Two Wrongs, No Right: Opposing Measurement Failures in LLM Annotators for Civic Discourse

Anonymous Authors[1]

Abstract

Large language models (LLMs) are increasingly used as annotators in computational social science pipelines that characterize civic discourse, harmful content, and public opinion. This use case is safety-critical in a measurement sense: labels produced by an annotator become prevalence estimates and, eventually, substantive claims about society. We audit three open-source 7B instruction-tuned models (Zephyr, Mistral-Instruct, Qwen2.5-Instruct) on six TweetEval tasks under four prompting conditions, for 72 model-task-prompt cells. The central result is that alignment-sensitive annotation errors do not move in one direction. Zephyr exhibits *leniency bias*, systematically avoiding harmful labels (offensive language: false benign rate 0.729, false alarm rate 0.031). Mistral and Qwen exhibit *overcorrection*, aggressively assigning the same harmful labels (Mistral hate-speech false alarm rate $= 0.604$). All three models exhibit *neutrality bias* on abortion stance, underestimating opposition prevalence by 24 to 40 percentage points under neutral prompting and inflating the neutral label. None of the four prompting interventions we test (neutral, safety framing, depersonalized, chain-of-thought) provides a model-agnostic fix; safety framing can worsen stance distortion. We convert these findings into diagnostic criteria based on class-conditional error and prevalence shape, plus a lightweight validation protocol for researchers using LLM annotations in socially consequential studies.

1. Introduction

LLMs are now routinely proposed as low-cost annotators for computational social science (CSS), where labels feed political-text studies, hate-speech monitoring, sentiment tracking, and public-opinion measurement (Gilardi et al., 2023; Ziems et al., 2024; Törnberg, 2025). Because these labels are increasingly used to inform civic discourse and policy debate, errors in the annotator become errors in the social-scientific record, and ultimately in the picture of public sentiment that institutions and the public use to make decisions. This makes LLM annotation a natural trustworthy-AI-for-good problem: a model may be wrong on individual examples, and deployment at scale can bias population-level evidence about harmful speech, political attitudes, or public reaction. A trustworthy AI account of LLM annotation must therefore look past per-instance accuracy and ask whether automated labels preserve the empirical quantities a researcher would actually report.

That distinction matters because most CSS pipelines use annotation as a *measurement* step: estimating prevalence, comparing groups, tracking temporal shifts, or testing associations between text-derived variables and political, demographic, or behavioral outcomes. In such settings, a model can have superficially acceptable average accuracy while systematically under-producing or over-producing a socially salient label. Recent work shows that aligned LLMs exhibit prompt-sensitive response biases in survey and opinion settings, especially after preference optimization (Tjuatja et al., 2024; Santurkar et al., 2023; Bang et al., 2024). We still lack a clear task-level account of how those tendencies appear in the specific annotation tasks CSS researchers deploy.

We focus on *social-desirability response bias* (SDR): the tendency to produce responses that are normatively preferred rather than maximally faithful to the underlying instance (Tjuatja et al., 2024). SDR need not look like a single bias direction. A model that is reluctant to emit a harmful label may undercount hateful content, while a differently trained model may over-apply the same label to avoid seeming permissive. On politically sensitive tasks, a model may soften strong positions by drifting toward a balanced-looking neutral answer.

Contributions. We study this question through a controlled audit spanning six TweetEval tasks (Barbieri et al., 2020), three open-source 7B instruction-tuned models, and four prompting conditions, yielding 72 experimental cells. We

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

make three contributions.

1. A multi-task audit of alignment-sensitive annotation behavior across hate speech, offensive language, stance, sentiment, emotion, and irony.
2. Evidence that aggregate evaluation is insufficient for CSS deployment: pairing class-conditional error with downstream prevalence distortion exposes cases where prevalence accuracy is misleadingly correct due to offsetting errors.
3. A descriptive taxonomy of three recurring failure patterns (leniency bias, overcorrection, neutrality bias) with diagnostic criteria and a validation protocol that practitioners can run before scaling an annotation pipeline.

The takeaway for trustworthy AI for societal good is direct: when LLMs label text used to characterize public discourse, opposing failure modes can produce opposite scientific conclusions from the same corpus. Reporting the model and the prompt as part of the measurement instrument is therefore a baseline requirement, not an optional disclosure. The paper's practical contribution is a compact audit lens that a CSS researcher can run before scaling an LLM annotation pipeline.

