Preprint of the paper:

Wessel, M., Spinde, T., Pfeffer, J. & Demartini, G., "Definitional Sensitivity in Media Bias Detection: A Multi-Definition Dataset and Benchmark", to appear in Findings of the Association for Computational Linguistics: EMNLP 2026.

Definitional Sensitivity in Media Bias Detection: A Multi-Definition Dataset and Benchmark

Martin Wessel[1,2], Timo Spinde[3], Jürgen Pfeffer[1], Gianluca Demartini[4]
[1]Technical University of Munich, [2]Center for Digital Technology and Management (CDTM), [3]National Institute of Informatics (NII), Tokyo, [4]University of Queensland
Correspondence: m.wessel@media-bias-research.org

Abstract

Media bias detection relies on definitions and examples that specify what counts as bias, yet these specifications often vary across datasets or remain implicit, even when given the same name. Such variation makes it unclear whether models trained for the same bias category learn the same construct or different phenomena, a problem largely overlooked in prior work. We examine how definition choice affects bias annotation in a between-subjects experiment with 354 participants and a parallel evaluation with four LLMs. Participants and models rate six news articles across four bias categories using definitions that vary in conceptual framing and elaboration. Across 8,496 human and 28,800 LLM ratings, we find that the conceptual target of a definition drives annotation divergence, while construct-preserving elaboration does not: conceptual framing significantly shifts annotations for humans and does so even more strongly for LLMs. We discuss implications for construct specification in annotation protocols and prompt-based measurement, and consider how definitional sensitivity may propagate to downstream classification beyond media bias. We also release MUDD, the Multi-Definition Bias Detection Dataset.

1 Introduction

Media bias, the slanting of media content through language, framing, or selection choices, can shape public opinion and political attitudes (Entman, 2007; Spinde, 2025). Detecting it at scale relies on classifiers trained on human-annotated datasets, where annotators apply a given definition of bias to concrete texts. That conceptual definition is the bridge between the abstract construct of "bias" and the labels that train and evaluate models. Yet a review of 115 datasets finds that most rely on varying, often unstated, definitions of what counts as bias (Wessel et al., 2023), and bias categories such as context bias or framing are overlapping and theoretically complex, leaving substantial interpretive room when a definition is applied to a full article. Figure 1 illustrates the resulting risk: the same article, rated under three standard concepts of linguistic bias drawn from the literature, receives substantially divergent ratings.

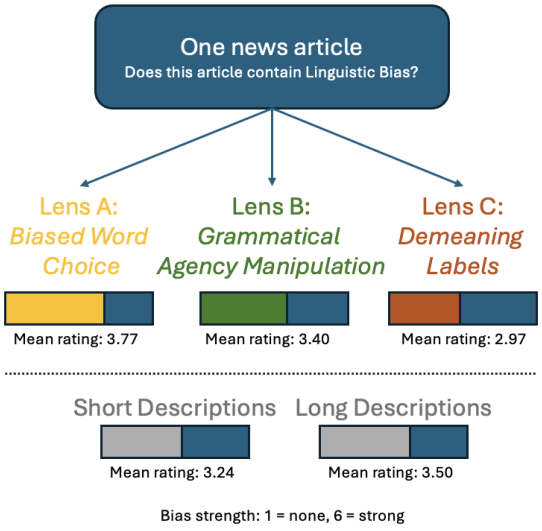

Figure 1: Same article, same bias category, three verdicts. Three standard conceptual definitions of *linguistic bias* from the literature (biased word choice, grammatical agency manipulation, demeaning labels) yield significantly different mean bias ratings on the same 300 articles.

If such definitional choices systematically push annotators toward different judgments on the same texts, then datasets that appear to measure the same construct may in fact capture different phenomena, undermining comparability across studies and the validity of classifiers trained on them. Whether this happens, and how strongly, is an open empirical question: prior work documents that definitions vary across datasets but does not isolate the effect of that variation on annotation behavior.

We address this gap with a pre-registered, between-subjects experiment (a within-subjects de-

sign would let annotators compare definitions, confounding the effect itself). In a 3 × 2 factorial design (concept × length), 354 U.S.-based participants each rate six of 300 news articles across four bias categories (linguistic, gender, political, and context bias) using one of six randomly assigned definitions per category. The three concepts per category are theoretically grounded; the two lengths vary construct-preserving elaboration while holding the concept fixed. We replicate the experiment with four large language models to compare human and LLM sensitivity to definitional variation.

Our results yield three contributions. (1) We provide the first controlled evidence that *what* a definition conceptually specifies, not *how much* construct-preserving elaboration it provides, is a primary driver of annotation divergence. We find significant concept effects across all four bias categories, robust to political orientation, gender attitudes, and clarity. (2) The tested LLMs amplify this definitional sensitivity by approximately 1.3–4× in absolute rating differences relative to humans, and the direction of prompt effects can reverse across architectures. (3) We release all 8,496 human annotations, 28,800 LLM annotations, 24 bias definitions, and full experimental materials as **MUDD**, the **Multi-Definition Bias Detection Dataset** for further research on definitional sensitivity in media bias detection.[1]

2 Related Work

2.1 Definitional Variation Beyond Prompt Optimization

Instruction and prompt wording are known to shift annotation and model behavior, but the effect of the underlying *conceptual definition* of a target construct has received far less attention. In media bias specifically, Wessel et al. (2023) find that the 115 datasets in the MBIB benchmark rely on inconsistent and often unstated definitions, a pattern they identify systematically. Spinde et al. (2023) and Rodrigo-Ginés et al. (2024) propose taxonomies that organize these conceptualizations, and Spinde et al. (2025) argue, via an automated extraction framework, that the sheer volume and heterogeneity of available definitions makes their impact on downstream annotation a relevant concern. In adjacent work, Röttger et al. (2022) show that descriptive versus prescriptive schemes yield fundamentally different hate speech datasets from the same content. Definitional divergence is documented, but its causal effect on media bias annotation, a domain whose concepts are overlapping and theoretically complex, has not been tested directly.

2.2 How Definitions Affect Human Annotators

Within human annotation, guideline specificity is the closest proxy to what we study. For media bias, Fan et al. (2019) and Hamborg et al. (2019) find that narrowly defined bias types yield more focused datasets, while Salavati et al. (2024) suggest clearer definitions improve quality at the cost of generalizability, and Spinde et al. (2021b) report persistent low inter-annotator agreement even with clear guidelines. Beyond media bias, Ross et al. (2017) find that providing a hate speech definition does not improve agreement, and Parmar et al. (2023) show that guidelines at different detail levels produce significantly different labels, with more detail not necessarily helping. Annotator identity adds a second source of variation: political and demographic characteristics predict subjective ratings (Sap et al., 2022; Al Kuwatly et al., 2020; Talat, 2016), and annotator disagreement often reflects genuine ambiguity rather than noise (Plank, 2022; Uma et al., 2021).

Together, these findings suggest definitions and annotator traits both move labels, but none isolates the causal effect of the definition itself while holding articles, interface, and annotators constant, motivating H1–H3:

H1 Annotators assigned different conceptual definitions of the same bias category produce significantly different bias strength ratings for the same articles, and annotators assigned longer, more detailed definitions produce significantly different ratings than those assigned shorter definitions of the same concept.

H2 Annotators assigned different conceptual definitions or definition lengths report significantly different confidence in their bias judgments.

H3 After controlling for conceptual definition and definition length, the effects of annotators' political orientation and gender attitude on bias ratings are smaller than the concept effect.

