since LLMs are inherently stochastic [5], the responses may differ. A majority vote is applied over the three responses to obtain one stable annotation per model.
Step 2 Consensus across models: After Step 1, each LLM contributes a single annotation. With three LLMs in total, a second majority vote is then applied across their outputs, yielding the final annotation for that item.

In short, each LLM first settles on its own answer by voting across three attempts, and then the models collectively agree through another vote. This hierarchical voting scheme ensures robustness both *within* and *across* models, providing reliable annotations.

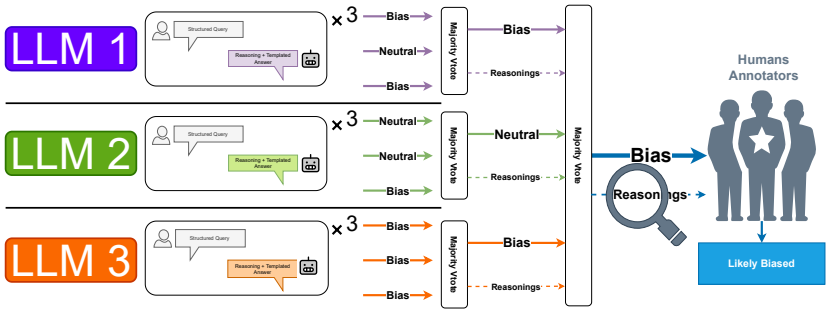

Figure 3: LLM-As-Annotators with Human-In-The-Loop (HITL) Bias Annotation Framework. Each LLM is queried three times, and its responses are majority-voted to remove stochasticity. These majority-voted labels are taken from three LLMs and then majority-voted (across LLMs) again to produce a final label, which is reviewed—along with LLM reasoning—by human annotators.

Human-In-The-Loop Human oversight remained essential for validating automated outputs and refining the annotation pipeline. We engaged 12 reviewers (three undergraduates, four masters students, two Ph.D. holders, and three senior researchers) spanning computer science, media studies, political science, and linguistics. Two subject-matter experts led the process using bias identification protocols, ensuring coverage of both linguistic and visual biases as defined by the U.S. Equal Employment Opportunity Commission [47]. Reviewers assessed all annotations, classifying them as *Acceptable*, *Needs Improvement*, or *Incorrect*, and provided rationales where needed. Their feedback informed iterative prompt and pipeline adjustments, with disagreements resolved through consensus. After refinement, the inter-annotator agreement (IAA) between automated labels and expert judgments, measured with Cohens Kappa, reached 0.72, indicating substantial alignment. Table 2 summarizes dataset schema, key statistics, and observed patterns.

3.4 Benchmarking Suite

We evaluate models on both closed-ended classification and open-ended reasoning tasks across three model families (LLMs, SLMs and VLMs) and two accessibility settings (open-ended and closed-ended). First, we consider open-source SLMs as text-only baselines, including encoder-only BERT-style transformers, CLIP-style joint encoders, decoder-only models such as GPT-2, and encoder–decoder models such as BART-base, all fine-tuned on our dataset. Second, we examine autoregressive decoder-only LLMs, tested under zero-shot, few-shot, and instruction fine-tuning (IFT) regimes. In addition to open-source models, we also evaluate closed-source APIs such as GPT-4o [15] and Gemini 1.5 Pro [44]. Third, we extend evaluation to VLMs, which are assessed not only on classification accuracy but also on their ability to generate faithful, open-ended rationales grounded in both textual and visual evidence. For classification tasks, we report standard metrics including precision, recall, F1 score, and accuracy. For open-ended reasoning, we employ an LLM-as-judge [57] framework to evaluate the quality, faithfulness, and consistency of generated rationales.

Table 2: BiasCorpus analysis summary.