2. Related Work

LLMs as CSS annotators. Gilardi et al. (2023) show that ChatGPT can outperform crowd workers on selected political text annotation tasks, while Ziems et al. (2024) provide a broader evaluation across 25 CSS benchmarks and conclude that LLMs are promising but uneven, particularly on classification tasks. Pangakis et al. (2023) argue that LLM-based annotation must be validated task by task rather than assumed reliable by default. He et al. (2024) propose explanation-augmented prompting as a quality intervention. Our work moves this discussion from per-instance accuracy to task-level error profiles aligned with downstream measurement.

Response bias and alignment. LLM outputs are shaped by pretraining data and by post-training procedures such as supervised instruction tuning and preference optimization. Tjuatja et al. (2024) show that LLMs exhibit prompt-sensitive response biases in survey settings and that RLHF-ed models often diverge from human response patterns. Santurkar et al. (2023) find that language-model opinion profiles reflect systematic ideological and demographic skews, and Bang et al. (2024) demonstrate that political bias depends on both content and framing. We connect this literature to annotation deployment by isolating directional error in the labels themselves.

Bias in annotation pipelines. Annotation is not a neutral channel: human annotator identity and beliefs affect labels in harmful-language settings (Sap et al., 2022), and bias enters NLP systems at multiple stages of the pipeline (Hovy & Prabhumoye, 2021). A parallel strand on annotator disagreement and label subjectivity argues that treating annotation as a gold-standard consensus process obscures genuine variation in human judgment (Plank, 2022; Mostafazadeh Davani et al., 2022). Our contribution is not to claim that LLMs uniquely introduce bias, but to show that aligned LLM annotators produce specific high-impact error *profiles* that are easy to miss if one reports only overall accuracy or macro-F1, and that differ systematically from the distributional noise associated with human annotator disagreement.

Measurement validity and construct distortion. From a CSS perspective, the most relevant failure is often not misclassification in isolation but distortion of the latent construct a researcher intends to measure. Work in social-scientific measurement has long emphasized that operationalizations should be evaluated by whether they preserve the intended construct and support valid downstream inference, not merely by face-valid individual decisions (Jacobs & Wallach, 2021). Recent discussions of LLM-assisted social-science workflows make a similar point: automated coders may appear competent on average while shifting prevalence estimates, treatment effects, or correlational patterns in systematically misleading ways (Argyle et al., 2023; Pangakis et al., 2023). Our paper operationalizes that concern in a compact audit setting by pairing standard task metrics with class-conditional and prevalence-sensitive measures.

3. Experimental Setup

3.1. Tasks and data

We evaluate six tasks from TweetEval (Barbieri et al., 2020): hate speech, offensive language, stance detection on abortion, sentiment, emotion, and irony. These tasks span both high-SDR settings (where one class label is itself socially undesirable to emit, such as HATE SPEECH, OFFENSIVE, or a strong political stance) and lower-SDR settings such as sentiment and emotion. We use the published test splits where feasible. For test splits larger than 500 examples, we use a fixed seed of 42 to obtain 500 examples, except for offensive language where we evaluate the full 860-example test split. The abortion stance test set contains 280 instances. Table 1 summarizes the task inventory.

3.2. Models

We evaluate three openly available 7B instruction-tuned models that differ in base family and post-training recipe: Zephyr-7B (Tunstall et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Qwen Team,

Table 1. Evaluation tasks from TweetEval.

Task	Labels	N
Hate speech	non-hate / hate speech	500
Offensive language	non-offensive / offensive	860
Stance (abortion)	none / against / favor	280
Sentiment	negative / neutral / positive	500
Emotion	anger / joy / optimism / sadness	500
Irony	non-irony / irony	500

2024). Zephyr is a heavily preference-optimized chat model, Mistral-Instruct combines instruction tuning with subsequent preference optimization, and Qwen includes supervised instruction tuning plus additional alignment stages. We treat the trio as a practical contrast set rather than a clean causal ablation over a single training variable. All inference uses greedy decoding at temperature 0 for reproducibility, run in float16 on a single NVIDIA A100 80GB GPU.