[1] Data, scripts and supplementary materials are available here and the preregistration here.

2.3 How Definitions Affect LLM Annotators

LLMs increasingly serve as scalable annotation alternatives: Horych et al. (2025) show LLM annotations can match or outperform expert data for media bias, and Gilardi et al. (2023) find ChatGPT outperforms crowd workers on several tasks at a fraction of the cost. Their reliability, however, depends heavily on how the task is framed: Pangakis et al. (2023) report substantial quality variation with task framing on subjective tasks, and Törnberg (2024) show performance relies on prompt design and coding-instruction specificity. What remains underexplored is sensitivity to the definition of the target construct itself, as opposed to surface-level prompt wording, which we test alongside the human experiment:

H4 Definitional variation produces smaller variance in bias ratings among human annotators than among LLM annotations on the same articles and scale.

3 Methodology

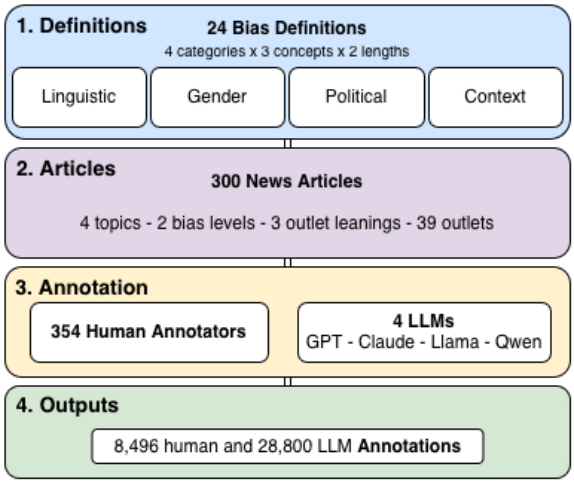

Figure 2: Experimental pipeline.

3.1 Research Design

Our design aims to (1) isolate the effect of a bias definition from surface features such as length, register, and example count, (2) vary only theoretically separable concepts within a category so that any rating differences can be attributed to conceptual content, and (3) ensure that all definitions are comprehensible to non-expert annotators. We run a pre-registered, between-subjects experiment using a 3×2 (concept $\times$ length) factorial design (Figure 2 outlines the pipeline), applied independently across four bias categories: linguistic, gender, political, and text-level context bias. We select these four to cover both mechanism-based bias (linguistic and context) and topic-based bias (gender and political), following the taxonomy of (Spinde et al., 2023). For each category, we review existing bias definitions, classify each by its underlying mechanism, and select three concepts that target separable mechanisms and can independently vary within a single text, so that an article can exhibit one concept while being neutral on the others. We fix the number of concepts to three *a priori* to balance experimental power with feasibility in a between-subjects design. We write the definition texts ourselves, based on the reviewed source conceptualizations, and standardize them across concepts for length, register, and example count, so that only the targeted mechanism varies. We pilot-test the resulting definitions on Prolific ($n = 24$) using the full main-study design supplemented with comprehensibility items; feedback from this pilot leads us to adopt easy-language phrasing throughout. We provide a detailed description of the definition-derivation process, and final definition texts in Appendix C.

We randomly assign each participant one of six definitions per category (three conceptual variants in either a short or long version) which use these to rate bias in news articles. Short- and long-definition groups see only definitions of their assigned length, with no within-participant mixing. For H3, we pre-specify a smallest effect of interest of ± 0.10 scale points for covariate effects. We operationalize length as construct-preserving elaboration: long definitions add explanation and examples that reinforce the same target mechanism.

Table 1 summarises the three conceptual sub-definitions (C1–C3) per category.

	Linguistic	Gender	Political	Context
C1	Word choice	Visibility	Slant	Omission
C2	Agency	Stereotypes	Coverage	Sources
C3	Labels	Irrelevant focus	Framing	Emotion

Table 1: Concepts per bias category. For a detailed description and theoretical sources see Appendix C.4.

3.2 Materials and Participants

Bias perception varies systematically with article topic, outlet leaning, and framing style, so a diverse selection is necessary to avoid confounding definition effects with stimulus effects (Spinde et al.,

2021a, 2023). We therefore select 300 English-language news articles from Raza (2024), sampled across four topic domains (politics, social issues, economy, science) and stratified by NLP-predicted bias label (biased vs. unbiased, provided by (Raza, 2024)) as well as source outlet political leaning (left, center, right), comprising 39 outlets spanning the ideological spectrum.

We run an a priori power analysis (power = .80, $\alpha = .05$, $f = 0.25$) which indicates that the most demanding test (a three-way interaction in the $3 \times 2 \times 3$ design) requires $N = 342$. We recruit 354 participants through Prolific (not including 84 who drop out and 18 we reject for insufficient response time). All participants are U.S. residents fluent in English, and we use Prolific's prescreening filters to approximate U.S. Census distributions for political orientation, gender, and age.

3.3 Procedure and Measures

Because both article characteristics and annotator characteristics shape bias judgments, we collect participant covariates alongside the annotation task (Spinde et al., 2021c; Sap et al., 2022). We administer the study via Qualtrics. After informed consent and a demographics questionnaire (age, gender, education, political orientation on an 8-point scale, two gender attitude items, news consumption habits), participants enter the annotation task.

Each participant reads six articles sequentially and rates each for bias across all four categories using their assigned definition. Bias ratings use a 6-point Likert scale (1 = Strongly disagree to 6 = Strongly agree) in response to: *"In my opinion, this article contains [category] bias."* Confidence is reported after each rating on a 7-point scale. Definitions are accessible via a collapsible element on each page. We embed an attention check that allows one retry; participants who fail twice are excluded.[2] After the task, participants rate the clarity of each definition using a six-item UEQ-adapted scale (covering understandability, value, supportiveness, quality, ease, and clarity, each on a 7-point semantic differential scale) and report overall reading attention. We record page-level timing and definition re-click behavior throughout. A full list of measures is provided in Appendix A.

[2]The attention check tests whether participants grasp the general concept of bias after the definitions are displayed. The exact item is in the supplementary materials.

We assign participants to conditions via a counterbalanced design: 27 covering sets of three definition profiles each, used twice (once per length group), ensure that all 300 articles are annotated under all six definitions across all four categories, yielding 7,200 article–category–definition combinations.

3.4 LLM Annotation

To ensure that the interpretation of definitions is directly comparable between humans and models, we replicate the human task across four LLMs chosen to span provider and cost tiers: two frontier commercial models, GPT-4o (OpenAI, 2024) and Claude Sonnet 4.6 (Anthropic, 2026), and two same-size open-weight models from different providers, Llama 3.1 8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Team, 2024). Each model receives a structured prompt containing the article text, one of the 24 definitions, and instructions to rate bias (1–6) and confidence (1–7).[3] Each model annotates all 300 articles under all 24 definitions (7,200 annotations per model) at temperature 0.0, with each article–definition pair submitted as an independent query. We access GPT-4o and Claude via commercial APIs and run Llama and Qwen via Hugging Face. We analyse LLM annotations using the same statistical framework as the human data.