Metric	**Value**
Rows	40,945
Field names	`unique_id`, `headline`, `url`, `article_text`, `image`, `multimodal_label`, `reasoning`
Top outlets	*Financial Times* (4,895), *USA TODAY* (3,320), *CNN* (3,067), *The Guardian* (2,833), *Forbes* (2,482)
Label classes	2 (*Biased*=22,974; *Not Biased*=17,971)
Unique images	40,945

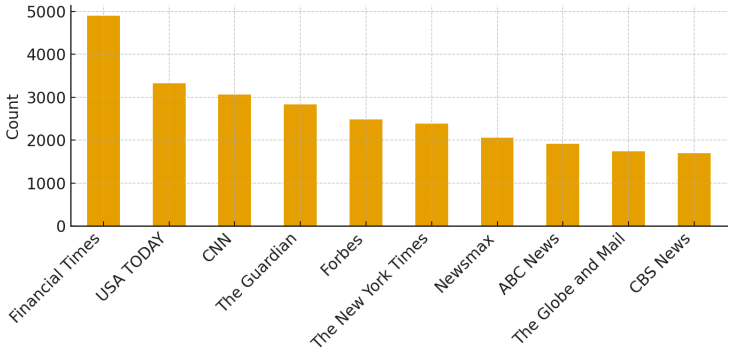

Figure 4: Top 10 outlets

4 Results and Discussion

4.1 Experimental Setting

Compute & software. We run our experiments on NVIDIA A40/A100 Tensor Core GPUs. The stack used Python 3.10, PyTorch 2.0, Hugging Face Transformers, and OpenCV.

Baselines and model families. We employed a suite of baselines to evaluate on our data, which are:

1. Text-only SLMs: BERT-base-uncased [10], DistilBERT [38], RoBERTa-base [23], GPT-2 (small) [30], and BART-base [18].

2. Decoder-only LLMs (zero-/few-shot and IFT): open-source LLaMA-3.2-Instruct [4], Mistral-7B-Instruct [45]; closed-source: GPT-4o [15], Gemini 1.5 Pro [44].

3. VLMs (text+image): Phi-3-Vision-128k-Instruct [36], LLaVA-1.6 [19], LLaMA-3.2-11B-Vision-Instruct [4], plus GPT-4o-mini [15] and Gemini 1.5 Pro [44].

For text-image annotation we used LLaVA-NeXT [21], Phi-3-Vision-128k-Instruct [24], and MiniCPM-V-2_6 [54], resolving conflicts via a two-step hierarchical voting scheme.

Protocols and metrics. We evaluated models under three regimes: zero-shot, five-shot, and full fine-tuning. The dataset was partitioned into training, validation, and test splits with an 80/10/10 ratio, and we additionally employed 5-fold cross-validation to assess robustness. For classification tasks, we report standard metrics including precision, recall, F1 score, and accuracy. To capture performance on open-ended outputs, we further evaluate reasoning quality using LLM-as-judge assessments, focusing on reasoning accuracy and faithfulness. We used GPT4o as the judge model.

We compare multimodal models in two complementary setups. In the closed-ended setting, models generate only a categorical label (biased or not biased), and evaluation relies on accuracy-based metrics against ground-truth annotations. In contrast, the open-ended setting (oVQA) requires models to produce both a label and a concise rationale. In this case, rationales are assessed by an LLM judge, which evaluates their correctness and consistency with the given evidence.

Exploratory Analysis. Figure 4 presents an overview of our dataset through exploratory data analysis.

4.2 Performance Evaluation

In this section, we present the performance of multimodals on our dataset and show main results.

Table 3: **Vision–Language (Text+Image) Performance.** Metrics are reported in percentages: Precision (Prec.), Recall, F1, and Accuracy (Acc.). Closed-source models are marked with †. Results are grouped into SLMs and VLMs. For large models (Phi, LLaVA, LLaMA), we report *Instruct* variants. Configurations include fine-tuning (FT), instruction fine-tuning (IFT), and few-shot (0-shot, 5-shot). All models use the same splits and both modalities. Bold means best values in each model family.

Model	Config	Prec. (%)	Recall (%)	F1 Score (%)	Acc. (%)
Small VLMs					
XLNet + VGG-19	FT	72.0	68.2	70.1	77.1
BERT + RN-34	FT	75.8	71.5	73.6	79.4
BERT + CLIP	FT	81.3	73.4	77.2	84.2
DistilBERT + CLIP	FT	68.5	64.0	66.2	74.9
RoBERTa + CLIP	FT	**84.5**	**81.4**	**82.9**	**83.6**
Large VLMs					
Phi-3-Vision	0-shot	70.4	66.0	68.1	69.8
	5-shot	73.2	71.0	72.1	70.5
	IFT	**76.8**	**78.1**	**77.4**	**74.0**
LLaVA-1.6	0-shot	62.5	60.8	61.6	62.7
	5-shot	68.1	67.0	67.5	65.2
	IFT	75.4	74.6	75.0	76.1
LLaMA-3.2-11B	0-shot	65.0	68.9	66.9	68.4
	5-shot	73.4	74.2	73.8	72.1
GPT-4o mini†	0-shot	71.8	74.6	73.2	72.9
	5-shot	**77.9**	**79.8**	**78.8**	**77.2**
Gemini-1.5 Pro†	0-shot	70.9	73.1	72.0	71.5
	5-shot	76.8	78.5	77.6	76.9

Table 4: Reasoning accuracy and faithfulness (LLM-judged) for open-ended rationales. Closed-source models are marked with †. Bold means best value in each model family.