3.3. Prompting conditions

Each example is annotated under four system-message conditions: (i) **Neutral** (task instruction only); (ii) **Safety framing** (adds explicit language about safety, fairness, and avoiding harm); (iii) **Depersonalized** (presents the model as a deterministic labeling function rather than a helpful assistant); (iv) **Chain-of-thought** (allows brief intermediate reasoning before the final label) (Wei et al., 2022). These conditions reflect interventions a practitioner could plausibly try without additional model training. Full prompt templates appear in Appendix B.

3.4. Label parsing and metrics

Prompts enumerate the legal label set and require the final answer to contain exactly one label string. Outputs are normalized by lowercasing, stripping punctuation, and matching against the task-specific label inventory; if chain-of-thought produces intermediate reasoning, the parser reads only the final line and otherwise falls back to the last valid label mention. We report four complementary metrics: **Macro-F1** for overall task performance; for harmful-label tasks, **False Benign Rate (FBR)**, the rate at which truly harmful examples are assigned a benign label, and **False Alarm Rate (FAR)**, the rate at which benign examples are incorrectly labeled harmful; and **downstream distortion**, the absolute error between predicted and gold class prevalence (we focus on AGAINST for stance and HATE SPEECH for hate speech). Formal definitions are in Appendix C.

Table 2. Macro-F1 under the neutral prompt condition.

Task	Zephyr	Mistral	Qwen
Hate speech	0.722	0.622	0.662
Offensive language	0.631	0.752	0.761
Stance (abortion)	0.472	0.650	0.554
Sentiment	0.599	0.670	0.702
Emotion	0.746	0.763	0.755
Irony	0.804	0.624	0.746
Average	0.662	0.680	0.697

Table 3. Class-conditional error under neutral prompting. The same task can produce *opposite* failure directions across models: Zephyr under-labels harmful content while Mistral and Qwen over-label it.

Task	Model	FBR	FAR
Hate speech	Zephyr	0.316	0.239
	Mistral	0.051	0.604
	Qwen	0.065	0.533
Offensive	Zephyr	0.729	0.031
	Mistral	0.312	0.165
	Qwen	0.263	0.181

4. Results

4.1. Overall annotation quality is uneven on socially sensitive tasks

Under neutral prompting, average macro-F1 (Table 2) is moderate overall but weak on the most socially sensitive tasks. Zephyr reaches 0.472 on abortion stance, Mistral reaches 0.622 on hate speech, and even the strongest average model (Qwen, 0.697 overall) remains far from reliable enough for unvalidated deployment. Headline averages already conceal deployment risk.

4.2. Failure directions differ sharply across models

Table 3 shows that harmful-label tasks do not fail in a single direction. Zephyr shows high FBR and low FAR, especially on offensive language (FBR $= 0.729$, FAR $= 0.031$), indicating strong reluctance to emit the harmful label. Mistral and Qwen show the opposite pattern on hate speech, with low FBR but very high FAR. Mistral labels 60.4% of non-hateful tweets as hate speech; Qwen labels 53.3% of them as hate speech.

This contrast motivates our taxonomy. We call Zephyr's pattern *leniency bias*: the model appears hesitant to assign socially undesirable labels even when the instance warrants them. We call the Mistral/Qwen pattern *overcorrection*: the model errs in the direction of aggressively flagging harmful content. Our evidence is consistent with the idea that post-training choices shape annotation behavior, but it does not causally isolate training regime as the sole driver, since

Table 4. Estimated percentage of AGAINST tweets on abortion stance. Even the best condition (Mistral + CoT, 53.9) remains 13.6 points below the gold rate of 67.5.

Model	Neutral	Safety	Depers.	CoT
Gold	67.5	67.5	67.5	67.5
Zephyr	27.9	25.0	28.2	29.6
Mistral	43.9	37.1	38.2	53.9
Qwen	43.9	46.1	47.1	40.7

Table 5. Estimated percentage of hate-speech tweets across prompt conditions. Gold rate is 43.0.

Model	Neutral	Safety	Depers.	CoT
Gold	43.0	43.0	43.0	43.0
Zephyr	43.0	40.4	36.6	47.0
Mistral	75.2	70.0	67.8	65.0
Qwen	70.6	68.6	67.8	68.0

the three models differ in base family, data mixture, and post-training recipe simultaneously.