4 Results

We collect 354 completed responses (8,496 valid ratings) from U.S. Prolific workers: 174 women, 174 men, 6 non-binary; mean age 44.9 ($SD = 15.9$, range: 18–84); mean political orientation 2.99 ($SD = 2.28$) on a 0–7 scale (0 = very liberal, 7 = very conservative); 78% consume news at least several times per week. Educational attainment is diverse, with the largest groups holding a bachelor's degree ($n = 123$) or some college education ($n = 76$). [4]

[3]The exact prompt template is available in the supplementary materials. We acknowledge that prompt form can influence LLM annotations; because all four models receive an identical prompt template that varies only in the injected definition text, any between-definition variation within a model is attributable to the definition itself rather than to scaffolding.

[4]Data quality was high: only 5 of 354 participants (1.4%) failed an attention check, and none failed both. Self-reported reading attention was high ($M = 5.64$, $SD = 0.66$ on a 7-point scale), with no difference between length conditions. Bias ratings show no systematic drift across article positions ($r = .008$, $p = .43$), despite a 41% decline in reading time from Article 1 (176.0 s) to Article 6 (103.5 s).

4.1 Concept Choice Dominates Definition Length

Category means range from M=2.12 (gender) to M=3.61 (context). Re-click rates are higher for linguistic (21.3%) and context (21.0%) than for gender (6.5%) and political (6.5%), suggesting definitional dependence is greatest where the construct is least intuitive. Because each participant is assigned a single definition per category, the concept and length factors are fully between-subjects and not crossed with participant; participant-level random intercepts are therefore largely unidentifiable from definition effects, and we model article as the primary source of non-independence.

Concept effects. One-way ANOVAs detect concept effects for all four bias categories. Linguistic bias shows the strongest effect ($F(2, 2115) = 48.57$, $p < .0001$, $\eta^2 = .044$), with all three concepts differing: Concept 1 (word choice, $M = 3.77$) > Concept 2 (grammatical agency, $M = 3.40$) > Concept 3 (demeaning labels, $M = 2.97$); all pairwise comparisons survive Bonferroni correction ($d = 0.25$–0.53). Gender bias also shows a meaningful concept effect ($F(2, 2115) = 11.64$, $p < .0001$, $\eta^2 = .011$), with Concept 1 (representational visibility, $M = 2.33$) rated higher than both Concept 2 (stereotypical traits, $M = 2.02$, $d = 0.22$) and Concept 3 (irrelevant focus, $M = 2.03$, $d = 0.22$), while Concepts 2 and 3 do not differ ($d = -0.001$, $p = .99$). Context ($F(2, 2115) = 3.97$, $p = .019$, $\eta^2 = .004$) and political bias ($F(2, 2115) = 3.22$, $p = .040$, $\eta^2 = .003$) also produce statistically detectable concept effects, but with smaller effect sizes.

Length effects. Definition length, in contrast, has negligible effects. Independent-samples t-tests reveal a length effect only for linguistic bias (long $M = 3.50$ vs. short $M = 3.24$; $t = -3.76$, $p = .0002$, $d = -0.16$); gender ($p = .27$), political ($p = .33$), and context bias ($p = .19$) show no length effects. Two-way ANOVA confirms that for linguistic bias the concept effect ($\eta^2 = .043$) substantially exceeds both the length effect ($\eta^2 = .006$) and the Concept × Length interaction ($\eta^2 = .005$). Significant Concept × Length interactions appear for linguistic bias ($F(2, 2112) = 5.98$, $p = .003$) and, more weakly, for gender bias ($F(2, 2112) = 2.83$, $p = .059$); political ($p = .272$) and context bias ($p = .537$) show none.

Mixed-effects confirmation. To account for repeated ratings of each article across definitions, we refit the main tests as linear mixed-effects models with a random intercept for article (`rating ~ concept * length + (1|article)`; likelihood-ratio tests). Concept remains significant in all four categories (linguistic $\chi^2(2) = 110.0$; gender $\chi^2(2) = 27.8$; political $\chi^2(2) = 11.0$; context $\chi^2(2) = 9.9$; all $p < .01$), while length is significant only for linguistic bias ($\chi^2(1) = 14.8$, $p < .001$; gender, political, and context all $p > .09$), reproducing the ANOVA pattern. The article random intercept absorbs the expected between-article variance, most strongly for political bias (ICC = .44) and least for the more subjective categories (ICC = .10–.17), consistent with political bias being a comparatively shared construct (Section 5).

Annotator characteristics. ANCOVA models bound political orientation and gender attitude effects within ± 0.10 scale points across all four categories (TOST, all $p < .05$), with definition variables accounting for virtually all explained variance. All 24 definitions receive comparable clarity ratings ($M \approx 3.84$–4.07), ruling out comprehension as a driver.

4.2 Within-Definition Reliability

Although the between-subjects design assigns most article–definition combinations to a single annotator, we produce an overlap on 15% of the 7,200 total combinations, each rated by 2–4 independent participants and balanced across the four categories. This subset permits estimating *within-definition* inter-annotator agreement. Overall Krippendorff's α (interval) = 0.39 [95% CI: 0.34, 0.44], with ICC(2,k) = 0.54, mean pairwise $|\Delta| = 1.36$ on the 1–6 scale, and 63% of pairs agreeing within one scale point; α rises to 0.47 on the 200 combinations with ≥ 3 raters, so the full-subset estimate is conservative. Agreement varies sharply by category (Table 2).

Longer definitions yield slightly higher agreement ($\alpha = 0.40$ vs. 0.35). Given this low within-definition agreement, the concept effects reported in Section 4.1 should be interpreted as aggregate shifts in the distribution of ratings across definitions, not as evidence that individual articles receive stable labels under any single definition; the systematic between-definition mean differences we observe arise in addition to, not instead of, substan-

Category	α [95% CI]
Political	0.56 [0.47, 0.63]
Context	0.18 [0.07, 0.28]
Linguistic	0.17 [0.06, 0.27]
Gender	0.11 [−0.04, 0.24]

Table 2: Within-definition Krippendorff's α (interval) by bias category, computed on the multi-rater subset. 95% confidence intervals from 1,000 combination-level bootstrap resamples.

Table 3: Concept effects across three concept levels, by annotator type and bias category: (a) η^2 from one-way ANOVAs, (b) largest pairwise absolute mean difference, (c) largest pairwise Cohen's d.

	Human	GPT-4o	Claude	Llama	Qwen
(a) Effect size (η^2)					
Linguistic	.044***	.171***	.165***	.142***	.346***
Gender	.011***	.234***	.265***	.007**	.104***
Political	.003*	.040***	.014***	.013***	.019***
Context	.004*	.045***	.075***	.040***	.150***
(b) Max pairwise $\lvert\Delta M\rvert$					
Linguistic	0.84	1.47	1.13	1.25	1.08
Gender	0.34	0.96	0.80	0.06	0.59
Political	0.19	0.76	0.40	0.50	0.37
Context	0.21	0.73	0.78	0.74	0.71
(c) Max pairwise $\lvert d\rvert$					
Linguistic	0.56	1.07	0.92	0.90	1.68
Gender	0.25	1.09	1.33	0.16	0.80
Political	0.12	0.50	0.28	0.27	0.35
Context	0.15	0.53	0.75	0.46	0.85

Note. Human n=8,496; LLM n=7,200 each. *** p<.001, ** p<.01, * p<.05.

tial within-definition article-level variability.

4.3 Confidence Reflects Decisiveness

Confidence differs by concept for all four categories (all $p < .05$) but not by length (all $p > .20$). Crucially, confident annotators produce substantially more extreme ratings across all four categories ($r = .42$–$.45$, all $p < .001$), indicating that confidence indexes decisiveness rather than inter-annotator agreement.