Model	Config	Reasoning Acc. (%)	Faithfulness (%)
Phi-3-Vision	0-shot	58.2	71.0
	5-shot	62.7	74.5
	IFT	67.9	79.8
LLaVA-1.6	0-shot	52.4	68.1
	5-shot	55.8	70.2
	IFT	**68.5**	**80.4**
LLaMA-3.2-11B	0-shot	59.6	72.3
	5-shot	65.4	78.0
GPT-4o mini †	0-shot	65.1	80.2
	5-shot	**70.8**	**84.6**
Gemini-1.5 Pro †	0-shot	63.4	79.0
	5-shot	70.2	83.1

4.2.1 Traditional fine-tuning is still competitive for classification

We compare different VLMs (both small vs. large) and across regimes (0-shot, 5-shot, IFT) in a multimodal setting and present results in Table 3. The results show clear trends across scales and regimes. Among small VLMs, RoBERTa+CLIP attains the best F1 (82.9) and Acc (83.6), followed by BERT+CLIP (77.2 F1), highlighting gains from strong language encoders with multimodal backbones. Large open-source models improve from 0-/5-shot to IFT (e.g., Phi-3-Vision from 68.1 to 77.4 F1; LLaVA-1.6 from 61.6 to 75.0 F1). Closed-source systems perform strongly in 5-shot (e.g., GPT-4o mini 78.8 F1, Gemini-1.5 Pro 77.6 F1).
Overall, our result findings show: (i) CLIP boosts small-model performance, (ii) IFT markedly improves large open-source models, and (iii) closed-source instruct models remain competitive.

4.2.2 Open-ended reasoning (oVQA): Strong classifiers also tend to be strong Reasoners

In the open-ended setting (oVQA), we evaluate autoregresive VLM models output both a label and a short rationale. We then use a judge LLM (GPT4o) to evaluate the rationale, reporting reasoning accuracy (the rate at which the rationale correctly supports the ground-truth decision) and faithfulness (the rate at which the rationale is supported by the given input). We also measure the correlation between closed-ended accuracy and open-ended reasoning accuracy across all models and regimes.

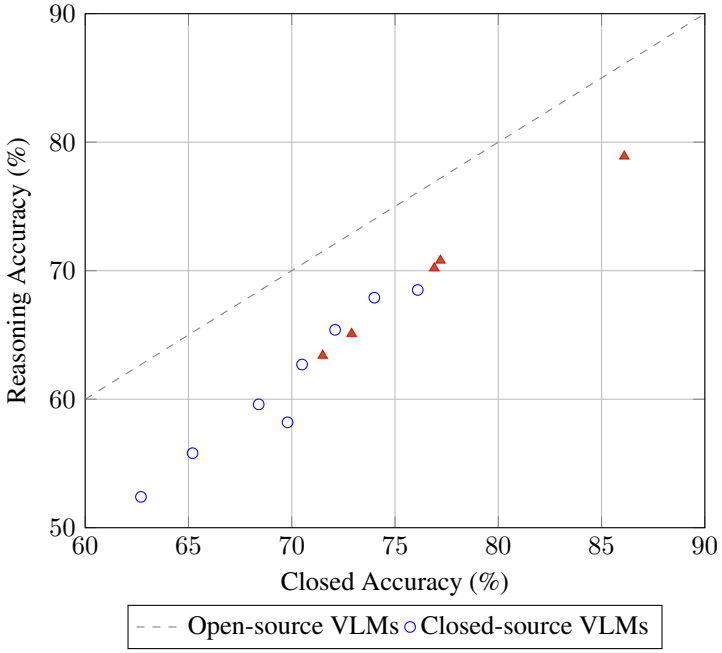

Figure 5: Closed vs open-ended reasoning accuracy across VLMs. The dashed diagonal indicates perfect alignment ($y = x$). Closed-source models (red triangles) cluster higher than open-source models (blue circles).

