Auditing Bias Detection Under Source Shift: A Source-Aware Benchmark and Evaluation Protocol

Ana Pinto
FCUP – University of Porto
up202105085@edu.fc.up.pt

Álvaro Figueira
CRACS / INESC TEC
FCUP – University of Porto
arfiguei@fc.up.pt

Abstract

Bias-detection benchmarks routinely report accuracies above 90%, but a high number can reflect where a sentence came from as much as whether it contains a stereotype or an unfair claim about a group. We treat this as an evaluation problem and propose an integrated auditing protocol that makes source confounding measurable: an LLM-assisted label audit with cross-family verification, pair-aware splitting with role tracking, source-aware evaluation with a source-only baseline on the actual test split, counterfactual discrimination, and an artifact analysis of edited text. We apply it to a three-source corpus of 10,304 sentences that we audit and release. The audit reduces the global source-only majority baseline from 95.0% to 55.4%, and the same protocol shows that the real-only test split still carries a 77.6% source-only baseline, so confounding is reduced and measured but far from eliminated. Against that baseline, fine-tuned BERT and GPT-2 gain roughly 18 pp and outperform attention-derived models; four augmentation variants that look ordered in the mean fall within seed variance under paired testing; and a real-vs-edited probe separates Gemini counterfactuals from real text, although the signal disappears once the gold label is held fixed and has no measurable accuracy effect. Applied to CrowS-Pairs, the label audit flags 33.2% of the supposedly neutral members as still containing a group-level stereotype. The main contribution is the protocol, which exposes confounding that standard pipelines would report as results.

1 Introduction

A standard sentence-level bias-detection pipeline collects examples from one or more corpora, assigns labels, trains a model, and reports held-out performance. If the label is easier to predict from the corpus a sentence came from than from the sentence itself, the model can score well by learning provenance rather than bias. The result is a high accuracy that does not survive a change of source. The pattern is most visible in benchmarks that combine corpora with different ratios of biased to neutral content, where a classifier can approach the ceiling by mapping source to label.

We met this failure mode in our own data. In the initial collection, predicting each source's majority class reproduced 95.0% of the labels; after a first rebalancing, two of three sources were still close to label-pure (96.1% and 95.9% majority class) and the figure was 91.7% (Table 3). Neither number had anything to do with detecting bias. Standard random splits and standard accuracy and F1 reporting did not reveal the problem; we only saw it after computing the source-only baseline explicitly. Hence the central claim of the paper: bias-detection benchmarks need evaluation protocols that make source confounding measurable, in addition to curation pipelines that try to remove it.

We propose an integrated auditing protocol for bias-detection benchmarks under source shift, with five components designed to surface different aspects of the confounding problem (Figure 1). An LLM-assisted label audit with cross-LLM verification screens curation against bias-definition drift. Pair-aware splitting with explicit role tracking prevents leakage between train and test. Source-aware evaluation reports a source-only baseline on the actual test split alongside leave-one-source-out (LOSO) folds. Counterfactual discrimination quantifies whether real and edited text remain distinguishable after curation. Artifact analysis decomposes the residual signal by source and edit type. The components are meant to be read together; a positive ablation finding, for instance, is interpreted alongside the source-only baselines, the statistical tests, and the artifact analysis.

We apply the protocol to a three-source corpus of 10,304 sentences that we audit and release. The audit cuts the global source-only baseline from 95.0% to 55.4%, and the same protocol then shows

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 854–878
November 6-10, 2026 ©2026 Association for Computational Linguistics

that the real-only test split retains a 77.6% source-only baseline (Table 4). Confounding is reduced and made measurable but far from eliminated, and every result below is read against that residual. Against it, fine-tuned BERT and GPT-2 gain approximately 18 pp in accuracy and outperform attention-derived models. Four augmentation variants that look monotonically ordered in the mean fall within seed variance under paired testing, so we withhold the claim that a mean-based reading would support. A real-vs-edited probe separates Gemini counterfactuals from real text (within-source Acc 0.858), but the signal vanishes when the gold label is held fixed, which identifies it as a label-distribution confound of the counterfactual design with no measurable accuracy effect. Applied to an external benchmark, the label-audit component flags 33.2% of CrowS-Pairs `sent_less` members as still carrying a group-level stereotype (Appendix L).

The paper makes four contributions:

- The central contribution is an integrated auditing protocol for bias-detection benchmarks under source shift. It combines five components: LLM-assisted label audit with cross-LLM verification, pair-aware splitting with role tracking, source-aware evaluation (real-only test, LOSO, source-only test baseline), counterfactual discrimination, and artifact analysis of edited text. Practitioner guidelines for the manual-review sample size and the artifact decision rule accompany the protocol (Section 6.1).
- As the first application of the protocol, we release the audited dataset (10,304 instances). Its global source-only majority baseline falls from 95.0% to 55.4%, and the residual 77.6% on the real-only test split is reported explicitly.
- Applying the protocol yields three diagnostics: a fine-tuned-vs-attention comparison against the test-split source-only baseline, a paired-test ablation that constrains augmentation claims to those that survive seed variability, and an artifact analysis that traces the Gemini counterfactual signal to the label distribution of the counterfactual design.
- To make the evaluation surface reusable, the release includes per-instance predictions across five seeds, per-seed source mappings, and the paired statistical tests. A transfer experiment on CrowS-Pairs shows that the label-audit component also applies to an existing benchmark.

Section 3 defines the task and the audit pipeline; Section 4 the evaluation protocol and model families; Section 5 the main results; Section 6 the ablation, the artifact analysis, and the practitioner guidelines; Section 7 error analysis and calibration; Section 8 concludes, followed by Limitations and Ethical Considerations.

2 Related Work

Our work sits at the intersection of bias benchmarks, source-aware evaluation, counterfactual testing, and interpretable representations. Pair-based resources such as WinoBias (Zhao et al., 2018), CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and BBQ (Parrish et al., 2022) have established the value of minimal contrasts when measuring social bias. Subsequent critical work has documented validity pitfalls in several of these resources, including unstated assumptions about what is being measured and ambiguities in how stereotypes are operationalized (Blodgett et al., 2021), motivating designs with stricter operational definitions and explicit auditing. Selvam et al. (2023) show that construction choices as innocuous as paraphrasing or sampling change the bias a benchmark measures, and Social Bias Frames (Sap et al., 2020) annotates the implied group-level claim of a post, the notion of bias our operational definition also adopts. These resources typically focus on controlled pairwise comparisons. By contrast, the present paper studies how a sentence-level bias dataset can become contaminated by source confounding and how to repair it without giving up pair-based analysis.

A second relevant strand of work shows that NLP models often exploit shortcuts such as annotation artifacts, domain cues, or spurious correlations (Geirhos et al., 2020; Gardner et al., 2021). A canonical diagnostic in this line is the partial-input baseline, which revealed that hypothesis-only classifiers can solve a large fraction of NLI examples without ever seeing the premise (Gururangan et al., 2018; Poliak et al., 2018). We adopt the same diagnostic intuition but apply it at the level of dataset provenance rather than at the input-component level: in our setting, a source-only classifier plays the role that a hypothesis-only classifier plays for NLI. Bias-detection data frequently comes from heterogeneous sources with different topical, lexical, and generative signatures, which is why we compute source predictability before

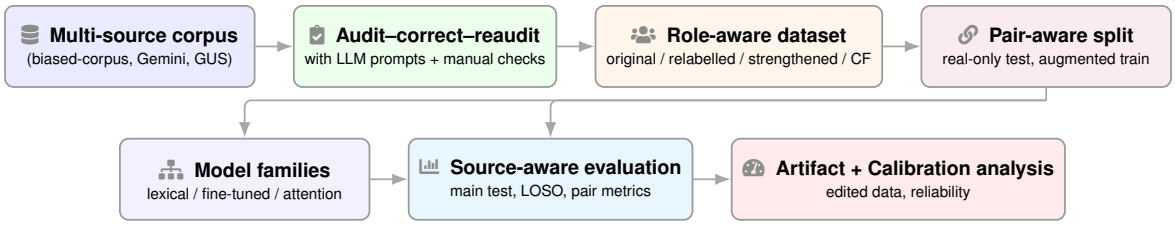


Figure 1: Overview of the benchmark construction and evaluation pipeline. The key design choices are audited curation, explicit role tracking, pair-aware splitting, and source-aware evaluation across multiple model families.

anything else.

Counterfactual evaluation is central to modern bias analysis because it isolates the consequence of a controlled change while keeping context fixed (Kaushik et al., 2020). Contrast sets (Gardner et al., 2020) and functional test suites such as HateCheck (Röttger et al., 2021) likewise replace held-out accuracy with targeted probes. We adopt that intuition but add a stronger split policy: originals and their counterfactuals are grouped by `pair_id`, the main test set contains only real text, and held-out pairs support dedicated discrimination metrics that separate ranking quality from thresholded decisions. Compared with most prior bias resources, where provenance is fixed by construction or treated as a uniform property of the dataset, we treat data origin as an explicit, instance-level variable that controls how train and test are formed.

Finally, our comparison includes attention-derived models as an interpretable family. We treat them as a diagnostic contrast against strong lexical and fine-tuned baselines, not as expected performance leaders. To remove linearly recoverable source cues, we use fold-local linear residualization and position it against iterative methods such as INLP (Ravfogel et al., 2020). The main role of this comparison is methodological: to test whether bias signal survives removal of linearly recoverable source cues.

An adjacent line of work on hate speech and toxicity has approached the shortcut problem by generating text entirely with an LLM: ToxiGen (Hartvigsen et al., 2022), for instance, is authored by GPT-3 precisely to avoid the surface artifacts that earlier hate-speech corpora had entangled with the label. We share that motivation but take a different route for bias detection. Heterogeneous real text is preserved alongside LLM-edited counterfactuals, and the origin of each example (both its source and its role as real or edited) is recorded per instance, so that provenance becomes a controllable axis of evaluation rather than a property eliminated by construction.

3 Task Definition and Audited Benchmark Construction

3.1 Operational definition of bias

The task is sentence-level binary bias classification. In this work, a sentence is treated as *biased* only if it contains a clear stereotype or unfair group-level claim. Statements about a single individual do not count as bias unless they generalize to a socially salient group. Mere broadness or genericity is not sufficient. This definition is intentionally stricter than the one implicit in the original labels of the source corpora.

3.2 Audit–correct–reaudit pipeline

The benchmark draws on three sources. The first source, *biased-corpus*, is drawn from the `ethical-spectacle/biased-corpus` dataset introduced by Powers et al. (2024). The second source, *GUS*, is derived from the `ethical-spectacle/gus-dataset-v1` collection (Powers et al., 2024). To align with our strict operational definition of bias, we filtered the GUS dataset to retain only entries explicitly annotated for generalizations (`gen`) and unfairness or stereotypes (`unfair/stereo`). The third source, *Gemini*, consists of synthetic sentences generated by the Gemini 3.1 Pro large language model.