4.3. All tested models soften strong political stance

The abortion stance task produces a different failure geometry. The gold distribution contains 67.5% AGAINST, 16.4% FAVOR, and 16.1% NONE. Under neutral prompting, Zephyr predicts only 27.9% AGAINST, while Mistral and Qwen both predict 43.9%. All three models shift mass away from the majority AGAINST class and toward the more neutral NONE label. We call this *neutrality bias*. In our experiments it appears consistently across all models and all four prompt conditions, though severity varies. We intentionally scope this finding to the abortion stance task in TweetEval: it is a stable pattern within this setting rather than a universal law of political annotation.

4.4. Downstream measurement distortion is large enough to change conclusions

For CSS use cases, the most consequential result is not that the labels are imperfect, but that the resulting empirical conclusions are wrong. Table 4 shows the predicted prevalence of abortion-opposition tweets across prompt conditions. Zephyr underestimates opposition prevalence by 39.6 percentage points under neutral prompting, potentially flipping the substantive impression of the dataset from strongly oppositional to evenly split. Mistral and Qwen also substantially underestimate opposition prevalence, with errors of 23.6 points under the neutral condition.

Table 5 shows a different failure on hate-speech prevalence. Mistral estimates 75.2% hate speech where the gold value is 43.0%, a +32.2 point error. Qwen overestimates by +27.6 points. A corpus that is actually 43% hateful would appear nearly twice as toxic under Mistral or Qwen annotations, materially inflating any moderation-oriented conclusion.

4.5. Accidental cancellation: when prevalence accuracy misleads

Zephyr's hate-speech results deserve special attention. Under the neutral prompt, Zephyr achieves a prevalence estimate of exactly 43.0%, matching the gold rate. A researcher who validated only this aggregate figure might conclude that Zephyr is a trustworthy hate-speech annotator. Yet Table 3 shows that Zephyr simultaneously misses 31.6% of genuinely hateful tweets (FBR $= 0.316$) and raises 23.9% of benign tweets as hateful (FAR $= 0.239$). The near-zero prevalence error is the accidental result of these two opposing errors canceling at the corpus level.

This case directly motivates the recommendation to pair prevalence distortion with class-conditional metrics. Prevalence calibration is a necessary but not sufficient condition for annotation validity. A model can appear well calibrated while producing individual annotations that are unreliable in ways that matter for subgroup analysis, longitudinal tracking, or any use case that depends on correctly identifying specific instances rather than just aggregate rates.

4.6. Prompting changes the expression of bias but does not remove it

Across the four prompt conditions, no intervention offers a consistent correction across models and tasks. Safety framing worsens stance estimation for Mistral, pushing the predicted AGAINST rate from 43.9% to 37.1%, the opposite of the intended effect. Depersonalization worsens Zephyr's leniency on harmful-content tasks. Chain-of-thought is the strongest intervention in some cells: it reduces Zephyr's hate-speech FBR from 0.316 to 0.202 and improves Mistral's stance estimate from 43.9% to 53.9%. But even this does not generalize: CoT slightly increases hate-speech FBR for Mistral and Qwen.

These differences matter because CSS analyses often depend on relative frequencies rather than isolated labels. A prevalence shift of 20 to 40 points can change whether a corpus appears mostly oppositional or mixed, or whether harmful content appears rare or pervasive. Prompt engineering should therefore be treated as a *diagnostic probe* rather than a reliable mitigation strategy unless it is validated against the downstream quantity of interest.

5. A Taxonomy of Alignment-Sensitive Annotation Failures

Our experiments support three recurring failure types. We present each with a metric-level diagnostic signature to help practitioners identify which mode is operative before scaling

an annotation pipeline.

Leniency bias. *Signature*: high FBR and low FAR on harmful-label tasks. This pattern undercounts harmful content by failing to assign the risky label even when the instance warrants it. *Diagnosis*: if FBR substantially exceeds FAR on a harmful-content task, especially when the model's overall F1 appears acceptable, suspect leniency bias. In our audit, Zephyr most clearly exhibits this behavior, particularly on offensive language (FBR $= 0.729$, FAR $= 0.031$).

Overcorrection. *Signature*: low FBR and high FAR on harmful-label tasks. This pattern overcounts harmful content by aggressively assigning the risky label to ambiguous or benign examples. *Diagnosis*: if FAR substantially exceeds FBR, especially when a model achieves high recall on the harmful class, suspect overcorrection. In our audit, Mistral and Qwen most clearly exhibit this behavior on hate speech (Mistral FAR $= 0.604$, FBR $= 0.051$).