The results in Table 4 show that reasoning accuracy is consistently lower than closed-ended accuracy, highlighting the difficulty of generating correct and grounded rationales compared to predicting labels. Across open-source models, reasoning accuracy ranges from 52–68% in the 0- and 5-shot settings, improving to nearly 69% under instruction tuning. Faithfulness scores follow a similar upward trend, with most models exceeding 75% once tuned, indicating that rationales are increasingly supported by the input. Closed-source systems (GPT-4o mini, Gemini-1.5 Pro) achieve higher reasoning accuracy and faithfulness overall, with GPT-4o IFT reaching close to 79% reasoning accuracy and nearly 89% faithfulness.

Despite the performance gap as seen in Table 4, the correlation analysis (Pearson $r = 0.91$) in Figure 5 shows a strong linear relationship between closed-ended accuracy and reasoning accuracy. This suggests that models which are strong classifiers also tend to be strong reasoners, although reasoning lags consistently behind classification by 6–12 percentage points. Instruction tuning narrows this gap for both open- and closed-source models, but does not eliminate it entirely.
Overall, these result findings show that that closed-source VLMs are stronger reasoners, yet *all* models face a consistent reasoning gap compared to their closed-ended performance.

5 Discussion

5.1 Practical and Social Impact

Our findings carry both practical and social significance, as shown in Figure 6. By applying LLMs and VLMs, we introduce a scalable framework for detecting biases within multimodal news media.

This is particularly relevant for content moderation, journalism, and public policy, where identifying subtle or underrepresented viewpoints can improve accountability and foster greater transparency.

Socially, our framework empowers more equitable media practices by revealing hidden biases that shape public opinion and perpetuate stereotypes. In practical terms, this means stakeholders can systematically uncover and address harmful or skewed narratives. Integrating humanmachine collaboration also allows for efficient annotations without sacrificing critical context or cultural sensitivity, lowering the chances of purely algorithmic errors.

Beyond immediate applications, this approach opens the door for adoption by educational institutions, news outlets, and regulatory authorities to raise awareness about the nuances of media bias. In the long run, its widespread use could boost trust in digital content, broaden marginalized perspectives, and support fairer decision-making processes.

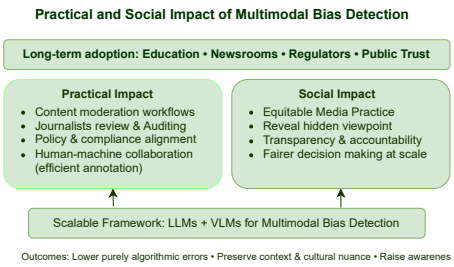

Figure 6: Practical and social impact of VIL-BIAS: a scalable LLM/VLM framework supports content moderation, journalism, and policy (practical), while advancing equity, transparency, and trust (social).

5.2 Limitations

Despite the strengths of our framework, several limitations remain. First, while our BiasCorpus spans a wide range of reputable outlets, its coverage is still skewed toward English-language and Western media. This restricts cross-cultural generalizability [34]and underrepresents media practices in regions such as Africa, South America, and South Asia. Similarly, the dataset employs binary labels (biased vs. not biased), which simplifies the inherently multi-dimensional nature of bias (e.g., partisan, sensationalist, representational). Extending to richer bias taxonomies remains an open challenge.

Second, although our hybrid annotation pipeline mitigates noise through majority voting and HITL oversight, annotation quality is constrained by the interpretive subjectivity of bias judgments. Agreement between automated labels and human reviewers (Cohens $\kappa = 0.72$) reflects substantial alignment but leaves room for disagreement, particularly in subtle or ambiguous cases. Moreover, because LLM as annotators may encode latent biases of their own [6], the benchmark risks inheriting systematic skew, raising the question of "bias in the bias benchmark'.'

Third, our evaluation relies partly on LLM-as-judge assessments to score rationales. While efficient, this introduces circularity, since another model evaluates the faithfulness of generated explanations [57]. Relatedly, reasoning judged as "faithful" may still omit deeper socio-cultural dynamics of framing, highlighting the gap between computational metrics and real-world fairness.