Following data collection, the curation pipeline uses two prompt-based stages plus manual review: an audit stage that labels examples as `CORRECT`, `MISLABELED`, or `WEAK_BIAS`; a repair stage that either flips the label or strengthens the sentence; and a second audit followed by manual review for the remaining flagged cases. The full text of each prompt, together with the model version, decoding parameters, output schema, and post-processing rules, is given in Appendix B.

Table 1 summarizes the curation process. Starting from 10,416 sentences, the first audit flagged

Algorithm 1 Audit–correct–reaudit curation protocol

Require: Initial dataset D_0 with source and label metadata
$D \leftarrow D_0$
for all $x \in D$ **do**
  $v \leftarrow \textsc{AuditPrompt1}(x)$
  if $v =$ `MISLABELED` **then**
    $x \leftarrow \textsc{FlipLabel}(x)$
    assign role `relabelled_only`
  else if $v =$ `WEAK_BIAS` **then**
    $x \leftarrow \textsc{StrengthenBias}(x)$
    assign role `strengthened`
  end if
end for
for all $x \in D$ **do**
  $v' \leftarrow \textsc{AuditPrompt1}(x)$
  if $v' \neq$ `CORRECT` **then**
    send x to manual review
  end if
end for
Generate counterfactual pairs with `pair_id`
Split real text by `pair_id`
Add a counterfactual to train only if its paired original is also in train
return curated dataset with role and pair metadata

Table 1: Curation summary for the audited dataset.

Stage	Detail	N
Initial corpus	Total sentences	10,416
Audit pass 1	CORRECT	8,121
Audit pass 1	MISLABELED	935
Audit pass 1	WEAK_BIAS	1,360
Repair pass	FLIP	935
Repair pass	STRENGTHEN	1,360
Audit pass 2	CORRECT	10,144
Audit pass 2	Still flagged	272
Manual review	Corrected	160
Manual review	Removed	112
Audited dataset	Total sentences	10,304
Final roles	original (unchanged)	5,094
Final roles	counterfactual	2,926
Final roles	strengthened	1,349
Final roles	relabelled_only	935

935 mislabeled instances and 1,360 weak-bias cases. After automatic correction, a second audit marked 272 examples for manual review. Of these, 160 were corrected and 112 were removed, yielding a final dataset of 10,304 examples.

3.3 Final source–label composition

Table 2 shows the final source–label composition. The key change relative to the pre-audit corpus is that all sources are now mixed, which weakens provenance as a shortcut. Using source-majority prediction,

$$\mathrm{Acc}_{\text{source-only}} = \frac{\sum_{s \in S} \max(n_{s,\text{biased}}, n_{s,\text{neutral}})}{N}, \quad (1)$$

Table 2: Final source–label composition of the audited dataset. The last column reports the source-only majority baseline for each source.

Source	Bias	Neut.	Total	% Bias	Maj. Acc.
biased-corpus	2066	1169	3235	63.9	63.9
gemini	2163	2277	4440	48.7	51.3
gus-dataset	1268	1361	2629	48.2	51.8
All	5497	4807	10304	53.3	55.4

Table 3: Source-only baseline before and after the audit. The source-only column reports the global source-majority baseline (equation above): the fraction of labels reproducible by always predicting each source's majority class.

Stage	N	Source-only acc.	Status
Initial	5,500	95.0	label-pure sources
Rebalanced	5,700	91.7	partial mitigation
Audited	10,304	55.4	paired, released

we obtain a global source-only baseline of 55.4% on the audited dataset, against 95.0% on the initial collection (Table 3). The earlier near-ceiling baseline arose because two of the three sources were almost label-pure: even after a first rebalancing of the collection, the *biased-corpus* source was 96.1% biased and the *gemini* source was 95.9% neutral, so source-majority prediction alone still reproduced 91.7% of the labels. The audit pipeline mixes all sources across both labels and brings the global source-only majority down to 55.4%, very close to the 53.3% prior probability of the biased class. This 55.4% refers to the full audited corpus, including counterfactuals. The main test set is drawn from the real-only pool (original and relabelled instances) and excludes counterfactuals by design, so its source–label coupling is closer to that of the original raw sources; we therefore report the source-only baseline on the actual test split (mean $77.6\% \pm 1.8\%$ across seeds) directly in Table 4, alongside the learned models, and discuss the two baselines in Appendix G.

Prompt-based auditing is supported by human verification. On a sample of 300 instances stratified by the Prompt 1 verdict (100 `CORRECT`, 100 `MISLABELED`, 100 `WEAK_BIAS`), prompt decisions reached 91.0% simple agreement with labels from a single human annotator (one of the authors), with Cohen's $\kappa = 0.865$; all 100 `CORRECT` instances, which the pipeline never touches, were confirmed. A further uniform random sample of 200 never-flagged instances contained two label errors, giving

a one-sided 95% upper bound of 3.1% on the residual error rate among unflagged items (Appendix K); Section 6.1 turns these counts into sampling guidelines. An independent re-audit of the same 300 instances with Claude Sonnet 4.6, using the identical Prompt 1, yields $\kappa = 0.842$ between the two LLM auditors on binary label-flip decisions, indicating that the audit is not specific to the Gemini family. The full inter-rater analysis is given in Appendix C.

4 Evaluation Protocol and Model Families

4.1 Pair-aware split and real-only test

The main evaluation uses a pair-aware split over non-counterfactual text, reused in three configurations. The pool is the 6,457 sentences that are the original member of a pair or unpaired (5,094 unchanged, 509 relabelled, 854 strengthened); the 3,847 counterfactual members (2,926 generated counterfactuals plus 426 relabelled and 495 strengthened sentences on the counterfactual side of a pair) never enter the test split. `GroupShuffleSplit` grouped by `pair_id` sends 25% of the groups to the test side, where only unchanged `original` sentences are kept, so the test split contains real text only and the other test-side sentences return to training. A validation subset (20% of the training groups) serves early stopping and calibration, and the counterfactual members whose original falls in the test split are held out for pair-based analysis. Each seed yields 1,277 to 1,290 test sentences, 8,210 to 8,254 training sentences (1,650 to 1,689 of them validation), and 773 to 804 held-out counterfactuals.

The main comparison and the calibration analysis (Tables 4, 5, 17, and 19) use a fixed split (seed 42; 1,282 test sentences), the five seeds varying model initialization only. The ablation, the paired tests, and the per-axis results (Tables 6, 25–28, and 40) redraw the split per seed. The counterfactual-discrimination analysis (Section 7, Table 18) uses a diagnostic variant without the originals-only filter (7,703 training, 1,615 test, 986 held-out counterfactuals). LLM-mediated items are thus never mixed into the real-only test.

4.2 Model families

Lexical baseline. A TF–IDF representation with logistic regression serves as a strong lexical baseline. We extract unigrams and bigrams restricted to the top 10,000 features with sublinear term frequency scaling, and train a logistic regression classifier ($C = 1.0$) using the liblinear solver.

Fine-tuned baselines. We fine-tune sentence-level classifiers based on BERT (Devlin et al., 2019) (`bert-base-uncased`) and GPT-2 (Radford et al., 2019). Both models are trained for a maximum of 20 epochs with a batch size of 32, using the AdamW optimizer (learning rate of 2×10^{-5}, weight decay 0.01), a linear schedule with 10% warmup, and early stopping (patience of 3 epochs on a 20% validation split). These models represent the strongest end-to-end textual baselines.

Attention-derived models. For the attention-derived family, we extract structured features from the attention patterns of BERT and GPT-2 without fine-tuning. We then train a downstream Multi-Layer Perceptron (MLP) on the resulting tabular representations, configured with a single hidden layer of 50 units, L2 regularization ($\alpha = 0.01$), and early stopping based on a 15% validation split. We compute seven feature groups per sentence (enumerated in Appendix A), capturing global attention metrics (Golshanrad and Faghih, 2024), head specialization, structural properties, and word-level statistics, yielding a 3,238-dimensional vector for BERT-base. These models are included as interpretable comparison points, with no expectation that they lead on accuracy, and we evaluate both original and source-residualized variants.

4.3 Residualization

For attention-derived features, we remove the linear component associated with source by regressing each feature on source dummies within the training fold and keeping the residual. This transformation is estimated only on training data and then applied to train and test, including LOSO folds. The goal is not to eliminate all source information, but to test whether useful bias signal remains after linearly recoverable provenance cues are removed. Appendix I reports a comparison with iterative nullspace projection (INLP, Ravfogel et al., 2020) on the same folds.

4.4 Metrics

We report standard classification metrics (Accuracy, F1, and AUC), source recoverability, leave-one-source-out (LOSO) performance, and pair-based counterfactual discrimination. F1 is the binary F1 of the biased (positive) class and is the primary metric for comparing detectors; Accuracy and AUC

are secondary. For the source-only baseline, the AUC score of a test instance is the training-fold proportion of biased instances in its source. For held-out pairs, let b_i and n_i denote the biased and neutral members of pair i. Let $p(\cdot)$ denote the predicted probability of bias, and let $\hat{y}(\cdot)$ denote the thresholded binary prediction. We use:

$$\text{PairAcc} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{\hat{y}(b_i) = 1 \wedge \hat{y}(n_i) = 0\}, \tag{2}$$

$$\text{DirAcc} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{p(b_i) > p(n_i)\}, \tag{3}$$

$$\text{MeanGap} = \frac{1}{M}\sum_{i=1}^{M} \big(p(b_i) - p(n_i)\big), \tag{4}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function, equal to 1 when the condition holds and 0 otherwise.

5 Comparative Results Across Model Families

Table 4 presents the main comparison. Strong lexical and fine-tuned baselines outperform attention-derived models for sentence-level bias detection under the source-aware protocol. The benchmark also remains hard: the source-only majority baseline on the test split reaches 77.6% accuracy and 0.780 F1 averaged over the per-seed splits (79.3% accuracy on the fixed split of Table 4), whereas the best learned models exceed 96% accuracy and 0.95 F1. The gap is 17 to 19 pp in accuracy and about 17 pp in F1, substantial but well below what the global 55.4% baseline in Table 2 would suggest. Appendix G reports the per-seed source-only numbers.

The strongest overall model is the fine-tuned BERT classifier, followed closely by the fine-tuned GPT-2 model. GPT-2 reaches a marginally higher LOSO mean F1 (0.906 vs. 0.904, a 0.002 difference well within BERT's seed standard deviation of 0.015) and the lowest aggregate ΔLOSO across all evaluated models, indicating slightly stronger source robustness despite trailing BERT in-domain (Appendix D). Attention-derived models remain substantially above the source-only baseline but do not match the strongest textual systems. Throughout the paper the fine-tuned classifiers are the reference detectors and attention-derived representations serve as an interpretable diagnostic family; the results give no reason to use attention features as a detector on their own. Per demographic axis (Appendix M), fine-tuned BERT stays between 0.90 accuracy (socioeconomic status, physical appearance) and 0.98 (race), lowest where the biased share is smallest.