4.4 LLMs Amplify Definitional Sensitivity

Table 3 compares concept effects across human annotators and four LLMs: GPT-4o, Claude Sonnet 4.6, Llama 3.1 8B, and Qwen2.5-7B, reporting η^2 from one-way ANOVAs alongside the largest pairwise absolute mean difference ($|\Delta M|$) and Cohen's d. Because η^2 denominators reflect different variance structures across multi-annotator and single-agent settings, $|\Delta M|$ and d enable more direct cross-population comparison.

Amplification across categories. All four LLMs exhibit larger concept effects than humans across almost all categories (except Llama for gender). Gender bias shows the largest amplification for GPT-4o and Claude ($\approx$2.4–2.8$\times$ in $|\Delta M|$), context the most consistent across models ($\approx$3.4–3.7$\times$); full magnitudes in Table 3. Category-level ordering is broadly preserved: linguistic produces the largest effects and political the smallest in every LLM (Spearman ρ=0.80 between Llama and Qwen on η^2). Magnitude, however, is not architecture-invariant: Qwen dominates on linguistic and context bias, GPT-4o and Claude on gender, while Llama is least sensitive throughout, with its near-zero gender effect reflecting floor compression (all concept means $M\approx$1.04–1.10) rather than a general pattern.

Rating level and inter-model agreement. All LLMs rate below humans (M=2.09–2.45 vs. 3.11); Llama's variance most closely matches humans, Qwen's is most compressed. Per-article rating correlations are highest between GPT-4o and Claude (r=.73–.82) and moderate for human–LLM pairs (r=.26–.51). Confidence scores are near-constant across all models (SD=0.69–0.94 vs. human 1.56), so substantial rating shifts are unaccompanied by any signal of uncertainty.

Length effects and interactions. The concept-over-length pattern persists within LLMs but with model-specific signatures and inconsistent directionality: GPT-4o shows small length effects for gender and context (d=0.12, p<.05); Claude shows none (all p>.09); Llama shows effects for linguistic (d=−0.09) and context (d=0.25, the largest across annotator types); Qwen shows long > short effects for gender (d=−0.19) and context (d=−0.26). Significant Concept × Length interactions cluster on context bias; no LLM shows an interaction for political bias, whereas humans do for linguistic bias.

Article-level sensitivity. Article-level rating ranges across definitions span up to 5.0 on the 6-point scale. Mid-scale articles (M=3.0–4.0) show significantly higher definitional sensitivity than extremes for linguistic, gender, and political bias (human d=0.29–1.00; LLM d=0.74–1.46), but not for context (d=−0.07, p=.57). A temperature robustness check on GPT-4o confirms concept effects persist under non-zero model temperatures (Appendix B).

5 Discussion

Across our four bias categories, we find that *what* a definition specifies consistently affects human and LLM ratings of the same articles, while adding construct-preserving detail to a fixed target does not.

5.1 Same Label, Different Constructs

Conceptual framing, not elaboration, drives divergence (H1). Conceptual framing drives annotation divergence, while construct-preserving elaboration of a fixed target does not. As we report in Section 4.1, concept effects are significant across all four bias categories, whereas length reaches significance only for linguistic bias and there with a far smaller effect size.

We interpret this as evidence that annotators form a mental working schema rapidly upon encountering a definition and then absorb additional elaboration into that schema rather than refining it. Re-click rates are consistent with this, suggesting that definitional dependence is greatest where the construct is least intuitive.

Implications for construct validity. We interpret these results as indicating that, in our data, if two researchers operationalise "media bias" using different concepts within the same category, they likely produce systematically different datasets. This happens not because their guidelines differ in specificity, but because they end up measuring different things, even when their wording implies the same category.

Our between-subjects design supports this interpretation. Several plausible confounds are bounded: political orientation and gender attitude effects on ratings are statistically equivalent to zero within a ± 0.10 bound after controlling for definition assignment (**H3**; all TOST $p < .05$), including the Concept $\times$ Gender Attitude interaction. Within-definition agreement (Section 4.2) further clarifies this: political bias combines the highest within-definition α (0.56) with the smallest between-definition effect, suggesting a relatively shared, stable construct, while gender and linguistic bias show large definitional shifts despite low intrinsic agreement, marking them as both intrinsically subjective *and* sensitive to how they are defined. Absolute rating shifts underline this: for linguistic bias, a single definitional choice moves ratings by nearly a full scale point ($|\Delta M| = 0.84$ on a six-point scale; Table 3).

Definitional choices matter most for borderline articles. Our article-level analysis indicates that definitional choices matter most where annotation is already difficult: for three of four categories, mid-scale articles show the largest rating shifts across definitions, while clear-cut cases are comparatively robust. Researchers using different operationalisations of the same bias category are therefore most likely to diverge on borderline cases, which are also the cases classifiers most need to get right.

Confidence indexes decisiveness, not reliability (H2). Confidence varies significantly across concept conditions, but tracks rating extremity ($r = .42$–$.45$) rather than inter-annotator convergence. Confident annotators commit more firmly to their individual reading without agreeing more closely with one another. We therefore interpret confidence as indexing decisiveness rather than certainty, a distinction that matters for aggregation schemes that treat confidence as a reliability proxy. All 24 definitions receive comparable clarity ratings ($M \approx 3.84$–4.07), with no short/long differences (all $p > .42$), so we can rule out comprehension difficulty as a driver.

5.2 Humans Generalise, LLMs Follow the Prompt

LLMs amplify concept effects across categories (H4). GPT-4o, Claude, and Qwen amplify the human concept effect across every bias category (Table 3), and Llama does so in three of four categories.

We propose that the amplification factor reflects the semantic distinctiveness of concepts within a category. For gender bias, our three definitions target semantically separable phenomena with limited lexical overlap, so an LLM prompted with one encounters few cues that would activate the others. For linguistic bias, the concepts share surface features that may activate similar internal representations even under distinct prompts, compressing between-definition differences. We suspect human annotators resist this pattern because they bring a schema-driven understanding of "bias" that partly activates neighbouring constructs regardless of which specific mechanism the definition targets, whereas LLMs lack this integrative prototype and map the provided wording more literally. The Concept $\times$ Length interactions are consistent with this reading: LLMs consistently show significant interactions for context bias, whereas humans do not

(Section 4.4), pointing to different feature integration.

This literalism can help or hurt depending on the goal. When the goal is to approximate the broader, everyday sense of "bias," the same tight coupling to wording becomes a disadvantage: small changes in phrasing can shift, or reverse, the direction of the effect.

LLMs are not drop-in replacements for human annotators. All four LLMs rate bias systematically lower than humans and report near-constant confidence, so substantial rating shifts across definitions come with no signal of uncertainty. GPT-4o and Claude correlate strongly with each other (r=.73–.82) but only moderately with humans (r=.26–.51), which means cross-validating LLMs against one another can produce reassuringly high reliability that masks a shared departure from human judgment. One might ask whether the observed effects depend on the specific concepts we select. We sample from a larger definitional space rather than claiming taxonomic completeness, and the significant, ordered rating gradients indicate that annotators treat the concepts as distinct. A GPT-4o robustness check at T=0.7 confirms concept effects persist under stochastic decoding (B).