Fourth, model and system constraints persist. Training and fine-tuning large VLMs remains computationally expensive, raising barriers for smaller labs and carrying environmental implications in line with ongoing Green AI debates [17]. Furthermore, our benchmark does not explicitly test robustness under adversarial manipulations (e.g., image perturbations, paraphrased captions), which are increasingly relevant for deployment. Likewise, generalization to other domains such as health or social media memes remains unverified.

Broader societal implications require caution. Automated bias detection tools are not value-neutral: if used irresponsibly by humans, they could reinforce rather than mitigate inequality. For example, governments or organizations may adopt such systems to suppress dissenting voices, while journalists or educators may over-rely on algorithmic judgments at the expense of critical human interpretation. These risks underscore the need for human accountability, transparency, and governance in every stage of system deployment [35]. In addition, our framework does not yet address fairness across demographic stereotypes (e.g., race, gender, religion), an important future direction for ensuring that both humans and machines act responsibly in mitigating bias.

In sum, VILBIAS offers a scalable and rigorous benchmark, but its current scope, annotation design, and evaluation protocols should be seen as an initial step. Addressing cultural diversity, adversarial robustness, sustainability, and human responsibility will be essential for making multimodal bias detection truly comprehensive and socially beneficial.

6 Conclusion

In this work, we introduced VILBIAS, a multimodal benchmark and framework for detecting and reasoning about bias in news media. By curating over 40,000 textimage pairs with hybrid LLM-as-annotator and HITL validation, we provide one of the first large-scale resources that evaluates both closed-ended classification and open-ended reasoning. Our experiments across SLMs, LLMs, and VLMs reveal consistent gains from multimodal grounding, demonstrate that parameter-efficient methods recover near full fine-tuning performance, and highlight the persistent gap between classification accuracy and reasoning quality. Together, these findings offer a rigorous foundation for advancing multimodal bias detection and evaluation.

At the same time, our study underscores the importance of recognizing limitations. The dataset, while diverse, does not fully capture cultural and linguistic variation, and rationales remain shaped by subjective interpretations of bias. Evaluation protocols that rely on LLM-as-judges raise questions of circularity, while computational requirements restrict accessibility for smaller labs. Most importantly, the broader societal risks of misuse are non-trivial: if deployed irresponsibly by governments, media organizations, or educators, such systems could reinforce rather than mitigate inequality.

Looking forward, we see three key directions. First, expanding coverage to more cultures, languages, and nuanced bias categories will strengthen representativeness [33]. Second, robustness to adversarial manipulations and temporal shifts should be incorporated to ensure adaptability. Third, governance frameworks are needed to ensure that bias detection tools remain transparent, accountable, and augment rather than replace human judgment. By advancing technical capabilities while acknowledging risks and responsibilities, VILBIAS takes a step toward fairer, more trustworthy multimodal media analysis.

References

[1] Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. Fairness and bias in multimodal ai: A survey, 2024. URL https://arxiv.org/abs/2406.19097.

[2] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.576.

[3] AllSides. Media bias, 2024. URL https://www.allsides.com/media-bias. Accessed October 22, 2024.

[4] AI at Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Technical report, Meta, 2024.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[6] Veronica Chatrath, Marcelo Lotif, and Shaina Raza. Fact or fiction? can llms be reliable annotators for political truths? *arXiv preprint arXiv:2411.05775*, 2024.

[7] Carmela Comito, Luciano Caroprese, and Ester Zumpano. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining*, 13 (101):1–20, 2023.

[8] Wei Cui and Mingsheng Shang. Migcl: Fake news detection with multimodal interaction and graph contrastive learning networks. *Applied Intelligence*, 55(78):1–15, 2024.

[9] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 392–401, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[11] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*, 2019.

[12] Fabrizio Gilardi et al. Chatgpt as a data collection tool: A critical reflection. *Political Analysis*, 2023.

[13] Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. Language Models Hallucinate, but May Excel at Fact Verification. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.62. URL https://aclanthology.org/2024.naacl-long.62.

[14] Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. Annollm: Making large language models to be better crowdsourced annotators, 2023.

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] S. Jindal, R. Sood, Richa Singh, Mayank Vatsa, and T. Chakraborty. Newsbag: A multimodal benchmark dataset for fake news detection. In *CEUR Workshop Proceedings*, volume 2560, pages 138 – 145. CEUR-WS, 2020.