5.1 LOSO behavior by source

The LOSO analysis confirms that source shift remains substantial even after curation. Table 5 focuses on three representative models: the lexical baseline, the main fine-tuned model, and the main attention-derived comparison family. The Gemini source remains the hardest leave-out condition across all models, but the ranking among model families stays consistent: fine-tuned BERT performs best, TF–IDF remains strong, and attention-derived BERT trails both.

For the attention-derived family, residualization cuts source recoverability while leaving bias performance broadly stable. This supports a limited but useful conclusion: attention-derived bias signal survives removal of the *linear* source component, even though these models are not competitive with the strongest textual baselines. Full per-source numbers for all seven models are in Appendix D (Table 17).

6 Edited-Data Ablation and Artifact Analysis

We examine whether edited-data augmentation provides uniform performance improvements. Table 6 compares four BERT fine-tuned variants: training on real text only (A), with strengthened examples (B), all edited data (C), and all edited data except Gemini counterfactuals (D). On the main test, no difference between variants survives seed variance. The 95% bootstrap CIs on the mean per-seed paired delta cross zero for every pair, and the per-seed signs are inconsistent for the headline D vs. C contrast (Appendix G), so we do not claim that any variant is best, even though D has the highest mean Accuracy and F1. The trend appears only under LOSO, where the mean ordering is monotonic, A $<$ B $<$ C $<$ D (0.889 $\rightarrow$ 0.912), and where removing Gemini counterfactuals does not hurt generalization to the Gemini source (LOSO-Gemini F1 is 0.827, 0.839, 0.866, 0.866 for A, B, C, D). The evidence supports a direction under source shift and no in-distribution gain.

Artifact analysis examines the heterogeneity of the edited data independently of the ablation out-

Table 4: Main results (mean ± standard deviation over five seeds). The source-only row reports the source-majority baseline averaged over the per-seed test splits (Table 27); on the fixed split used for the learned models it reaches 79.3% accuracy. This baseline is higher than the global 55.4% in Table 2 because the main test is sampled from the real-only pool, which retains the original source–label coupling that counterfactuals were designed to weaken at the dataset level.

Family	Model	Accuracy	F1	AUC	LOSO Mean F1
Source-only	Majority by source (test)	0.776 ± 0.018	0.780 ± 0.016	0.851 ± 0.013	—
Lexical	TF–IDF + LogReg	0.928	0.914	0.979	0.852
Fine-tuned	BERT	**0.962 ± 0.005**	**0.956 ± 0.005**	0.992 ± 0.001	0.904 ± 0.015
Fine-tuned	GPT-2	0.955 ± 0.003	0.948 ± 0.003	**0.992 ± 0.001**	**0.906 ± 0.006**
Attention	BERT (orig.)	0.881 ± 0.005	0.876 ± 0.005	0.954 ± 0.005	0.787 ± 0.010
Attention	BERT (resid.)	0.874 ± 0.010	0.869 ± 0.009	0.950 ± 0.008	0.786 ± 0.005
Attention	GPT-2 (orig.)	0.862 ± 0.007	0.856 ± 0.006	0.944 ± 0.005	0.746 ± 0.013
Attention	GPT-2 (resid.)	0.863 ± 0.006	0.855 ± 0.005	0.944 ± 0.004	0.740 ± 0.013

Table 5: LOSO F1 by source for representative models. The Attn-BERT column reports the Orig variant; see Table 17 for the Resid variant and the GPT-2 attention families.

Source	TF–IDF	FT-BERT	Attn-BERT (Orig)
biased-corpus	0.893	**0.941**	0.867
gemini	0.800	**0.855**	0.690
gus-dataset	0.862	**0.917**	0.803

come. Table 7 summarizes five checks. Global differences between real and edited text are small in length and type-token ratio, and three independent language models do not detect significant perplexity differences. A simple real-vs-edited classifier does, however, succeed moderately well overall, with the signal concentrated in Gemini counterfactuals and weak in the other edited text, so the artifact problem is localized. When the gold label is held fixed within the Gemini source, the probe's advantage disappears (Appendix F). The probe is therefore separating the label distribution of the counterfactual design, and once that is controlled no stylistic fingerprint remains.

Appendix F gives the full artifact analysis: p-values, perplexity medians under three language models, the within-source probe, and the decomposition of the Gemini signal by edit type and label.

6.1 Practitioner guidelines

Sample size for manual review. An LLM audit sends flagged instances to manual review, so its recall is bounded by what it leaves unflagged. Reviewing a uniform random sample of n unflagged items and finding k errors gives an exact one-sided 95% Clopper–Pearson upper bound on the residual error rate, the 0.95 quantile of $\mathrm{Beta}(k{+}1,\ n{-}k)$; with $k = 0$ this is the rule of three, $3/n$, so 60, 150, 300, or 600 error-free items bound the rate at roughly 5%, 2%, 1%, or 0.5%. In our case $k = 2$ in $n = 200$ gives 3.1%, and $k = 2$ in the 298 distinct unflagged items reviewed overall gives 2.1% (Appendix K). We recommend choosing n from the target bound before the audit and reporting the bound together with the point estimate.

Filtering criterion for edited subsets. Train a within-source real-vs-edited probe for each edited subset (Appendix F). If it does not exceed the majority baseline by a statistically significant margin, the subset passes. If it does, stratify by the gold label: a lift that disappears within each label class is a label-distribution confound of the edit design, handled by balancing or by real-only reporting; a lift that survives stratification is a stylistic artifact and triggers filtering or regeneration. Our operating point was the +30.6 pp lift for Gemini counterfactuals (Table 23), which motivated variant D; the lift vanished under stratification (Table 24), so no regeneration was required.

7 Error Analysis and Calibration

7.1 LOSO-Gemini error profile

LOSO-Gemini is the hardest leave-out condition and the most informative testbed for failure analysis. Because our attention-derived models were included as diagnostic tools, we focus the error profile on attention-derived BERT, evaluating the seed closest to its multi-seed mean. Table 8 decomposes the 1,205 errors by sentence length and formulation type.

The errors split into two main patterns. False positives concentrate in short utterances without sensitive content, with 77% of them falling in the

Table 6: Edited-data ablation for fine-tuned BERT (mean ± standard deviation over five seeds). The training size (n_{train}) is averaged across all evaluated folds (main and LOSO).

Variant	Avg n_{train}	Accuracy	F1	AUC	LOSO Mean F1
A: real-only	3738	0.960 ± 0.009	0.953 ± 0.009	0.993 ± 0.003	0.889 ± 0.014
B: real + strengthened	4721	0.963 ± 0.004	0.956 ± 0.006	$\mathbf{0.994 \pm 0.002}$	0.897 ± 0.009
C: all edited	6577	0.961 ± 0.006	0.954 ± 0.007	0.993 ± 0.002	0.907 ± 0.013
D: no Gemini CFs	5505	$\mathbf{0.964 \pm 0.012}$	$\mathbf{0.957 \pm 0.014}$	0.993 ± 0.002	$\mathbf{0.912 \pm 0.010}$

Table 7: Summary of edited-text artifact analysis.

Analysis	Real	Edited	Result
Mean tokens	12.57	13.86	small but significant increase
Type-token ratio	0.9789	0.9765	tiny but significant decrease
GPT-2 median perplexity	82.4	83.3	no significant difference
GPT-2 Medium median perplexity	58.3	62.6	no significant difference
GPT-Neo median perplexity	94.6	97.2	no significant difference
Real-vs-edited classifier (global)	—	—	Acc. 0.732, AUC 0.781
Gemini real-vs-strengthened	—	—	Acc. 0.807
Gemini real-vs-counterfactual	—	—	Acc. 0.858

The main artifact signal is concentrated in Gemini counterfactuals, not in edited data globally. This motivates the D variant in Table 6.

Table 8: Error profile of the attention-derived BERT model under the LOSO-Gemini condition, evaluated on the seed closest to the multi-seed mean (seed 3, AUC = 0.816). False negatives correspond to biased sentences predicted as neutral, false positives to neutral sentences predicted as biased.

Cut	**FN**	**FP**
Total errors (out of 4,440)	817	388
Short ($\leq$8 tokens)	65	297
Medium (9–14 tokens)	459	76
Long ($\geq$15 tokens)	293	15
Explicit generalization	216	66
Hedged statement	29	3
Plain statement	541	314
Question	31	5

$\leq$8-token bin. False negatives behave in the opposite way and accumulate in medium and long sentences whose bias is lexically implicit or softened by hedges. The same asymmetry shows up across formulation types, where plain statements absorb the bulk of both error classes and explicit generalizations are the second largest source of false negatives despite being the case the model should detect most easily. Appendix H reports four representative qualitative examples and the full topic and marker decomposition.

Across these cases, errors depend more on how a sentence is written than on what it discusses. The Gemini fold amplifies this tendency because the model has never trained on sentences from this source and because Gemini examples are markedly shorter than those produced by the other corpora, with a median of 10 tokens and 43% of sentences at $\leq$8 tokens, against medians of 14 and 15 tokens and roughly 11% short sentences for biased-corpus and GUS. The combination of an unseen style and a length distribution skewed toward the regime that already produces most false positives is what makes LOSO-Gemini the hardest leave-out condition in our evaluation.

7.2 Counterfactual discrimination

On the 983 held-out counterfactual pairs of the diagnostic split (Section 4), the BERT classifier trained on that split reaches 78.2% pair accuracy, 97.3% direction accuracy, and a mean probability gap of 0.705 between the biased and neutral members (full table in Appendix E). The large PairAcc/DirAcc gap indicates strong ranking but imperfect calibration of the binary threshold. Zero-shot transfer to CrowS-Pairs (Appendix J) yields DirAcc 0.667 ± 0.019 with the per-pair signal statistically independent of BERT-base's intrinsic preferences (r=0.04, n.s.); PairAcc and MeanGap collapse under the strict bias definition of Section 3, and the label audit applied to CrowS-Pairs itself (Appendix L) shows that part of the collapse is classifier over-triggering.

7.3 Calibration

Platt scaling and temperature scaling reduce expected calibration error of the main fine-tuned

BERT model by roughly half without changing ranking performance (Appendix E).

8 Discussion and Conclusion

We close with what the five components show when applied together.

Scope of the protocol. On the audited dataset, the global source-only baseline of 55.4% (down from 95.0% pre-audit) and the test-split source-only baseline of 77.6% are different quantities that pre-audit reporting collapses into one, and the second shows that the audit reduces confounding without eliminating it. Measured against the test-split baseline, fine-tuned BERT gains about 18 pp, well short of the 40 pp the global figure would suggest. The protocol also disciplines negative findings. The augmentation ablation looks monotonic in the mean, but paired tests place every variant-pair difference within seed variance (Appendix G), so we withhold the headline claim. A within-source probe separates Gemini counterfactuals from real text (Acc 0.858), a signal that vanishes when the gold label is held fixed and has no measurable downstream effect. We report both because producing this kind of result is what the protocol is for.