5.3 Beyond Media Bias

Media bias is a domain where definitional fragmentation is already documented: existing taxonomies map the contested concepts (Spinde et al., 2023; Rodrigo-Ginés et al., 2024; Spinde et al., 2025), and dataset heterogeneity (Wessel et al., 2023). Our results add a causal layer to this picture. Prior work asks how to phrase bias questions for annotators and how to organise the field's taxonomies; we show that, even once the phrasing is fixed, the definition itself moves the labels. Datasets such as MBIC (Spinde et al., 2021c) and the 115 collections surveyed in MBIB (Wessel et al., 2023) are not directly comparable simply because they share a category label.

We expect the same logic to apply to other subjective labeling tasks whose target admits multiple theoretically grounded concepts. A practical consequence for NLP is that, for such constructs, the choice of concept appears to matter more than the choice of surface phrasing, which shifts attention from prompt wording to construct specification in both human guidelines and LLM prompts.

5.4 Practical Recommendations

We synthesise four recommendations from the analysis above. First, **definitional reporting should become standard**: we recommend that researchers publish exact definition texts and state which conceptual tradition they draw from. Second, **concept selection matters more than construct-preserving elaboration**: our null length effects imply diminishing returns from lengthening a definition once its target construct is fixed. Third, **LLM-based annotation requires sensitivity testing**: we recommend annotating with at least two conceptually distinct definitions per category and reporting the variance, since sensitivity in one category does not generalise to others. Fourth, **within media bias detection, definitions should be treated as experimental interventions, not neutral metadata**: the inconsistency across 115 media bias datasets (Wessel et al., 2023) reflects, a situation in which researchers are running different experiments under the same label.

5.5 Future Work

We see several natural extensions. Deliberately varying definition quality would test whether clarity moderates outcomes when free to vary. Because each participant rates each item once, our design cannot separate intra-annotator noise from definitional effects; embedding hidden duplicates or a test–retest subset would bound how much within-definition variability reflects response noise. Since sensitivity in one category does not predict another, a further step is testing whether concept effects transfer across categories, and whether the definition-dependent model rankings we observe make LLM leaderboards for subjective annotation definition-contingent. More broadly, our findings invite replication across other subjective NLP tasks, larger open- and closed-weight models, and multilingual settings.

6 Conclusion

Across 354 humans and four LLMs, we show that *what* a bias definition specifies, not *how much* construct-preserving detail it adds, drives annotation divergence, with LLMs amplifying this sensitivity architecture-dependently. We release the Multi-Definition Bias Detection Dataset (37,296 annotations across 24 definitions) to support treating definitions as experimental interventions.

Limitations

Our experiment covers four bias categories (linguistic, gender, political, and text-level context bias) selected from a broader taxonomy (Spinde et al., 2023); whether the observed definitional sensitivity generalizes to other bias types (e.g., racial bias, reporting-level bias, framing bias) remains an open question for future work. Our length manipulation is construct-preserving: long definitions elaborate a fixed target without introducing edge cases or negative examples that could redirect interpretation. Varying the interpretation-redirecting scaffolding might risk blurring the concept/length distinction our design deliberately separates, a distinction already visible in annotation-guideline work outside media bias (Parmar et al., 2023; Röttger et al., 2022).

Our sample consists of U.S.-based, English-speaking Prolific crowdworkers. Because bias perception varies with language, culture, and expertise (Spinde et al., 2021a,c), and bias measurement often fails to transfer cleanly to other languages, our findings should be read as concerning U.S. English-speaking annotators; replication with trained journalists, linguists, or non-U.S. participants may yield different absolute levels or amplification patterns. On the stimulus side, the NLP-predicted labels we use to stratify articles (Raza, 2024) are not ground-truth bias indicators and serve only to balance predicted-biased and predicted-unbiased items across our four topic domains (politics, social issues, economy, science), which exclude areas such as health or entertainment. Each participant rates only six articles under a single definition, which strengthens causal identification through the between-subjects design but restricts reliability estimation to the 15% multi-rater subset. Our three concepts per category are a small sample of the much larger definitional space (Spinde et al., 2023; Rodrigo-Ginés et al., 2024; Spinde et al., 2025; Wessel et al., 2023); a broader or differently chosen set could plausibly produce larger effects. We run the main LLM experiment at temperature 0.0 to maximise reproducibility (Horych et al., 2025), and our T=0.7 robustness check covers GPT-4o but not Claude, Llama, or Qwen. Results are bound to the four tested versions; larger models in either the open-weight (for example, 70B-class) or closed-weight tier remain untested, and prior work indicates that framing and prompt effects can interact with model capability in non-monotonic ways (Pangakis et al., 2023; Törnberg, 2024). We also do not manipulate the prompt scaffold itself (for example, chain-of-thought or role prompting); this keeps prompt form identical across conditions and annotator types but leaves open whether a different scaffold would compress or amplify the observed concept effects. Finally, our covariates (political orientation, gender attitudes, news consumption, attention) are subject to social-desirability bias (Spinde et al., 2021a). Reading times decline by 41% across the session, from 176.0 s on Article 1 to 103.5 s on Article 6, which is consistent with fatigue; although we find no rating drift across article positions ($r = .008$, $p = .43$), we cannot rule out fatigue-by-definition interactions our design does not detect. Our study also captures one-shot annotation, and prior work on inter-annotator agreement suggests training can shift but not eliminate definition-dependent variation (Spinde et al., 2021b).

Ethical Considerations

The study is approved by the ethics review board of the University of Queensland. Prior to participation, all annotators receive a detailed information and consent page describing the purpose of the study, the nature of the task, the data collected, the expected duration, and the compensation, and can only proceed after giving explicit informed consent. Participation is fully voluntary, and participants can withdraw at any point without providing a reason. We recruited participants through Prolific and compensated them at an average rate of £14.34 per hour, well above Prolific's recommended fair-pay threshold. We collect no personally identifying information beyond the platform-mediated pseudonymous IDs, which we remove before analysis and before release; all data released as part of the MUDD Dataset is fully anonymised. The news articles used are drawn from an existing, publicly available dataset (Raza, 2024), and we do not redistribute article text beyond what that source already makes available. We intend the released dataset and materials to support research on definitional sensitivity and annotation methodology in subjective NLP tasks; we do not endorse their use for training classifiers that label outlets or individual journalists as biased in production settings, given the definitional sensitivity our results demonstrate.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Anthropic. 2026. System Card: Claude Sonnet 4.6.

Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: Global warming and the us prestige press. *Global environmental change*, 14(2):125–136.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Barbara L Fredrickson and Tomi-Ann Roberts. 1997. Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of women quarterly*, 21(2):173–206.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The quarterly journal of economics*, 120(4):1191–1237.

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

Nancy M Henley, Michelle Miller, and Jo Anne Beazley. 1995. Syntax, semantics, and sexual violence: Agency and the passive voice. *Journal of Language and Social Psychology*, 14(1-2):60–84.

Tomas 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 2025 Conference of the The Nations of the Americas Chapter of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Shanto Iyengar. 1994. *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, 9(3).

Sarah Macharia. 2020. Who makes the news?: 6th global media monitoring project.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

OpenAI. 2024. GPT-4o System Card.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 10671–10682.

Shaina Raza. 2024. Newsmediabias-plus dataset (version 1). https://doi.org/10.5281/zenodo.13961155. Dataset.

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.

Francisco-Javier Rodrigo-Ginés, Laura Plaza, and Jorge Carrillo-de Albornoz. 2023. UnedMediaBiasTeam @ SemEval-2023 task 3: Can we detect persuasive techniques transferring knowledge from media bias detection? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 787–793, Toronto, Canada. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 175–190.