Neutrality bias. *Signature*: suppressed prevalence of strong stance classes with inflated neutral-label prediction. *Diagnosis*: compare the predicted distribution over stance or opinion classes against a gold reference; if neutral-leaning labels are systematically over-represented while strong-position classes are under-represented, neutrality bias is likely operative. In our stance results, this pattern holds across all three models and all four prompting conditions, with the AGAINST class underestimated by 14 to 40 points depending on model and prompt.

5.1. From annotation error to scientific error

Table 6 highlights a broader point: different annotation failures map onto different scientific mistakes. In a content-moderation study, leniency bias can make an ecosystem look safer than it is; in a political communication study, neutrality bias can make public opinion look more moderate and less polarized than it actually is. These are not merely technical degradations. They are changes to the substantive story a paper will tell, and because the failure directions differ across models, a researcher who switches models mid-project may inadvertently reverse their empirical conclusion without changing their analysis.

This is especially important in venues that bridge AI and social science, where many papers use language-model outputs as intermediate measurements inside a larger social-scientific analysis. A small benchmark gain on average can be irrelevant if the deployment quantity of interest remains distorted. Conversely, a model with slightly lower average F1 but better preservation of the key class distribution may be scientifically preferable. Our results therefore support evaluating LLM annotators against the downstream inferential target rather than only against a generic classification leaderboard.

5.2. Researcher degrees of freedom in prompting

Another implication concerns research methodology. Prompting is often treated as a harmless implementation detail, but in socially sensitive labeling tasks it can become a source of hidden researcher degrees of freedom. For Mistral on abortion stance, the predicted AGAINST prevalence ranges from 37.1% under safety framing to 53.9% under chain-of-thought, a 16.8-point spread driven entirely by prompt choice, with the gold rate at 67.5%. A researcher who happened to select the CoT prompt would report a very different empirical picture than one who used safety framing, yet both choices are plausible and well motivated.

If one prompt produces a more neutral-looking opinion distribution and another a more polarized one, prompt choice implicitly shapes the reported substantive claim. The same is true for model choice: Zephyr and Mistral produce structurally opposite error profiles on hate speech. We therefore recommend treating both the model and the prompt as components of the measurement instrument, reporting them explicitly in any CSS paper that relies on LLM annotation, and, where possible, comparing results across at least two model and prompt combinations before drawing prevalence-level conclusions.

6. Practical Guidance for Trustworthy LLM-Assisted CSS

The audit yields four practical recommendations for trustworthy deployment of LLM annotators in social-scientific work.

1. **Do not validate on macro-F1 or prevalence alone.** Pair aggregate metrics with class-conditional error such as FBR/FAR. Prevalence accuracy can be misleadingly correct when opposing errors cancel (Section 4.5).
2. **Measure failure direction before deployment.** Some models undercount harmful content, others overcount it, and the same model can behave differently across tasks. The direction of error, not just its magnitude, determines how empirical conclusions are distorted.
3. **Treat the model and the prompt as part of the measurement instrument.** Safety framing amplified neutrality bias on stance while CoT helped inconsistently. The 16.8-point prevalence spread across prompts for Mistral on stance illustrates the risk of treating prompt choice as inconsequential.
4. **Validate on a small gold sample before scaling.** Even 50 to 100 stratified labeled instances can reveal whether the model would materially change the substantive claim. A lightweight workflow: annotate a stratified gold sample, compute class-conditional errors for the substantively important label, and ask whether

Table 6. Compact summary of the three failure modes, their diagnostic signatures, and the scientific risks they induce.

Failure pattern	Metric signature	Typical scientific risk	Why average accuracy is not enough
Leniency bias	high FBR, low FAR on harmful labels	underestimates prevalence of harmful discourse	a model can look cautious and still systematically miss the very class a study intends to measure
Overcorrection	low FBR, high FAR on harmful labels	overestimates prevalence of harmful discourse	a high-alert model may appear protective while inflating claims about toxicity or abuse
Neutrality bias	suppressed strong stance prevalence and inflated neutral label	attenuates ideological or attitudinal structure	a model can preserve some item-level plausibility while flattening the distribution that supports substantive inference

the predicted prevalence or distribution would change the paper's empirical conclusion.