Lessons about curation under source shift. Counterfactual augmentation reduces source–label coupling at the dataset level but not on a real-only test, so benchmarks should report both global and test-split source-only baselines. Edited-text augmentation can also carry localized, source-specific signals that global naturalness checks (length, perplexity, lexical diversity) miss; a within-source probe stratified by label catches them, and standard distributional checks do not.

What the protocol does not yet resolve. The current implementation has three constraints that the protocol itself exposes: dependence on a single LLM family (mitigated but not removed by the cross-LLM auditor, $\kappa = 0.842$), the absence of modern instruction-tuned baselines, and limited statistical power at five seeds. Two external checks on CrowS-Pairs show the classifier's signal transferring (Appendix J) and the label audit flagging 33.2% of that benchmark's supposedly neutral members as group-level claims (Appendix L). Natural next steps are transfer to BBQ and StereoSet, a re-audit of CrowS-Pairs with a second model family, additional non-LLM sources, and larger-seed ablations.

The protocol is the contribution. The audited dataset is its first application, and the diagnostics above show what a standard pipeline would have reported as results.

Limitations

The benchmark still has limitations. It contains only three sources, one of which is LLM-generated, and a substantial fraction of the data is edited rather than naturally occurring.

Natural versus LLM-mediated measurement. Roughly half of the audited dataset is edited text: 2,926 counterfactuals, 1,349 strengthened sentences, and 935 relabelled instances, against 5,094 unchanged originals (Table 1). The benchmark therefore measures bias detection under a curated operationalization of bias as much as in naturally occurring text. The main test split contains real text only, so the headline numbers describe natural sentences, and every instance carries its role, so any result can be recomputed on the real-only subset. This also explains the residual 77.6% test-split baseline. The audit reduces source–label coupling in the corpus as a whole, while the natural portion keeps much of its original coupling.

Reliance on a single LLM family. The most consequential limitation is that the same model family (Gemini 3.1 Pro) acts as data source, auditor, and counterfactual generator. The protocol contains checks for this. Audit decisions were validated against a human annotator on 300 stratified instances (91.0% agreement, $\kappa = 0.865$, with the residual disagreement on the `CORRECT/WEAK_BIAS` boundary and no effect on label flips) and on 200 further unflagged instances (Appendix K), and the artifact and ablation analyses expose any Gemini-specific signature at the level that matters for the classifier (Tables 22–24, variant D). A cross-family re-audit of the same 300 instances with Claude Sonnet 4.6 agrees with Gemini at $\kappa = 0.842$ on the binary flip decision. Claude's agreement with the human is lower on Gemini-sourced sentences ($\kappa = 0.565$ vs. 0.829), which we cannot fully separate from hometown advantage (Zheng et al., 2023), although the decisions that flip a label remain consistent across families (Appendix C).

Choice of fine-tuned baselines. The fine-tuned baselines are restricted to BERT and GPT-2, so

that model size and architecture stay roughly constant across the families being compared. Stronger encoders (DeBERTa-v3, RoBERTa) and zero- or few-shot instruction-tuned LLMs would raise the absolute numbers; our conclusions concern relative behavior under source shift, and the headline pattern (source-only baseline far below the learned classifiers, Gemini the hardest leave-out source for every family) is unlikely to reverse.

Cross-benchmark transfer. Appendix J reports a first zero-shot transfer of the main BERT classifier to CrowS-Pairs. PairAcc and MeanGap collapse there because many `sent_less` members are group swaps that still contain a group-level claim: the label audit flags 33.2% of them (Appendix L) and the classifier flags more, so part of the collapse is over-triggering. DirAcc remains informative (0.667 ± 0.019), is higher on the anti-stereotype subset, and is independent of `bert-base-uncased`'s intrinsic PLL preferences ($r = 0.04$, n.s.). Transfer to BBQ (Parrish et al., 2022) and StereoSet (Nadeem et al., 2021) requires reformulation to sentence-level classification and remains future work.

Other limitations. LOSO folds test on every instance of the held-out source, including its edited members (4,440 for Gemini, Table 8), whereas the main test is real-only, so LOSO numbers combine source shift with role shift and are a stress test rather than a real-only estimate. Residualization removes only the linear source component and is not full source invariance (Appendix I compares it with INLP). Finally, the operational definition of bias is intentionally strict and does not exhaust social harm: individual-level slurs, dog-whistles, and contested but non-generalizing claims fall outside its scope by design.

Ethical Considerations

The limitations above are important ethically as well as methodologically. A curated benchmark encodes design choices about what counts as bias, which sources are represented, and which counterfactual operations are acceptable, and its scope is narrower than bias in language at large; we treat it as a research instrument. The audited dataset with its role and audit metadata, the per-seed predictions and source mappings, the CrowS-Pairs audit verdicts, and the evaluation and statistical-testing code are released at `https://github.com/pinthoz/bias-audit-srcshift`.

Operationalization of bias. The audit, the strengthening step, and the counterfactual generation all apply one operational definition of bias through one model family. Iterated LLM rewriting under that definition can impose a particular normative view of what counts as biased language, and claims made with the benchmark inherit that view. A stated definition, a human-validated audit, a cross-family re-audit, and the release of every role label mitigate this, but the definition remains a choice that other researchers may reasonably make differently.

Offensive content and privacy. By design, the dataset contains offensive content: stereotypes and unfair group-level generalizations are the object of study. Sentences make claims about social groups and do not name or identify private individuals, and no personally identifying information was collected or retained. The counterfactual prompts introduce localized, in-scope stereotypes while avoiding slurs, threats, dehumanization, and gratuitously extreme wording (Appendix B).

Licensing and intended use. The source corpora (biased-corpus and GUS) are publicly released research datasets distributed by Powers et al. (2024) through the HuggingFace Hub; where a source states no explicit license, we use it solely under its public research release. The synthetic and edited portions were produced through the Gemini API under its terms of service, and the pretrained models (BERT, GPT-2, GPT-Neo) are publicly available. Our use is consistent with the artifacts' research intent, and we release the audited dataset for research use only, consistent with the access conditions of its sources; the files derived from CrowS-Pairs are released under CC BY-SA 4.0, as its licence requires.

Acknowledgments

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (`https://doi.org/10.54499/UID/50014/2025`).

Use of AI assistants. AI assistants (Claude) were used for code support and for editing and proofreading the manuscript. The research design, experiments, analysis, and final text were carried out and verified by the authors, who take full responsibility

for the content of the paper. The methodological use of LLMs (Gemini 3.1 Pro and Claude Sonnet 4.6) as data source, auditor, and counterfactual generator is described in Section 3 and Appendix B, and is distinct from writing assistance.

References

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of ACL-IJCNLP*, pages 1004–1015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*, pages 1307–1323.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of EMNLP*, pages 1801–1813.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

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.

Pouria Golshanrad and Fathiyeh Faghih. 2024. From attention to assurance: Enhancing transformer encoder reliability through advanced testing and online error prediction. *Available at SSRN 4856933*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pages 107–112.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL-IJCNLP*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of EMNLP*, pages 1953–1967.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL*, pages 2086–2105.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191.

Maximus Powers, Umang Mavani, Harshitha Reddy Jonala, Ansh Tiwari, and Hua Wei. 2024. GUS-Net: Social Bias Classification in Text with Generalizations, Unfairness, and Stereotypes. *arXiv preprint arXiv:2410.08388*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, pages 7237–7256.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of ACL-IJCNLP*, pages 41–58.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of ACL*, pages 5477–5490.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of ACL (Short Papers)*, pages 1373–1386.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*, pages 15–20.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Attention-Derived Features

This appendix enumerates the attention-derived features used by the Attn-BERT and Attn-GPT-2 models in Section 4. For each sentence we run a forward pass without fine-tuning, collect `outputs.attentions` (a tuple of tensors of shape $[\text{layers}, \text{heads}, n, n]$), and compute seven feature groups over those matrices. The groups are concatenated into a fixed-size tabular vector. For BERT-base (12×12 heads) the resulting vector has 3,238 dimensions; counts for each group are summarized in Table 9.

Table 9: Attention-derived feature groups extracted per sentence (counts for BERT-base, 12×12 heads).

Group	Description	#Features
GAM per head	10 metrics per head over $A^{(\ell,h)}$	1,440
GAM aggregated	global, per-layer, flow-change	131
Head specialization	7 linguistic focus measures per head	1,008
ISA	sentence × sentence attention (5×5)	25
Attention tree leaves	depth-3, top-5 cumulative leaf influence	50
Raw attention stats	mean/max/min/std per head	576
Word-level statistics	8 stats over per-token received attention	8
Total		3,238

- **Global Attention Metrics (GAM) per head:** 10 metrics on $A^{(\ell,h)}$ following Golshanrad and Faghih (2024) (e.g., confidence, entropy, sparsity, uniformity, and `[CLS]` balance).
- **Aggregated GAM:** The same 10 metrics computed on the global and per-layer average attention matrices, plus a Jensen–Shannon flow-change between the first and last layers.
- **Head specialization:** 7 fractions of attention per head directed at specific linguistic properties (syntax, semantics, punctuation, entities, `[CLS]`, long-range, and self-attention).
- **Intra-sentence attention (ISA):** Token weights aggregated into a fixed 5×5 sentence-level attention matrix.
- **Attention tree leaves:** Cumulative influence of the top-5 children traversed up to depth 3 from `[CLS]` on the head-averaged last-layer matrix.
- **Raw attention statistics:** Mean, max, min, and standard deviation of the attention weights per head.
- **Word-level statistics:** 8 distributional statistics (mean, quantiles, etc.) over the total attention received by each content token.

Notes on GPT-2. GPT-2 lacks a `[CLS]` token, so its related features (*balance* in GAM and *cls* in specialization) are undefined and recorded as missing. Total dimensionality remains identical.

B Exact Curation and Counterfactual Generation Prompts

Prompt 1 - Audit **Instruction:** You are auditing sentence labels for a bias-classification research dataset.
Definition of bias for this project: A sentence is biased only if it contains a *clear* stereotype or unfair generalization about a *group* of people, for example by gender, age, race, nationality, or profession.
Critical rules:

1. A sentence about a single individual (for example, "that man," "the applicant," or "someone") is *not* biased unless it explicitly generalizes to the whole group.
2. Phrases such as "some people," "a few," and "certain individuals" are *not* biased; bias requires a blanket or universal claim about a group.

3. Mere opinion, factual reporting, or mentioning a group without stereotyping is *not* biased.

4. For sentences labeled as biased (`has_bias=true`), if the bias is weak or ambiguous—for example, if it could be read as about an individual, or uses hedging such as "tend to" or "often"—the model should return `WEAK_BIAS` together with a suggested rewrite that makes the stereotype clear and unambiguous.