[17] Tahniat Khan, Soroor Motie, Sedef Akinli Kocak, and Shaina Raza. Optimizing large language models: Metrics, energy efficiency, and case study insights. *arXiv preprint arXiv:2504.06307*, 2025.

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[20] Luyang Lin, Lingzhi Wang, Jinsong Guo, Jing Li, and Kam-Fai Wong. Inditag: An online media bias analysis and annotation system using fine-grained bias indicators. *arXiv preprint arXiv:2403.13446*, 2024.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[22] Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*, 2024.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach, 2019. URL http://arxiv.org/abs/1907.11692.

[24] Microsoft. Phi3 cookbook. Technical report, Microsoft, 2024.

[25] Yeo-Gyeong Noh, MinJu Han, Junryeol Jeon, and Jin-Hyuk Hong. Biassist: Empowering news readers via bias identification, explanation, and neutralization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025.

[26] OpenAI. GPT-4o System Card, August 2024. URL https://cdn.openai.com/gpt-4o-system-card.pdf. White-paper style system card, version released August 8, 2024. Accessed 2025-04-24.

[27] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 3644, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701870. doi: 10.1145/3592572.3592842. URL https://doi.org/10.1145/3592572.3592842.

[28] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.

[29] Shrimai Prabhumoye, Rafal Kocielnik, Mohammad Shoeybi, Anima Anandkumar, and Bryan Catanzaro. Few-shot instruction prompts for pretrained language models to detect social biases. *arXiv preprint arXiv:2112.07868*, 2021.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[31] Faiaz Rahman. Multimodal fake news detection via clip-guided learning. In *Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)*, pages 1234–1243. IEEE, 2022.

[32] Shaina Raza, Oluwanifemi Bamgbose, Veronica Chatrath, Shardule Ghuge, Yan Sidyakin, and Abdullah Yahya Mohammed Muaad. Unlocking bias detection: Leveraging transformer-based models for content analysis. *IEEE Transactions on Computational Social Systems*, 2024.

[33] Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, 237:121542, 2024.

[34] Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Mukund S Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya. Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*, 2025.

[35] Shaina Raza, Rizwan Qureshi, Anam Zahid, Joseph Fioresi, Ferhat Sadak, Muhammad Saeed, Ranjan Sapkota, Aditya Jain, Anas Zafar, Muneeb Ul Hassan, et al. Who is responsible? the data, models, users or regulations? responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650*, 2025.

[36] Microsoft Research. Phi-3-mini-128k-instruct, 2024. URL https://huggingface.co/microsoft/Phi-3-mini-128k-instruct. Dense 3.8B parameter, decoder-only Transformer model with supervised fine-tuning and alignment techniques, supporting up to 128,000 tokens of context.

[37] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 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, 2024.

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019. URL http://arxiv.org/abs/1910.01108.

[39] Karen Santos-DAmorim and Májory K Fernandes de Oliveira Miranda. Misinformation, disinformation, and malinformation: clarifying the definitions and examples in disinfodemic times. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 26, 2021.

[40] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3):171–188, June 2020. ISSN 2167-6461. doi: 10.1089/big.2020.0062. URL https://www.liebertpub.com/doi/abs/10.1089/big.2020.0062. Publisher: Mary Ann Liebert, Inc., publishers.

[41] KMA Solaiman. Biaslab: Toward explainable political bias detection with dual-axis annotations and rationale indicators. *arXiv preprint arXiv:2505.16081*, 2025.

[42] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. Neural media bias detection using distant supervision with babe–bias annotations by experts. *arXiv preprint arXiv:2209.14557*, 2022.

[43] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[45] Mistral AI Team. Mistral 7b the best 7b model to date. Technical report, Mistral, 2023.

[46] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524, 2018.

[47] U.S. Equal Employment Opportunity Commission (EEOC). Discrimination by type, 2024. URL https://www.eeoc.gov/discrimination-type. Accessed: 2024-12-19.

[48] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness, 2024.

[49] Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model, 2024.

[50] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067.

[51] Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. Introducing mbib-the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2765–2774, 2023.

[52] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. 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 1819–1862, 2024.

[53] Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, pages 5285–5296, 2025.

[54] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL `https://arxiv.org/abs/2408.01800`.

[55] YouGov. Trust in media 2024: Which news outlets americans trust, 2024. URL `https://today.yougov.com/politics/articles/49552-trust-in-media-2024-which-news-outlets-americans-trust`.

[56] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.