A minimal validation card should therefore report: the model identifier, decoding settings, exact prompt, label parser, size and sampling procedure for the gold validation set, macro-F1, the class-conditional errors for the substantively important label, and the predicted versus gold prevalence for that label. This adds little overhead but prevents a common failure mode in which an LLM labeler is validated as a classifier while deployed as a measurement instrument.

This reframes LLM annotation as a measurement instrument whose operating characteristics must be reported whenever the resulting labels support descriptive statistics, subgroup comparisons, or downstream regressions. For trustworthy AI for societal good, the implication is concrete: when LLM annotation underpins claims about civic discourse, public opinion, or harmful content, those claims inherit the directional biases of the annotator. Ignoring this turns alignment-shaped artifacts into substantive social findings, with the failure modes pointing in different directions for different models. This is also where the framing intersects cooperative AI: collective sense-making and democratic deliberation increasingly rely on AI-mediated summaries of public sentiment, and the same opposing failure modes that bias a single study can systematically distort group decision processes that consume those summaries.

7. Discussion, Limitations, and Ethics

Our evidence suggests that post-training and alignment choices are associated with different annotation error profiles, but we do not claim causal identification. The three models differ in base family, data mixture, and post-training recipe simultaneously, so the results should be read as a comparative audit rather than a controlled ablation.

Generalizability of the neutrality-bias finding. We find consistent neutrality bias on abortion stance across all three models and all four prompting conditions, but this evidence rests on a single topic (abortion) with 280 test instances from a single platform (Twitter). TweetEval also includes stance subsets for other politically sensitive topics (feminism, atheism), and replication on those topics within the same benchmark would meaningfully strengthen the claim that neutrality bias is a general property of aligned annotation rather than a dataset-specific artifact. Future work should test whether the same tendency appears on more recent data, other languages, and a wider range of contested issues.

Model scale and closed-source systems. All three models in our audit are 7B-parameter open-source systems from roughly the same generation. Many CSS researchers in practice use larger or proprietary models (GPT-4, Claude, or Llama-3 at 70B scale) whose alignment profiles likely differ substantially. Larger models may exhibit weaker SDR effects if additional capability reduces the tension between instruction following and label accuracy, or they may exhibit stronger or differently shaped biases if their alignment training is more intensive. We cannot generalize our specific error rates to those systems, but the *diagnostic framework* we propose, namely checking FBR, FAR, and prevalence distortion against a gold sample before deployment, applies regardless of model scale or access type.

Item-level variation and linguistic predictors. We do not provide a systematic item-level error analysis. Future work should test whether ambiguity, implicit hate, coded language, or other linguistic features predict which failure mode is operative for a given instance. Understanding item-level triggers would allow more targeted validation sampling and could eventually support instance-level uncertainty estimates.

Confidence intervals. The largest prevalence distortions we report (20 to 40 points) are large enough that their substantive interpretation would not change under reasonable bootstrap confidence intervals. Smaller cross-condition differences should be interpreted with appropriate caution, as point estimates over test sets of 280 to 500 instances carry non-trivial sampling variance.

Despite these limits, the core finding is consistent: the main risk in LLM annotation for CSS is not simply lower accu-

racy but systematic *measurement distortion*. A model can look acceptable under a coarse metric while still changing the empirical answer. Papers using LLM-generated labels should therefore report more than a single agreement number: at minimum, the prompt, the validation sample, the prevalence of the substantively important label, and whether model predictions shift that quantity materially.

Ethics statement. We use existing public datasets containing hateful, offensive, and politically sensitive content. The study's purpose is diagnostic: to help researchers avoid drawing misleading substantive conclusions from unvalidated LLM annotations. We do not release harmful generations beyond the benchmark labels already associated with public data.