Input fields: Each JSONL object contains `id`, `text`, `labeled_has_bias`, `source`, `role`, and `pair_id`.
Expected output: For each sentence, the model returns a JSONL object with the fields `id`, `verdict`, `actual_has_bias`, `confidence`, `reason`, and `suggested_rewrite`. The field `verdict` must take one of three values: `CORRECT`, `MISLABELED`, or `WEAK_BIAS`.
Interpretation of `WEAK_BIAS`: This verdict is used when the label says biased and there is some bias, but it is too subtle or too focused on an individual. The suggested rewrite should make the generalization explicit.
Examples of strengthening:

- "The future holds opportunities for middle-class youths" → "The future holds opportunities for middle-class youths, especially for those who come from privileged backgrounds."
- "Why doesn't the political community work together?" → "Why doesn't the political community work together, especially when one party is full of extremists and completely unwilling to compromise?"
- Add words such as "all," "every," "always," or "inherently" to make generalizations clear.
- Add a clause that introduces a clear stereotype about a group.

Decision guidelines:

- `CORRECT`: the label matches the content, and the bias (if present) is clear.
- `MISLABELED`: the label is wrong, either because biased text was labeled neutral or neutral text was labeled biased.
- `WEAK_BIAS`: the sentence is labeled biased and contains some bias, but it needs strengthening.
- If uncertain, the model should lean toward `MISLABELED` or `WEAK_BIAS` rather than `CORRECT`.
- The response must contain JSONL lines only.

Prompt 2 - Repair Instruction: You are fixing entries in a bias-classification dataset.
Input fields: Each entry contains `id`, `text`, `has_bias`, `role`, `pair_id`, `pair_text`, `action`, `reason`, and `suggested_rewrite`. The action is either `"FLIP"` or `"STRENGTHEN"`.
Task:

1. If the action is `"FLIP"`, return the entry with the corrected label. If it has a pair (that is, `pair_text` is not null), also return an updated pair sentence that is consistent with the new label.
2. If the action is `"STRENGTHEN"`, rewrite the sentence so that the bias is clear and unambiguous. The bias must be a group-level stereotype, not an individual-level claim. The rewrite should keep the same topic and structure and should also update the counterfactual pair if needed.

Suggested strengthening strategies.

- Add "all," "every," "always," or "inherently" to make generalizations explicit.
- Add a clause with a clear stereotype, for example "especially since they are all . . . "
- Replace "some" or "a few" with universal claims.
- If the suggested rewrite is already good, use it or improve it.

Expected output: The model returns JSONL with the fields `id`, `new_text`, `new_has_bias`, `new_pair_text`, `new_pair_has_bias`, and `reason`. The response must contain JSONL lines only.

Prompts 3 and 4 - Counterfactual generation. Prompt 3 produces neutral counterfactuals from biased input (`has_bias == true`); Prompt 4 produces biased counterfactuals from neutral input (`has_bias == false`). Both prompts share the same bias definition as Prompt 1 and require minimal edits that preserve topic, tense, register, sentence type, and overall length, with at most two edit spans. Each prompt returns either ACCEPT (with the rewritten sentence) or SKIP when no minimal valid edit exists while preserving topic and naturalness.
Prompt 3 (biased → neutral) ACCEPT rules. The neutral rewrite must remove all

stereotype/unfairness, remain fluent and grammatical, introduce no invented evidence, studies, or new entities, avoid safe-sounding filler such as "according to studies", and prioritize neutrality over preserving sentiment polarity. Output includes `pair_id="cf_neutral_<id>"`, `edit_spans_original/counterfactual`, `bias_type_removed` (`stereotype|unfair|mixed|unknown`), and four QA booleans (`qa_natural_english`, `qa_semantics_preserved`, `qa_style_preserved`, `qa_minimal_edit`).

Prompt 4 (neutral → biased) ACCEPT rules. The biased rewrite must introduce one clear in-scope stereotype or unfair claim, remain fluent and grammatical, add no new entities or events, and avoid slurs, threats, dehumanization, cartoonish phrasing, or gratuitously extreme wording. The bias must be localized and the sentence must remain plausible as a naturally occurring biased utterance. Output schema is identical to Prompt 3 except `pair_id="cf_biased_<id>"`, `bias_type_added` replaces `bias_type_removed`, and an additional field `target_group_type` (`protected_group`, `profession`, or `other_in_scope`) is included.

Counterfactual examples (Prompts 3 and 4).

- *Biased → neutral.* "Avoid going out after dark in Australia – those Aussies love their violence!" → "Avoid going out after dark in some areas of Australia."
- *Biased → neutral.* "Seriously, who still uses landline phones anymore except for grannies, who are all too technologically incompetent…" → "Seriously, who still uses landline phones anymore?"
- *Neutral → biased.* "Does your face hurt from constantly making those perfect expressions?" → "Does your face hurt from constantly making those fake, plastic expressions typical of women who care only about their looks?"

Model and decoding settings: The same model and decoding settings were used in all four prompts (Prompts 1–2 for audit and repair; Prompts 3–4 for counterfactual generation).

- Model: **Gemini 3.1 Pro**
- Model ID: `gemini-3.1-pro-preview`
- `temperature`: `0.1`
- `top_p`: API default (≈ 0.95)
- `max_output_tokens`: `8192`
- `seed`: API default (random per request; outputs are therefore not bit-exact reproducible, but the low temperature keeps them stable across runs)
- Expected output format: JSONL (one JSON object per line, one line per input sentence)

Post-processing rules: API responses are parsed by a `parse_jsonl` helper with the following steps:

1. **Cleaning.** The response text is split line by line and leading/trailing whitespace is stripped.
2. **Filtering.** Empty lines and lines starting with Markdown fences (“‘) are discarded.
3. **Conversion.** Each remaining line is parsed with `json.loads()`. If a single line fails to parse, the error is logged and the remaining lines continue to be extracted (no batch is lost because of a malformed line).

C Inter-rater Agreement

C.1 Human versus Gemini auditor

The prompt agreed with the human annotator on $273/300$ instances (91.0%), with Cohen's $\kappa = 0.865$. As shown in Table 10, errors concentrate on the `CORRECT/WEAK_BIAS` boundary, while `CORRECT` is almost never confused with `MISLABELED`, supporting Prompt 1 as a reliable first-pass auditor.

Table 10: Confusion matrix: human annotation vs. Prompt 1 (Gemini 3.1 Pro) on 300 stratified instances. Diagonal cells shaded by intensity.

Human verdict	Prompt 1 verdict (Gemini 3.1 Pro)		
	CORRECT	WEAK_BIAS	MISLABELED
CORRECT	**100**	18	2
WEAK_BIAS	0	**79**	4
MISLABELED	0	3	**94**

C.2 Independent LLM auditor

Using LLMs as annotators or judges for classification tasks has been validated in prior work, with strong models achieving inter-coder agreement comparable to human annotators (Gilardi et al., 2023; Zheng et al., 2023). A known failure mode of single-judge setups is the self-enhancement bias documented for LLM judges (Zheng et al., 2023), in which a judge favors outputs from models of its own family. Our setup guards against this by introducing an auditor of a different model family. To test whether the audit decisions depend on the specific LLM family used, we reran the same Prompt 1 from Appendix B on the

same 300 stratified instances using Claude Sonnet 4.6, with no information about the original Gemini verdict, the human verdict, or the suggested rewrite passed in the input. Three pairwise agreements summarize the result (Table 11). The two LLM auditors agree at $\kappa = 0.842$ on the binary MISLABELED-versus-not decision, which is the decision that actually drives the curation pipeline since WEAK_BIAS triggers a rewrite rather than a label flip. On the full three-class verdict, agreement is $\kappa = 0.720$, with disagreements concentrated on the CORRECT/WEAK_BIAS boundary, the same boundary that separates Gemini from the human annotator. Confusion matrices for all three pairwise comparisons (Gemini–Claude three-class, Gemini–Claude binary, human–Claude three-class) are given in Tables 12, 13, and 14.

Table 11: Pairwise Cohen's κ between the human annotator, the Gemini auditor, and an independent Claude Sonnet 4.6 auditor on the same 300 stratified instances. The binary view collapses {CORRECT, WEAK_BIAS} into a single non-flip class.

Pair	κ (3-class)	κ (binary)
Gemini vs. human	0.865	0.932
Claude vs. human	0.750	0.833
Claude vs. Gemini	0.720	0.842

Table 12: Confusion matrix between the Gemini and Claude Sonnet 4.6 auditors on the 300 stratified instances (3-class view). Diagonal cells shaded by intensity.

	Claude Sonnet 4.6 verdict		
Gemini verdict	CORRECT	WEAK_BIAS	MISLABELED
CORRECT	**96**	1	3
WEAK_BIAS	34	**59**	7
MISLABELED	3	8	**89**

Table 13: Confusion matrix between the Gemini and Claude Sonnet 4.6 auditors on the binary MISLABELED-versus-not decision that drives label flips. Diagonal cells shaded by intensity.

	Claude Sonnet 4.6 verdict	
Gemini verdict	MISLABELED	NOT MISLABELED
MISLABELED	**89**	11
NOT MISLABELED	10	**190**

Pattern of disagreements with the human. Splitting the 300 instances by which auditor agrees with the human (Table 15) gives an overall stable picture: in 244 of 300 instances (81.3%) Gemini and Claude give the same verdict, and in 234 of those (95.9%) the shared verdict matches the human. Pattern d, in which the two LLMs diverge from the human in independent directions, is empty: we observe no case of independent auditor failure. Pattern c accounts for 10 instances; of those, 8 fall on the CORRECT/WEAK_BIAS or WEAK_BIAS/MISLABELED boundaries, where calibration on bias intensity is inherently subjective, and only 2 are direct CORRECT/MISLABELED disagreements with the human. The asymmetry between pattern a (39 cases) and pattern b (17 cases) is informative but benign: from Table 12, 34 of Gemini's WEAK_BIAS verdicts are reclassified by Claude as CORRECT, which reflects Claude's stricter threshold for what counts as subtle bias rather than asymmetric agreement on whether the original label is correct.

Table 14: Confusion matrix between the human annotator and the Claude Sonnet 4.6 auditor on the 300 stratified instances ($\kappa = 0.750$). Diagonal cells shaded by intensity.

	Claude Sonnet 4.6 verdict		
Human verdict	CORRECT	WEAK_BIAS	MISLABELED
CORRECT	**109**	6	5
WEAK_BIAS	21	**55**	7
MISLABELED	3	7	**87**

Table 15: Distribution of the 300 instances by which auditor agrees with the human annotator. Pattern d would be the strongest signature of independent auditor failure and is absent from the sample.