Chiman Salavati, Shannon Song, Willmar Sosa Diaz, Scott A Hale, Roberto E Montenegro, Fabricio Murai, and Shiri Dori-Hacohen. 2024. Reducing biases towards minoritized populations in medical curricular content via artificial intelligence for fairer health outcomes. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1269–1280.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Norbert Schwarz, Herbert Bless, Fritz Strack, Gisela Klumpp, Helga Rittenauer-Schatka, and Annette Simons. 1991. Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, 61(2):195.

Timo Spinde. 2025. *Automated Detection of Media Bias: From the Conceptualization of Media Bias to its Computational Classification*. Springer Vieweg Wiesbaden.

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, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021a. Do you think it's biased? how to ask for the perception of media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Timo Spinde, Luyang Lin, Smi Hinterreiter, and Isao Echizen. 2025. Leveraging large language models for automated definition extraction with taxomatic — a case study on media bias. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'25)*, volume 19, Copenhagen, Denmark. AAAI.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021b. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Spinde, Lada Rudnitckaia, Sinha Kanishka, Felix Hamborg, Bela, Gipp, and Karsten Donnay. 2021c. MBIC – a media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference 2021*, Beijing, China (Virtual Event). Tex.pubstate: published tex.tppubtype: inproceedings.

Zeerak Talat. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*. ACM. ISBN 978-1-4503-9408-6/23/07.

A Measures

Dependent variables.

- *Bias rating:* 6-point Likert scale per article per category (1 = Strongly disagree to 6 = Strongly agree)
- *Confidence:* 7-point scale per rating (1 = Not at all confident to 7 = Extremely confident)

Independent variables.

- *Concept:* Three levels per category (between-subjects)
- *Length:* Two levels: short vs. long definition (between-subjects)

Covariates and control measures.

- Political orientation (8-point continuous scale)
- Gender attitudes (two items)
- Definition clarity (six-item UEQ-adapted scale, 7-point semantic differential)
- Self-reported reading attention (7-point scale)
- Page-level reading time (seconds)
- Definition re-click frequency (count per article per category)
- Attention check pass/fail

B LLM Robustness Check

To separate prompt sensitivity from sampling variance, we re-run GPT-4o on a stratified 60-article $\times$ 24-definition subset at $T{=}0.7$ with five samples per cell. Within-cell SD ranges from 0.22 (political) to 0.45 (context), while the largest between-concept $|\Delta M|$ is $3.7\times$, $3.5\times$, $4.1\times$, and $1.3\times$ the within-cell SD for linguistic, gender, political, and context bias. Concept effects remain significant at $T{=}0.7$ (linguistic $\eta^2{=}.230$, gender .205, political .067, context .041; all $p{<}.001$), matching or exceeding the $T{=}0$ results.

C Definitions

C.1 Definition Derivation Procedure

The four bias categories are selected from a broader taxonomy of media bias types compiled from the literature (Wessel et al., 2023; Spinde et al., 2023) along two dimensions. As *mechanism-based* categories, we select **linguistic bias** and **text-level context bias**, which capture complementary levels of textual organization: local linguistic choices versus global informational structure. As *topic-based* categories, we select **gender bias** and **political bias** for their prominence in the media bias literature.

C.2 Selection of Concepts

When designing the bias definitions, we have three goals. First, we want to isolate conceptual content from surface features such as length, register, and example count, so that any observed rating differences can be attributed to what a definition specifies rather than how it is phrased. Second, we want each definition to be rooted in an existing research tradition, so that the concepts we compare reflect genuine theoretical divides in the literature rather than variants we invented ourselves. Third, we want the concepts within a category to be analytically independent, so that a given article can in principle exhibit one while remaining neutral on the others, letting us interpret between-definition rating differences as effects of the definition rather than of inherent article properties. To meet these goals, we follow a two-step process grounded in existing reviews of the media bias, communication, and computational linguistics literature (Wessel et al., 2023; Spinde et al., 2023; Rodrigo-Ginés et al., 2023; Spinde et al., 2025). For each of the four categories, we first collect the conceptualisations of that category used in prior work and classify each by its underlying mechanism. In all four categories, this process yields at least three separable concepts which adhere to two selection criteria:

1. **Theoretical grounding.** Each sub-concept is rooted in a distinct research tradition, ensuring that conceptual differences reflect genuine theoretical divides rather than surface-level variation.
2. **Independent variability.** The three concepts target analytically distinguishable mechanisms such that a text can in principle exhibit one while remaining neutral on the others, and the three can vary independently within a single article without logical entailment.

We note that these concepts sample from a larger definitional space rather than claiming taxonomic completeness. As discussed in Section 5.4, the significant and ordered rating gradients observed across just three concepts per category suggest that the full range of operationalizations used in the field would yield at least comparable heterogeneity.

C.3 Definition Authoring

The definition *texts* shown to participants are author-constructed. This choice ensures that definitions within each category match in length, example count, register, and readability, isolating conceptual content as the sole variable. Each sub-concept is written in a short version (one to two sentences, one example) and a long version (extended explanation, three examples), with the long version adding elaboration and additional examples without altering the target construct. We deliberately optimise the three concept definitions within each category for diversity rather than minimal contrast:

we choose examples and mechanism phrasings that sit as far apart as is defensible within the concept, so that any residual rating differences we observe are driven by the conceptual core rather than by a single contestable wording choice. If three separated operationalisations produce the divergences we report, small wording perturbations within any one of them are unlikely to erase the pattern, although they could shift absolute effect sizes.

We then subject all 24 resulting definitions to two validation steps. The authors review each definition for conceptual accuracy, length parity within a category, and register consistency. We then pilot-test the full set on Prolific ($n = 24$) using the same study design as the main experiment, supplemented with open-ended comprehensibility questions after each category. The pilot confirms that participants can reliably distinguish between concepts within each category and leads us to adopt an easy-language register throughout.

We acknowledge that even with this process, the specific wording within each sub-concept represents one defensible realisation rather than the only one. Our claims concern between-sub-concept differences, not absolute rating levels.

C.4 Rationale per category

C.4.1 Linguistic Bias

For linguistic bias, we find three levels of analysis: semantic content, syntactic form, and social labeling. We select **biased word choice** (C1), rooted in lexical framing and Entman's framing theory (Entman, 1993), targeting how adjectives, verbs, and metaphors carry implicit valence; **grammatical agency manipulation** (C2), rooted in transitivity and responsibility attribution research (Henley et al., 1995), targeting how active/passive constructions obscure or highlight who acts; and **demeaning or stigmatizing labels** (C3), rooted in labeling theory and dangerous speech research (Leader Maynard and Benesch, 2016), targeting how identity-linked terms marginalize groups.

C.4.2 Gender Bias

Gender bias definitions in the literature organize around three stages of representation. We select **representational visibility** (C1), rooted in media representation and gender sourcing research (Macharia, 2020), targeting systematic over- or under-representation in expert roles and quotes; **stereotypical trait attribution** (C2), rooted in the Stereotype Content Model (Fiske et al., 2018) and gender schema theory, targeting how identical behavior is framed differently by gender; and **irrelevant personal focus** (C3), rooted in objectification theory (Fredrickson and Roberts, 1997), targeting the foregrounding of appearance, family status, or likeability when irrelevant to professional role.