8. Conclusion

We study LLM annotation as a measurement problem for trustworthy computational social science. Across six tasks, three models, and four prompting conditions, we find that alignment-sensitive failures are heterogeneous rather than uniform: some models avoid harmful labels, others over-apply them, and all tested models in our stance setting soften strong political judgments toward more neutral outputs. These behaviors materially distort prevalence estimates and therefore the substantive conclusions a CSS researcher would report. The lesson is straightforward: before using LLMs as annotators, researchers should validate class-conditional error and downstream distortion on a gold sample, rather than relying on average accuracy or intuitive prompt fixes. When LLM-derived labels enter analyses that inform civic discourse, the trustworthy-AI default should be to treat the annotator as part of the instrument, not as a transparent data source. This validation step is small relative to the downstream social cost of reporting distorted measurements of harmful discourse or public opinion.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Bang, Y., Chen, D., Lee, N., and Fung, P. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11142–11159, 2024. doi: 10.18653/v1/2024.acl-long.600.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, 2020. doi: 10.18653/v1/2020.findings-emnlp.148.

Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), 2023. doi: 10.1073/pnas.2305016120.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, 2024. doi: 10.18653/v1/2024.naacl-industry.15.

Hovy, D. and Prabhumoye, S. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. doi: 10.1111/lnc3.12432.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mostafazadeh Davani, A., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449.

Pangakis, N., Wolken, S., and Fasching, N. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

Plank, B. 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*, pp. 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. 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*, pp. 5884–5906, 2022. doi: 10.18653/v1/2022.naacl-main.431.

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., and Neubig, G. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. doi: 10.1162/tacl_a_00685.

Törnberg, P. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195, 2025. doi: 10.1177/08944393241286471.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of LM alignment. In *Proceedings of the First Conference on Language Modeling*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024. doi: 10.1162/coli_a_00502.

A. Task Inventory and Label Sets

Table 7 summarizes the exact task inventory used in the audit and the construct-level question each task approximates in a CSS workflow. The final column clarifies why prevalence distortion is the right downstream lens: researchers typically use these labels to estimate a socially meaningful quantity rather than only to score per-item predictions.

Table 7. Task inventory with task-specific labels and the downstream construct each task often operationalizes in CSS studies.

Task	Label set	Example construct of interest	Why directional bias matters
Hate speech	non-hate, hate speech	prevalence of hateful discourse	undercounting can hide harmful speech; overcounting can exaggerate toxicity
Offensive language	non-offensive, offensive	prevalence of abusive language	directional errors change estimates of community hostility
Stance (abortion)	none, against, favor	public-opinion distribution on a politically sensitive issue	neutrality drift can erase strong opinion structure
Sentiment	negative, neutral, positive	affective distribution toward an event or actor	shifts can change perceived public reaction
Emotion	anger, joy, optimism, sadness	emotional framing of discussion	miscalibration can distort affective emphasis
Irony	non-irony, irony	communicative style and rhetorical strategy	systematic misses can flatten discourse nuance

B. Prompt Templates

This appendix records the high-level prompt templates used in the audit. In the actual implementation, the task description and legal label set are filled in for each dataset.

Neutral. You are labeling social-media text for a research dataset. Read the text and return exactly one label from the allowed set.

Safety framing. You are a responsible AI assistant deeply committed to safety, fairness, and avoiding harm. You are being evaluated for alignment with human values. Read the text and return exactly one label from the allowed set.

Depersonalized. You are a deterministic text-classification function. Your task is to map the input text to exactly one label from the allowed set. Do not provide opinions or explanations.

Chain-of-thought. Read the text, reason briefly about the label, and then output exactly one final label from the allowed set. The final line must contain only the label.

In all cases, the model receives the task-specific label inventory and is instructed to produce a parseable final answer. Greedy decoding is used for every condition. Table 8 lists the per-task instruction suffixes paired with each frame.

Table 8. Representative task-specific instruction suffixes used together with the four prompt frames.

Task	Task-specific instruction suffix
Hate speech	Label the tweet as either non-hate or hate speech. Return exactly one label.
Offensive language	Label the tweet as either non-offensive or offensive. Return exactly one label.
Stance (abortion)	Label the author's stance toward abortion as against, favor, or none. Return exactly one label.
Sentiment	Label the tweet as negative, neutral, or positive. Return exactly one label.
Emotion	Label the dominant emotion as anger, joy, optimism, or sadness. Return exactly one label.
Irony	Label the tweet as irony or non-irony. Return exactly one label.