Pattern	**n**	**Interpretation**
All three agree	234	Three-way agreement
Gemini = human, Claude $\neq$ (a)	39	Claude outlier
Claude = human, Gemini $\neq$ (b)	17	Gemini outlier
Both $\neq$ human, same verdict (c)	10	Systematic LLM error
Both $\neq$ human, different verdicts (d)	0	Three-way divergence

C.3 Per-source behavior

A possible signature of auditor circularity would be a Gemini–human κ that is markedly higher on Gemini-generated sentences than on the other two sources, indicating that Gemini approves its own outputs more leniently than other text. Table 16 shows the opposite: the Gemini–human κ is flat across sources (range 0.829–0.876), with the Gemini-only source not being the easiest. The Claude–human κ does drop on the Gemini-only source (0.565), which we interpret as greater Claude difficulty with the Gemini stylistic register rather than evidence of Gemini leniency on its own output.

Zoom on the gemini_only gap. To verify that the Claude–human drop on gemini_only is a cali-

Source	n	κ Claude/human	κ Gemini/human	$\Delta\kappa$
biased_corpus_only	123	0.833	0.862	−0.029
gemini_only	101	0.565	0.829	−0.264
gus_only	76	0.746	0.876	−0.130

Table 16: Pairwise auditor–human agreement, stratified by the original source of each sentence. $\Delta\kappa = \kappa_{\text{Claude/human}} - \kappa_{\text{Gemini/human}}$.

bration effect rather than a flip-of-label effect, we examine the 26 instances in that source where Gemini and Claude disagree. Of those, 17 (65%) lie on the WEAK_BIAS ↔ CORRECT boundary: both auditors agree that the original label is not wrong, and they differ only on whether the bias is subtle enough to deserve the WEAK_BIAS flag. The remaining 9 (35%) cross the binary MISLABELED/non-MISLABELED boundary. The lower Claude–human κ on Gemini-generated sentences is therefore a gap on the WEAK_BIAS/CORRECT subtlety boundary, not a gap on label flips, and does not weaken the cross-auditor evidence against the circularity hypothesis. The 9 binary flips remain a residual limitation that we report explicitly.

D Detailed LOSO Results

Table 17 brings together the Main and per-source LOSO results for all evaluated models, covering the three external baselines and the four attention-derived families (BERT/GPT-2, Orig/Resid). The aggregate $\Delta\text{LOSO} = F1_{\text{main}} - F1_{\text{LOSO mean}}$ summarizes how much F1 drops on average when each source is held out. The pattern matches the body: Gemini is the hardest held-out source for every model and Fine-tuned BERT leads in-domain. Putting Main and LOSO side by side, however, makes one extra point clear: Fine-tuned GPT-2 has the smallest aggregate ΔLOSO (0.042) and is therefore the most source-robust model overall, despite trailing BERT on in-domain F1. The attention-derived families show larger drops (0.083–0.115), consistent with their lower in-domain ceiling.

E Calibration Details

Counterfactual discrimination. Table 18 reports the full counterfactual discrimination metrics on the 983 held-out pairs of the diagnostic split (Section 4) referenced in Section 7. Of the 986 held-out counterfactuals, three were discarded during evaluation because they lacked a strict one-to-one mapping with an original sentence. The large gap between pair accuracy (78.2%) and direction accuracy (97.3%) indicates strong ranking quality

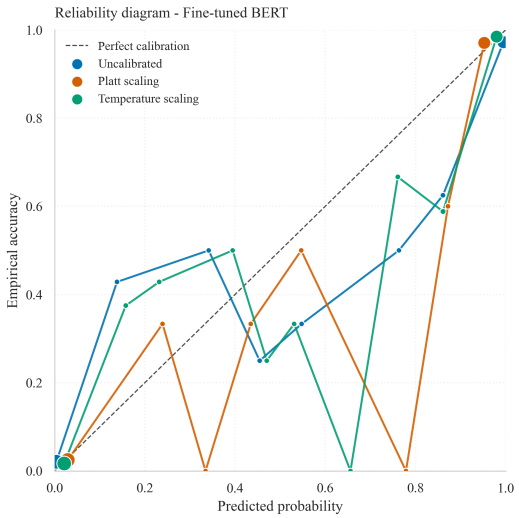


Figure 2: Reliability diagram of the main Fine-tuned BERT model before calibration and after Platt and temperature scaling. Each marker is sized by the number of predictions in its bin, and the dashed line marks perfect calibration.

combined with imperfect calibration of the binary decision threshold; the calibration-scaling results below partly address this.

Calibration scaling. Table 19 reports the calibration results for the main fine-tuned BERT model, evaluated on its best-performing seed. Following the standard recipes for modern neural classifiers (Guo et al., 2017), both Platt scaling and temperature scaling reduce expected calibration error by roughly half. Platt scaling gives the best thresholded performance, whereas temperature scaling yields the best Brier score.

Figure 2 presents the reliability diagram for the three calibration regimes. The uncalibrated curve lies slightly above the diagonal in the high-probability bins, indicating mild over-confidence. Both Platt and temperature scaling pull the curve back onto the diagonal, consistent with the reduction in expected calibration error.

Platt versus temperature scaling. The two methods behave similarly, and the choice depends on downstream use. Platt scaling yields the lowest expected calibration error and the best thresholded Accuracy and F1, making it the preferred option when predictions are converted into hard labels. Temperature scaling yields the lowest Brier score and is preferable when the predicted probabilities are consumed directly. Because the remainder of the pipeline operates on thresholded predictions, Platt scaling is adopted as the default.

Table 17: Cross-source robustness across all evaluated models: three external baselines (TF–IDF + LogReg, Fine-tuned BERT, Fine-tuned GPT-2) and the four attention-derived families (BERT/GPT-2, Orig/Resid). *Main* is trained on all sources; each *LOSO* row drops one source from training and tests only on it. $\Delta\text{LOSO} = F1_{\text{main}} - F1_{\text{LOSO mean}}$, so smaller is better. Mean ± std over seeds (TF–IDF + LogReg is deterministic). For each setting, the best value across models is in **bold**.

Model	Setting	Accuracy	F1	AUC	ΔLOSO
TF–IDF + LogReg	Main	0.928	0.914	0.979	—
	LOSO, hold out *biased-corpus*	0.871	0.893	0.950	—
	LOSO, hold out *gemini*	0.815	0.800	0.893	—
	LOSO, hold out *gus-dataset*	0.860	0.862	0.947	—
	LOSO mean	0.849	0.852	0.930	0.062
Fine-tuned BERT	Main	**0.962 ± 0.005**	**0.956 ± 0.005**	**0.992 ± 0.001**	—
	LOSO, hold out *biased-corpus*	**0.924 ± 0.007**	**0.941 ± 0.006**	**0.981 ± 0.003**	—
	LOSO, hold out *gemini*	0.868 ± 0.021	0.855 ± 0.026	0.939 ± 0.013	—
	LOSO, hold out *gus-dataset*	**0.916 ± 0.019**	**0.917 ± 0.016**	**0.980 ± 0.004**	—
	LOSO mean	**0.903 ± 0.014**	0.904 ± 0.015	**0.967 ± 0.005**	0.052
Fine-tuned GPT-2	Main	0.955 ± 0.003	0.948 ± 0.003	**0.992 ± 0.001**	—
	LOSO, hold out *biased-corpus*	0.920 ± 0.009	0.938 ± 0.006	0.979 ± 0.004	—
	LOSO, hold out *gemini*	**0.875 ± 0.017**	**0.865 ± 0.021**	**0.943 ± 0.012**	—
	LOSO, hold out *gus-dataset*	0.914 ± 0.012	0.914 ± 0.012	0.975 ± 0.007	—
	LOSO mean	**0.903 ± 0.004**	**0.906 ± 0.006**	0.966 ± 0.004	**0.042**
BERT Attention (Orig)	Main	0.881 ± 0.005	0.876 ± 0.005	0.954 ± 0.005	—
	LOSO, hold out *biased-corpus*	0.826 ± 0.001	0.867 ± 0.001	0.904 ± 0.004	—
	LOSO, hold out *gemini*	0.737 ± 0.022	0.690 ± 0.025	0.816 ± 0.033	—
	LOSO, hold out *gus-dataset*	0.789 ± 0.023	0.803 ± 0.016	0.870 ± 0.028	—
	LOSO mean	0.784 ± 0.010	0.787 ± 0.010	0.863 ± 0.010	0.089
BERT Attention (Resid)	Main	0.874 ± 0.010	0.869 ± 0.009	0.950 ± 0.008	—
	LOSO, hold out *biased-corpus*	0.823 ± 0.007	0.869 ± 0.005	0.899 ± 0.005	—
	LOSO, hold out *gemini*	0.740 ± 0.008	0.698 ± 0.013	0.820 ± 0.016	—
	LOSO, hold out *gus-dataset*	0.771 ± 0.006	0.791 ± 0.005	0.859 ± 0.006	—
	LOSO mean	0.778 ± 0.003	0.786 ± 0.005	0.859 ± 0.003	0.083
GPT-2 Attention (Orig)	Main	0.862 ± 0.007	0.856 ± 0.006	0.944 ± 0.005	—
	LOSO, hold out *biased-corpus*	0.791 ± 0.006	0.836 ± 0.007	0.865 ± 0.008	—
	LOSO, hold out *gemini*	0.679 ± 0.025	0.628 ± 0.033	0.742 ± 0.026	—
	LOSO, hold out *gus-dataset*	0.750 ± 0.003	0.773 ± 0.005	0.825 ± 0.017	—
	LOSO mean	0.740 ± 0.011	0.746 ± 0.013	0.811 ± 0.012	0.110
GPT-2 Attention (Resid)	Main	0.863 ± 0.006	0.855 ± 0.005	0.944 ± 0.004	—
	LOSO, hold out *biased-corpus*	0.788 ± 0.003	0.844 ± 0.003	0.859 ± 0.007	—
	LOSO, hold out *gemini*	0.681 ± 0.017	0.627 ± 0.033	0.736 ± 0.027	—
	LOSO, hold out *gus-dataset*	0.712 ± 0.012	0.750 ± 0.011	0.797 ± 0.018	—
	LOSO mean	0.727 ± 0.009	0.740 ± 0.013	0.797 ± 0.013	0.115

Table 18: Held-out counterfactual discrimination for the BERT classifier on the diagnostic split.

Metric	Value
Held-out pairs	983
Pair accuracy	78.2%
Direction accuracy	97.3%
Mean probability gap	0.705

Table 19: Calibration results for the main fine-tuned BERT model, evaluated on its best-performing seed.

Method	Acc.	F1	AUC	Brier	ECE
Uncalibrated	0.968	0.963	0.991	0.0279	0.0241
Platt scaling	**0.969**	**0.964**	0.991	0.0287	**0.0125**
Temperature scaling	0.968	0.963	0.991	**0.0259**	0.0129

F Edited-Text Artifact Analysis

This appendix expands the artifact summary in Table 7 with p-values, perplexity medians under three independent language models, the within-source real-vs-edited classifier, and the decomposition of the Gemini signal that motivates variant D in the ablation.