C.4.3 Political Bias

Political bias definitions map onto functions that operate at different stages of news production. We select **evaluative slant** (C1), rooted in partisan slant and tone analysis research (Groseclose and Milyo, 2005), targeting explicit favorability through adjectives and verdict-like conclusions; **selective issue coverage** (C2), rooted in agenda-setting theory (McCombs and Shaw, 1972) and gatekeeping research, targeting which issues are spotlighted or suppressed for each side; and **interpretive framing** (C3), rooted in framing theory (Entman, 1993) and interpretive journalism research (Iyengar, 1994), targeting how identical facts are cast as, e.g., "success" versus "failure."

C.4.4 Text-Level Context Bias

Context bias definitions cluster around modes of informational slant that correspond to omission, asymmetry, and addition. We select **omission of relevant facts** (C1), rooted in information omission research (Entman, 2007) and journalism ethics, targeting facts a reader would need to interpret the central claim fairly; **asymmetric source treatment** (C2), rooted in source balance and false balance research (Boykoff and Boykoff, 2004), targeting whether opposing sides receive comparable space, quotes, and argumentative weight; and **gratuitous emotional detail** (C3), rooted in emotional priming and irrelevant information research (Schwarz et al., 1991), targeting background details included not for informational value but to steer the reader's emotional response.

C.5 Full Definition Texts

Tables 4–7 present all 24 bias definitions used in the experiment. Each of the four bias categories comprises three conceptual sub-definitions (C1–C3), each in a short version (S: one to two sentences, one example) and a long version (L: extended explanation, three examples). Definition texts are shown as presented to participants.

Table 4: Linguistic bias: complete definition texts.

Sub-concept	Ver.	Definition text
Biased word choice (C1)	S	Linguistic bias is when an article's word choice (adjectives, verbs, labels, metaphors) nudges the reader toward liking or disliking a person or group, even if the basic facts could be described neutrally. *Example:* Neutral: "Police dispersed the crowd after midnight." Biased: "Police crushed a mob after midnight."
	L	Linguistic bias is when an article's word choice (adjectives, verbs, labels, metaphors) nudges the reader toward liking or disliking a person or group, even if the basic facts could be described neutrally. This can happen through a single loaded word or through many small word choices that add up across the article, slowly tilting the tone in one direction. *Example 1:* Neutral: "Police dispersed the crowd after midnight." Biased: "Police crushed a mob after midnight." *Example 2:* Neutral: "The senator questioned the proposal." Biased: "The senator attacked the proposal." *Example 3:* Neutral: "Protesters gathered outside the courthouse." Biased: "Agitators swarmed the courthouse steps."
Grammatical agency manipulation (C2)	S	Linguistic bias is when grammar (active vs. passive voice, who is named first, who is left unnamed) shifts responsibility or blame without changing the facts. *Example:* "The minister cut disability benefits by 8%." (clear) vs. "Disability benefits were cut by 8%." (decision-obscuring)
	L	Linguistic bias is when grammar (active vs. passive voice, who is named first, who is left unnamed) shifts responsibility or blame without changing the facts. This can happen when an article consistently uses sentence structures that highlight certain people as decision-makers while making others seem like bystanders, even though both played an active role. *Example 1:* "The minister cut disability benefits by 8%." (clear) vs. "Disability benefits were cut by 8%." (decision-obscuring) *Example 2:* "The company fired 200 workers before the holiday." (agency clear) vs. "200 jobs were lost before the holiday." (agency hidden) *Example 3:* "Officers shot the suspect during the chase." (responsibility assigned) vs. "The suspect was shot during the chase." (responsibility obscured)
Demeaning or stigmatizing labels (C3)	S	Linguistic bias is when an article uses demeaning, stigmatizing, or "othering" labels for people (nicknames, slurs, dehumanizing terms, mocking descriptors) instead of standard, respectful terms. *Example:* Biased: "The neighborhood is being overrun by junkies." Neutral: "The neighborhood has seen an increase in people with opioid addiction."
	L	Linguistic bias is when an article uses demeaning, stigmatizing, or "othering" labels for people (nicknames, slurs, dehumanizing terms, mocking descriptors) instead of standard, respectful terms. This also covers labels that reduce a person to one negative trait or that make certain groups sound like outsiders or threats, especially when a neutral alternative would work just as well. *Example 1:* Biased: "The neighborhood is being overrun by junkies." Neutral: "The neighborhood has seen an increase in people with opioid addiction." *Example 2:* Biased: "Illegals continue to flood across the border." Neutral: "Undocumented immigrants continue to cross the border." *Example 3:* Biased: "The tree-huggers blocked the construction site again." Neutral: "Environmental activists blocked the construction site again."

Table 5: Gender bias: complete definition texts.

Sub-concept	Ver.	Definition text
Representational visibility (C1)	S	Gender bias is when people of one gender are systematically more visible in an article: quoted more, treated as experts more often, or centered as the main actors, while others are absent or relegated to minor roles. *Example:* The article quotes six "experts" (CEO, professor, analyst, doctor, spokesperson, union leader): all men, while women appear only as "local residents" with no expert role.
	L	Gender bias is when people of one gender are systematically more visible in an article: quoted more, treated as experts more often, or centered as the main actors, while others are absent or relegated to minor roles. This is especially telling when the underrepresented gender is equally present and qualified in the field being covered, but editorial choices make it seem like expertise belongs mainly to one gender. *Example 1:* The article quotes six "experts" (CEO, professor, analyst, doctor, spokesperson, union leader): all men, while women appear only as "local residents" with no expert role. *Example 2:* A technology article interviews five male founders about industry trends while the only woman quoted is an office manager commenting on workplace culture. *Example 3:* A political analysis piece features four male commentators debating policy while the sole female voice is a voter sharing a personal story rather than analysis.
Stereotypical trait attribution (C2)	S	Gender bias is when the article assumes or reinforces gender stereotypes (who is "naturally" caring, emotional, tough, logical, suited to leadership, etc.) when describing people or evaluating their actions. *Example:* "He delivered a firm, strategic response." vs. "She responded emotionally and struggled to stay calm." (With no evidence that emotionality differed, just a gendered interpretive lens.)
	L	Gender bias is when the article assumes or reinforces gender stereotypes (who is "naturally" caring, emotional, tough, logical, suited to leadership, etc.) when describing people or evaluating their actions. This includes cases where the same behavior is described differently depending on gender: for example, calling it "confidence" in one person and "aggression" in another. *Example 1:* "He delivered a firm, strategic response." vs. "She responded emotionally and struggled to stay calm." (With no evidence that emotionality differed, just a gendered interpretive lens.) *Example 2:* "The male nurse showed unusual compassion for his patients." (implying compassion is unexpected for men) vs. "The nurse showed compassion for her patients." (implying it is natural for women) *Example 3:* "She managed to lead the negotiation successfully." (implying surprise) vs. "He led the negotiation successfully." (stated as expected)
Irrelevant personal focus (C3)	S	Gender bias is when the article focuses on appearance, sexuality, family status, or "likeability" in ways that distract from competence or actions when this focus would be unlikely for another gender in the same role. *Example:* "The lawyer, wearing a tight red dress, arrived in court…" Her legal arguments are summarized in one line, while her look gets a full paragraph.
	L	Gender bias is when the article focuses on appearance, sexuality, family status, or "likeability" in ways that distract from competence or actions when this focus would be unlikely for another gender in the same role. This means drawing attention to personal traits like looks or family life that have nothing to do with the person's job or achievements, especially when a person of another gender in the same position would not get that treatment. *Example 1:* "The lawyer, wearing a tight red dress, arrived in court…" Her legal arguments are summarized in one line, while her look gets a full paragraph. *Example 2:* "The CEO, a mother of three who still finds time to stay fit, announced the merger." A male CEO's family and fitness would rarely be mentioned in the same context. *Example 3:* "The senator, known for her warm smile and approachable demeanor, introduced the defense bill." Her male colleague's bill introduction mentions only his policy credentials.