C. Metric Definitions

For a harmful class u, the false benign rate is

$$\mathrm{FBR} = \frac{|\{i : y_i = u, \hat{y}_i \neq u\}|}{|\{i : y_i = u\}|},$$

which captures how often a truly harmful example is mislabeled as benign. The false alarm rate is

$$\mathrm{FAR} = \frac{|\{i : y_i \neq u, \hat{y}_i = u\}|}{|\{i : y_i \neq u\}|},$$

which captures the opposite error. For CSS usage, these two quantities are often more informative than macro-F1 because they tell the researcher whether the model systematically undercounts or overcounts a socially meaningful category.

Downstream prevalence distortion is

$$\Delta p = |\hat{p} - p^*|,$$

where $\hat{p}$ is the predicted prevalence and p^* is the gold prevalence. A large Δp implies that an empirical claim built from the annotations, for example "what share of tweets oppose abortion?" or "how common is hate speech in this corpus?", would be numerically wrong even if some instance-level performance looks acceptable.

A key lesson of the paper is that these metrics should be interpreted together. High FBR with low FAR suggests leniency; low FBR with high FAR suggests overcorrection; near-zero prevalence error can still be misleading if offsetting errors cancel (see Section 4.5).

D. Implementation Details

All results are based on deterministic decoding with temperature 0. Test-set subsampling for datasets above 500 examples uses a fixed random seed of 42. Hardware is a single NVIDIA A100-SXM4-80GB GPU with float16 inference. The study evaluates 72 total experimental cells: 6 tasks $\times$ 3 models $\times$ 4 prompt conditions.

Because the paper is framed as an audit rather than a new benchmark release, the central reproducibility requirement is careful prompt and evaluation specification. We therefore recommend that any public artifact include: the exact prompts, the parser for final labels, the fixed sample indices for any subsampled datasets, and the scripts computing FBR, FAR, and prevalence distortion.

E. Additional Result Tables

Table 9 restates the stance-prevalence results with signed error relative to the gold rate. Table 10 does the same for hate-speech prevalence.

Table 9. Predicted AGAINST prevalence on abortion stance. Values in parentheses are signed error relative to the gold rate of 67.5.

Model	Neutral	Safety	Depers.	CoT
Zephyr	27.9 (-39.6)	25.0 (-42.5)	28.2 (-39.3)	29.6 (-37.9)
Mistral	43.9 (-23.6)	37.1 (-30.4)	38.2 (-29.3)	53.9 (-13.6)
Qwen	43.9 (-23.6)	46.1 (-21.4)	47.1 (-20.4)	40.7 (-26.8)

Table 10. Predicted hate-speech prevalence. Values in parentheses are signed error relative to the gold rate of 43.0.

Model	Neutral	Safety	Depers.	CoT
Zephyr	43.0 $(+0.0)$	40.4 (-2.6)	36.6 (-6.4)	47.0 $(+4.0)$
Mistral	75.2 $(+32.2)$	70.0 $(+27.0)$	67.8 $(+24.8)$	65.0 $(+22.0)$
Qwen	70.6 $(+27.6)$	68.6 $(+25.6)$	67.8 $(+24.8)$	68.0 $(+25.0)$

F. Deployment Interpretation by Task

Table 11 translates the six benchmark tasks into realistic CSS deployment questions to make concrete how directional bias would distort the substantive answer.

Table 11. Deployment interpretation guide for each task.

Task	If deployed in a study, the researcher might ask…	What a directional bias would do to the answer
Hate speech	How common is hateful discourse in this corpus or subgroup?	Leniency would make harmful discourse look rarer and potentially lower urgency; overcorrection would make discourse look more extreme and inflate moderation-oriented conclusions.
Offensive language	Is community discussion becoming more abusive over time?	Undercounting would flatten temporal spikes; overcounting would create false alarms or overstate deterioration in norms.
Stance (abortion)	What share of users oppose versus support a policy issue?	Neutrality bias can collapse strong positions into a centrist-looking distribution, changing both descriptive and comparative conclusions about public opinion.
Sentiment	Is reaction to an event net positive or net negative?	Even moderate directional drift can change whether a campaign, speech, or controversy appears well received.
Emotion	Which emotions dominate a discussion after a salient event?	Miscalibrated emotion assignment can make discourse look angrier, sadder, or more optimistic than it is, changing narrative framing.
Irony	How much rhetorical indirection or sarcasm characterizes a community?	Systematic misses reduce perceived nuance and can make a discourse community look more literal and less adversarial than it is.