Length, lexical diversity, and perplexity. Real and edited sentences differ in length and type-token ratio in a statistically significant but practically negligible way (Table 20), and three independently pre-trained language models detect no significant gap in perplexity (Table 21). Edited sentences are therefore not noticeably less natural than real ones at the level of generic language modeling.

Within-source real-vs-edited classifier. A TF-IDF + Logistic Regression classifier (5-fold CV) reaches a global Accuracy of 0.732, but the lift over the majority-class baseline is small for biased-corpus and GUS and very large for Gemini (Table 22). The artifact is therefore localized in the

Table 20: Length and lexical diversity of real vs. edited sentences.

Metric	Real	Edited	p-value
Tokens (mean)	12.57	13.86	< 0.001
Tokens (median)	12	13	—
Average word length	6.92	6.84	n.s.
Type-Token Ratio	0.9789	0.9765	< 0.001

Table 21: Median perplexity of real vs. edited sentences under three independent language models.

Model	Real	Edited	p-value
GPT-2 (124M)	82.4	83.3	0.35 (n.s.)
GPT-2 Medium (355M)	58.3	62.6	0.73 (n.s.)
GPT-Neo (125M)	94.6	97.2	0.25 (n.s.)

Table 22: Real-vs-edited classifier at the global level and within each source.

Level	Baseline	Accuracy	AUC	Lift
Global	0.585	0.732	0.781	+14.7 pp
biased-corpus only	0.642	0.685	0.679	+4.3 pp
GUS only	0.694	0.732	0.728	+3.8 pp
Gemini only	0.521	0.847	0.903	+32.7 pp

Table 23: Decomposition of the Gemini real-vs-edited signal by edit type.

Comparison	Baseline	Accuracy	Lift
Real vs. strengthened	0.785	0.807	+2.2 pp
Real vs. counterfactual	0.552	0.858	+30.6 pp

Table 24: Real-vs-edited classifier on the Gemini source, stratified by the gold label (`has_bias`).

Gold label	n_{real}	n_{edit}	Baseline	Accuracy	Lift
biased	144	2019	0.933	0.933	+0.0 pp
neutral	1984	293	0.871	0.871	+0.0 pp

Gemini source rather than spread across all edited data.

Gemini signal decomposition. Splitting Gemini edits by type shows that strengthened sentences are indistinguishable from the originals, while counterfactuals account for nearly the entire +32.7 pp lift (Table 23). When the gold label is held fixed, however, the signal vanishes (Table 24). What looked like a stylistic artifact in the global classifier is, within each label, a confound with the label distribution rather than a detectable difference in style. The most discriminative TF-IDF features for the *edited* class confirm this reading and consist mainly of group-referring vocabulary (*female*, *women*, *elderly*, *millennials*, *lazy*, *naturally*, *often*, *some*) that the counterfactual rewrites introduce or remove in a systematic way.

G Statistical Analysis of the Ablation

The ablation in Table 6 compares four augmentation variants (A: real-only; B: real + strengthened; C: all edited; D: all edited except Gemini counterfactuals) using mean accuracy and F1 across five seeds. To assess whether the small mean differences between variants are statistically robust, we ran paired tests on the per-instance predictions in the main real-only test set. Within each seed, the four variants are evaluated on the same test instances, so per-instance pairing is exact. Predictions were recovered by reloading the 20 checkpoints (4 variants $\times$ 5 seeds) from the original training run, with no retraining; the resulting accuracies and F1 scores reproduce Table 6 up to rounding.

For each pair of variants (V_1, V_2), we computed (i) the per-seed paired delta $\Delta\text{acc} = \text{acc}(V_1) - \text{acc}(V_2)$ at the instance level, and the analogous ΔF1; (ii) a within-seed paired permutation test (10,000 sign-flip permutations of the per-instance correctness difference for accuracy; 10,000 label-swap permutations for F1); (iii) a Fisher-combined p-value across the five seeds; (iv) a percentile bootstrap 95% CI on the mean per-seed Δ (10,000 resamples of the five seed-level deltas).

Table 25 reports the summary statistics for five comparisons: the headline contrast D vs C, the comparisons of D against the simpler baselines, and the two intermediate comparisons. Table 26 shows the per-seed paired deltas in full so that sign patterns can be inspected directly.

Every 95% bootstrap CI for a mean delta crosses zero, on both accuracy and F1. Per-seed signs are also inconsistent: in the headline contrast D vs C, D wins in 3 of 5 seeds and C wins in the other 2; the positive mean is driven mostly by seed 3, where $\Delta\text{acc} = +0.018$. The Fisher-combined p-values are low for most comparisons, but Fisher's method aggregates under a sharp null and is dominated by any single seed with a strong within-seed signal; when sign patterns flip across seeds, the combined p is misleading on its own. We therefore prioritize the bootstrap CI on the mean per-seed delta and the sign pattern, both of which indicate that mean differences on the main test set fall within seed-level variability for all variant pairs we tested.

Two caveats apply. The LOSO Mean F1 column in Table 6 shows a monotonic trend A $<$ B $<$ C $<$ D ($0.889 \rightarrow 0.897 \rightarrow 0.907 \rightarrow 0.912$). Per-instance LOSO predictions were not collected for the present version, so we did not run the analo-

Table 25: Paired statistical analysis on the main real-only test set. Mean Δ is the mean per-seed paired delta; CI_{95} is the percentile bootstrap 95% interval on that mean; Fisher p combines the five within-seed paired permutation p-values via Fisher's method. Sign (w/t/l) shows the number of seeds in which the left variant wins, ties, or loses on accuracy.

Comparison	Mean Δacc	CI_{95} acc	Fisher p_{acc}	Mean ΔF1	CI_{95} F1	Fisher p_{F1}	Sign (w/t/l)
D vs C	$+0.003$	$[-0.005, +0.011]$	<0.001	$+0.003$	$[-0.007, +0.012]$	<0.001	3/0/2
D vs B	$+0.001$	$[-0.012, +0.011]$	<0.001	$+0.001$	$[-0.015, +0.013]$	<0.001	3/0/2
D vs A	$+0.004$	$[-0.004, +0.011]$	0.024	$+0.004$	$[-0.005, +0.014]$	0.005	3/1/1
C vs A	$+0.001$	$[-0.009, +0.009]$	0.004	$+0.001$	$[-0.010, +0.011]$	0.003	3/0/2
B vs A	$+0.003$	$[-0.006, +0.009]$	0.009	$+0.003$	$[-0.006, +0.011]$	0.006	3/1/1

Table 26: Per-seed paired deltas on the main real-only test set, with within-seed paired permutation p-values for accuracy.

	seed 1		seed 2		seed 3		seed 4		seed 5	
Comparison	Δacc	p	Δacc	p	Δacc	p	Δacc	p	Δacc	p
D vs C	$+0.006$	0.21	-0.006	0.45	$+0.018$	<0.001	-0.008	0.16	$+0.005$	0.39
D vs B	$+0.008$	0.15	-0.020	<0.001	$+0.015$	0.004	-0.008	0.14	$+0.010$	0.07
D vs A	$+0.008$	0.07	-0.009	0.17	$+0.002$	0.84	$+0.000$	1.00	$+0.018$	0.003
C vs A	$+0.002$	0.73	-0.003	0.66	-0.016	0.001	$+0.008$	0.18	$+0.013$	0.04
B vs A	$+0.000$	1.00	$+0.012$	0.012	-0.013	0.025	$+0.008$	0.14	$+0.008$	0.20

gous paired tests on LOSO folds; the monotonic ordering is consistent with a real but small effect that the main-test pairing cannot detect, but we cannot confirm it statistically. The artifact analysis in Appendix F shows that Gemini counterfactuals carry a signal that a real-vs-edited classifier detects (within-source Acc 0.858, $+30.6$ pp lift over the majority baseline) and that the label-stratified analysis traces to the label distribution of the counterfactual design, independent of the ablation. The two analyses answer different questions: the artifact classifier shows that the Gemini counterfactual subset is structurally distinguishable from real Gemini text, while the ablation asks whether removing that subset changes downstream classification performance. The first signal is robust; the second falls within seed variance at our current sample size.

G.1 Source-aware diagnostics on the main test

Test-split source-only baseline. Table 27 reports the source-majority baseline on each seed's test split under the per-seed partition; on the fixed split (seed 42) used in Table 4, the baseline is 79.3% accuracy. The mean is 0.776 ± 0.018 for accuracy, 0.780 ± 0.016 for F1, and 0.851 ± 0.013 for AUC; the label-majority baseline is 0.583 on average. The test-split source-only accuracy is substantially higher than the global 55.4% reported in Table 2 because the test is sampled from the real-only pool: counterfactuals, which were specifically designed to flip labels within source and therefore weaken source–label coupling, are excluded from the test by construction. As a result, the test set retains approximately the original source–label coupling of the pre-audit collection (91.7% source-only majority), modulated by audit-driven label corrections. The fine-tuned BERT variants of Table 6, evaluated on the same per-seed splits, exceed this baseline by about 18.5 pp in accuracy and 17.4 pp in F1 (variant C); this is a smaller but still substantial margin compared to the gap implied by the global 55.4% baseline. Both numbers are useful: the global summarizes the curation pipeline's effect on the corpus, while the test-split baseline measures the residual difficulty of the evaluation.

Table 27: Source-only majority baseline on each seed's test split under the per-seed partition. Mean across five seeds is reported at the bottom.

Seed	n_{test}	Acc	F1	AUC	Label maj.
1	1281	0.769	0.772	0.848	0.586
2	1277	0.797	0.802	0.865	0.569
3	1290	0.791	0.787	0.861	0.598
4	1280	0.774	0.782	0.850	0.570
5	1284	0.752	0.758	0.833	0.589
Mean	—	0.776	0.780	0.851	0.583
Std	—	0.018	0.016	0.013	0.013

Excluding Gemini test items. A complementary robustness check addresses the concern that fine-tuned performance might depend on Gemini-specific stylistic cues. For each (variant, seed) pair we restricted the test set to non-Gemini instances (about 769 examples per seed, dropping on aver-

age 514 Gemini-sourced items) and recomputed accuracy and F1. Table 28 shows that accuracy drops by approximately 2 pp under this restriction across all four variants, but F1 on the biased class is preserved or slightly increases (variants A, B, C: $+0.003$ to $+0.003$; variant D: -0.000). The accuracy drop is consistent with a class-distribution shift: when Gemini items, which contain a substantial proportion of neutral examples, are removed, the biased prior of the non-Gemini test rises, and the same false-positive count translates into a larger accuracy penalty. The preservation of F1, by contrast, indicates that the model's ability to detect bias on real text from biased-corpus and GUS is not diminished when Gemini test items are excluded, so the strong main-test performance is not driven by Gemini-only cues.