Table 6: Political bias: complete definition texts.

Sub-concept	Ver.	Definition text
Evaluative slant (C1)	S	Political bias is when the article systematically favors or disfavors a political party or ideology through positive or negative evaluation (tone, adjectives, verdict-like conclusions), rather than sticking to balanced reporting. *Example:* "The minister's plan is a bold fix that will finally restore order." vs. "The opposition's proposal is a reckless gamble."
	L	Political bias is when the article systematically favors or disfavors a political party or ideology through positive or negative evaluation (tone, adjectives, verdict-like conclusions), rather than sticking to balanced reporting. This includes patterns where praise and criticism are handed out unevenly: one side's actions are consistently described with positive words while the other side's similar actions get negative ones. *Example 1:* "The minister's plan is a bold fix that will finally restore order." vs. "The opposition's proposal is a reckless gamble." *Example 2:* "The governor's swift action saved thousands of jobs." vs. "The mayor's hasty decision put the budget at risk." (Both describe fast action, but with opposite evaluative framing.) *Example 3:* "The party delivered on its promise to reform healthcare." vs. "The party pushed through a controversial overhaul of healthcare." (Same policy outcome, different tone.)
Selective issue coverage (C2)	S	Political bias is when political coverage repeatedly highlights certain issues or scandals for one side while downplaying or ignoring comparable issues for another side, shaping what audiences think matters. *Example:* An article highlights Party A's misuse of travel funds, but Party B's similar audit finding is mentioned briefly in a late paragraph or not followed up.
	L	Political bias is when political coverage repeatedly highlights certain issues or scandals for one side while downplaying or ignoring comparable issues for another side, shaping what audiences think matters. Within a single article, this can show up when one side's problems are described in detail while the other side's similar problems are mentioned only briefly or left out entirely. *Example 1:* An article highlights Party A's misuse of travel funds in detail, but Party B's similar audit finding is mentioned only in a single sentence in the last paragraph. *Example 2:* An article spends six paragraphs detailing one candidate's policy contradictions with quotes and timelines, while a rival candidate's comparable flip-flop is summarized in one line without evidence. *Example 3:* An article thoroughly documents lobbying ties of one party's officials with names and amounts, while the other party's similar connections are acknowledged only as "both sides have faced scrutiny."
Interpretive framing (C3)	S	Political bias is when the article frames political events so that readers are pushed toward one interpretation (e.g., "success" vs. "failure," "defense" vs. "attack"), even if the same facts could support multiple reasonable readings. *Example:* "The government finally caved and reduced the fee." (implies weakness) vs. "The government responded to public concerns and reduced the fee." (implies responsiveness)
	L	Political bias is when the article frames political events so that readers are pushed toward one interpretation (e.g., "success" vs. "failure," "defense" vs. "attack"), even if the same facts could support multiple reasonable readings. This goes beyond single words to how the overall story is told: the order of information and the explanations chosen can make the reader lean toward one conclusion when other conclusions are just as reasonable. *Example 1:* "The government finally caved and reduced the fee." (implies weakness) vs. "The government responded to public concerns and reduced the fee." (implies responsiveness) *Example 2:* "The president was forced to abandon the policy after mounting pressure." (implies defeat) vs. "The president revised the policy in light of new evidence." (implies rational adaptation) *Example 3:* "The opposition blocked the bill, plunging the legislature into gridlock." (implies obstruction) vs. "The opposition exercised its oversight role by challenging the bill." (implies democratic function)

Table 7: Text-level context bias: complete definition texts.

Sub-concept	Ver.	Definition text
Omission of relevant facts (C1)	S	Context bias is when an article leaves out important facts or constraints that a reasonable reader would need to interpret the main claim fairly (even if everything stated is technically true). *Example:* Article says: "City spending on 'administration' rose 20% this year." Missing context: The budget category was renamed/reclassified, so the increase is mostly an accounting change, not new spending.
	L	Context bias is when an article leaves out important facts or constraints that a reasonable reader would need to interpret the main claim fairly (even if everything stated is technically true). This also covers missing background like how the data was collected, historical context, or exceptions to a rule that would change how a reader judges the claim. *Example 1:* Article says: "City spending on 'administration' rose 20% this year." Missing context: The budget category was renamed/reclassified, so the increase is mostly an accounting change, not new spending. *Example 2:* Article says: "Crime in the district increased by 15% this quarter." Missing context: The police department changed its reporting method to include incidents previously counted differently. *Example 3:* Article says: "The drug was found to be effective in clinical trials." Missing context: The trial had only 30 participants, lasted two weeks, and was funded by the manufacturer.
Asymmetric source treatment (C2)	S	Context bias is when an article mostly presents one side's reasons, quotes, or evidence while giving the other side little space, weaker paraphrases, or no serious explanation, creating a tilted overall picture. *Example:* 8 paragraphs: detailed quotes from one side ("We have three reasons. . . ," data, anecdotes). 1 sentence for the other: "Critics disagreed," with no reasons, no data, no quote.
	L	Context bias is when an article mostly presents one side's reasons, quotes, or evidence while giving the other side little space, weaker paraphrases, or no serious explanation, creating a tilted overall picture. This imbalance shows not just in how much space each side gets but also in how their arguments are presented: one side gets direct quotes and named sources while the other is summed up in a single vague sentence. *Example 1:* 8 paragraphs: detailed quotes from one side ("We have three reasons. . . ," data, anecdotes). 1 sentence for the other: "Critics disagreed," with no reasons, no data, no quote. *Example 2:* A healthcare article quotes three named doctors supporting a treatment with detailed explanations while opposition is summarized as "some experts have raised concerns" without naming anyone or explaining what the concerns are. *Example 3:* A labor dispute article gives the union representative two full paragraphs of direct quotes with statistics while the company's response is paraphrased in a single sentence as "the company denied the allegations."
Gratuitous emotional detail (C3)	S	Context bias is when an article adds background details that are not necessary to understand the event but are likely to shape the reader's emotion or moral judgment (sympathy, anger, disgust), steering interpretation. *Example:* About a consumer complaint lawsuit, the article includes: "The plaintiff has filed for bankruptcy twice and was once evicted," without explaining why this is relevant to the specific complaint.
	L	Context bias is when an article adds background details that are not necessary to understand the event but are likely to shape the reader's emotion or moral judgment (sympathy, anger, disgust), steering interpretation. This includes personal history or past events that have nothing to do with the story but make the reader feel a certain way about the people involved, especially when such details are only added for some people but not others. *Example 1:* About a consumer complaint lawsuit, the article includes: "The plaintiff has filed for bankruptcy twice and was once evicted," without explaining why this is relevant to the specific complaint. *Example 2:* In reporting a traffic accident, the article mentions: "The driver, who had a prior conviction for petty theft ten years ago, collided with the barrier." The prior conviction has no relevance to the accident. *Example 3:* In covering a whistleblower's complaint, the article notes: "The employee, who went through a contentious divorce last year, filed the report on Monday." The personal history has no bearing on the validity of the complaint.