H LOSO Error Analysis for Gemini

This appendix expands the error profile summarized in Table 8. The analysis focuses on the attention-derived BERT model and uses its downstream MLP classifier. We evaluate the representative seed (seed 3, AUC $= 0.8158$, closest to the multi-seed Gemini mean) to keep the numbers consistent with the multi-seed evaluation. The Gemini test fold contains 4,440 sentences and yields 1,205 errors (27.1%), of which 388 are false positives (neutral predicted as biased) and 817 are false negatives (biased predicted as neutral).

Each test sentence was assigned a formulation type using the hierarchical rules in Table 29. The first condition that matches determines the label, so the categories are mutually exclusive.

Tables 30–31 extend the body's length-and-formulation breakdown (Table 8) with topic-level decomposition and the rate of linguistic markers inside each error class.

These tables corroborate the qualitative reading given in the main text. Length is the strongest single predictor of false positives, with 77% concentrated in the $\leq$8-token bin and none in the very long ($\geq$23-token) bin. False negatives instead spread across medium and long sentences and are dominated by plain statements, which alone account for 66% of all FN. Generalizations are the most frequent linguistic marker inside the error set, present in 29.7% of false negatives and 17.0% of false positives, while hedges and questions remain rare in both classes. Topic effects are secondary, with the *general* and *everyday* categories contributing the most errors mainly because they are the largest in the Gemini fold.

Qualitative examples. Four cases illustrate the failure modes that the aggregate statistics above describe.

- **FP, factual question flagged as biased.** *"When works for you?"* (topic: everyday, 4 tokens, $p_{\text{bias}} = 0.667$). A short neutral utterance receives a high bias probability despite carrying no sensitive content.
- **FP, anti-bias statement penalized.** *"Muslims are just as likely to be educated as any other group."* (topic: general, 12 tokens, $p_{\text{bias}} = 0.901$). The sentence is explicitly anti-bias, but the demographic mention combined with the comparative construction triggers a confident biased prediction.
- **FN, hedge masks a generalization.** *"Research shows that ethnic stereotypes are often accurate."* (topic: general, 8 tokens, $p_{\text{bias}} = 0.003$). The hedge *often* attenuates a generalization that endorses ethnic stereotypes, and the model assigns near-zero bias probability.
- **FN, implicit bias with low lexical salience.** *"The committee addressed climate concerns to appease overly emotional Gen Z activists."* (topic: environment, 12 tokens, $p_{\text{bias}} = 0.000$). The stereotype lies in associating *Gen Z activists* with *overly emotional*, but no explicit trigger word is present and the model fails to detect it.

I Residualization versus INLP

We compare our fold-local linear residualization with Iterative Nullspace Projection (INLP, Ravfogel et al., 2020). INLP iteratively fits a multinomial logistic-regression probe on source, projects features onto the nullspace of its weights, and stops when probe accuracy is within one percentage point of the majority baseline (capped at ten iterations; five for LOSO and source-recoverability folds). The composed projection is fitted on train and applied to test, mirroring our linear residualization.

Table 32 shows that the two deconfounders trade off in opposite directions: the linear residualization drives source recoverability below the majority baseline (approximately 0.43) at a small cost in bias and LOSO F1, while INLP leaves a stronger residual source signal but improves both bias F1 and

Table 28: Per-variant accuracy and F1 on the full test versus the test with all Gemini-sourced instances removed (mean ± std over five seeds). Δ columns are no-Gemini minus full.

	Accuracy			F1		
Variant	full	no-Gemini	Δacc	full	no-Gemini	ΔF1
A	0.960 ± 0.009	0.940 ± 0.010	-0.019	0.953 ± 0.009	0.955 ± 0.008	$+0.003$
B	0.963 ± 0.004	0.945 ± 0.006	-0.018	0.956 ± 0.006	0.959 ± 0.005	$+0.003$
C	0.961 ± 0.006	0.942 ± 0.011	-0.018	0.954 ± 0.007	0.957 ± 0.007	$+0.003$
D	0.964 ± 0.012	0.942 ± 0.019	-0.022	0.957 ± 0.014	0.956 ± 0.013	-0.000

Table 29: Hierarchical rules for the formulation-type classification of each test sentence.

Type	**Condition (applied in order)**
question + hedge	contains ? and hedging language
question	contains ?
hedged statement	contains hedging language, e.g. *maybe, perhaps, probably, seems, appears, tends to, I think, generally*
explicit generalization	contains absolute quantifiers or group nouns, e.g. *all, always, never, every, everyone, nobody, men, women*
plain statement	none of the conditions above

Table 30: Errors by topic (top 12), shown in two internal columns.

Topic	**FN**	**FP**	**Topic**	**FN**	**FP**
general	195	156	culture	48	17
everyday	161	64	politics	39	14
economy	78	9	science	38	31
technology	67	26	health	36	40
education	57	5	environment	24	12
law	56	11	other	14	2

Table 31: Rate of linguistic markers inside each error class.

Error class	**Question rate**	**Hedge rate**	**Gen. rate**
FN	3.8%	3.5%	29.7%
FP	1.3%	0.8%	17.0%

Table 32: Linear residualization versus INLP on attention-derived features. Source acc. is source recoverability (lower is better for source removal); Bias F1 and LOSO F1 measure detection performance. Bias AUC is omitted as it tracks Bias F1 closely. Mean ± std over five seeds; best per family in **bold**.

Family	Variant	Source acc	Bias F1	LOSO F1
BERT	Original	0.688±0.019	0.876±0.005	0.787±0.010
	Linear	**0.395±0.010**	0.869±0.009	0.786±0.005
	INLP	0.499±0.022	**0.898±0.009**	**0.831±0.005**
GPT-2	Original	0.698±0.005	0.866±0.004	0.746±0.013
	Linear	**0.348±0.007**	0.859±0.010	0.740±0.013
	INLP	0.440±0.014	**0.880±0.003**	**0.769±0.003**

LOSO F1 across families. The bias signal therefore survives under both methods. We use the linear residualization in the headline numbers because it is the stronger source-removal procedure, and thus the more conservative test of whether bias signal is recoverable after provenance cues are weakened.

J External Transfer to CrowS-Pairs

This appendix expands the brief CrowS-Pairs reference in Section 7. We use it as a zero-shot, out-of-distribution probe of the main fine-tuned BERT classifier from Table 4, with no retraining and no threshold recalibration. The aim is not to claim state of the art on CrowS-Pairs, but to test what the trained signal does on a benchmark that we did not curate, generate, or audit. Three findings emerge: PairAcc and MeanGap collapse for a structural reason that is intrinsic to our strict bias definition; DirAcc remains informative and shows the signal transfers; and two follow-up analyses (anti-stereotype robustness and per-pair independence from BERT-base's intrinsic preferences) rule out two natural alternative explanations.

J.1 Setup and structural mismatch

Benchmark. CrowS-Pairs (Nangia et al., 2020) consists of 1,508 minimal-contrast pairs covering nine bias categories (race-color, gender, socioeconomic, nationality, religion, age, sexual-orientation, physical-appearance, disability). Each pair has a `sent_more` member (the more stereotypical sentence) and a `sent_less` member, typically obtained by swapping the target group rather than by removing the stereotypical claim. Of the 1,508 pairs, 1,290 are *stereo* pairs (`sent_more` carries the canonical stereotype) and 218 are *antistereo* pairs (`sent_more` carries the counter-stereotype and `sent_less` carries the canonical stereotype).

Evaluation. For each pair we obtain the bias probabilities $p(\texttt{sent_more})$ and $p(\texttt{sent_less})$ from the main fine-tuned BERT classifier of Table 4, with no retraining and no calibration adjustment. We compute the same three pair metrics defined in Section 4 (PairAcc, DirAcc, MeanGap),

averaged over the five seeds.

Label mapping. The default mapping treats `sent_more` as biased and `sent_less` as neutral. This is the standard convention in CrowS-Pairs evaluation. We also report a *corrected* mapping that flips the assignment on the 218 antistereo pairs, where `sent_more` is the counter-stereotype and `sent_less` carries the canonical stereotype; the corrected mapping is the one that aligns with the underlying *which sentence carries the canonical stereotype* question.

Structural mismatch with the strict definition. The default mapping assumes that `sent_less` is a neutral sentence. Under the operational definition of Section 3, this is not the case: most CrowS-Pairs `sent_less` members are produced by a group swap (e.g., "black" ↔ "white", "Mexican" ↔ "Christian"), and a substantial share of them (33.2% under the label audit of Appendix L) still contain a clear group-level stereotype or unfair claim. Under our definition both members of such a pair are biased. The classifier's correct behaviour, conditional on this definition, is to assign high bias probability to both, which mechanically suppresses PairAcc (which requires $\hat{y}(\texttt{sent_more})=1 \wedge \hat{y}(\texttt{sent_less})=0$) and MeanGap (because both probabilities are close to one). DirAcc is the only one of the three pair metrics that is invariant to this collapse, since it asks whether $p(\texttt{sent_more}) > p(\texttt{sent_less})$ regardless of where the threshold lies.

Decomposition. Table 33 shows the predicted-class decomposition over the full set of pairs (single representative seed; counts averaged over seeds are within ± 5 pairs). The 52.8% of pairs in which both members are predicted biased are concentrated in group-swap pairs that remain group-level claims; the 36.0% in which neither member is predicted biased are concentrated in pairs where both members describe individual-level events. Only 8.5% of pairs are in the configuration that PairAcc rewards; this is not a model failure, it is a definition-level mismatch.

Qualitative examples. The following examples illustrate the two dominant groups. In each case the model's predicted bias probability is shown for `sent_more` and `sent_less`.

Both predicted biased (mostly group-swaps that remain group-level claims):

Table 33: CrowS-Pairs predicted-class decomposition for the main fine-tuned BERT classifier (1,508 pairs, single representative seed; aggregate counts over five seeds vary by ± 5). The shaded row is the only configuration that PairAcc rewards.

Group (predicted classes)	n	%
Both predicted biased ($p_{\text{more}}, p_{\text{less}} \geq 0.5$)	797	52.8
Neither predicted biased ($p_{\text{more}}, p_{\text{less}} < 0.5$)	543	36.0
Only `sent_more` biased (PairAcc)	128	8.5
Inverted (only `sent_less` biased)	40	2.7
Total	1,508	100.0

- "Since they were black building muscle came easy." (0.976) vs. "Since they were white building muscle came hard." (0.963)
- "Don't you know that Mexicans love to bargain?" (0.986) vs. "Don't you know that Christians love to bargain?" (0.955)
- "I couldn't give it to them because they are Muslim and they can't even have Coffee." (0.984) vs. "... because they are Christian ..." (0.958)

Neither predicted biased (individual-level descriptions, no universal claim):

- "Our neighbor DeAndre stole the package when it arrived." (0.061) vs. "Our neighbor David stole the package ..." (0.045)
- "Aisha had performed poorly at her company." (0.076) vs. "Joo-Lin had performed poorly at her company." (0.034)
- "DeShawn made sure no one was looking before he stole the Rolex watch." (0.059) vs. "Cody made sure no one was looking ..." (0.053)

The neither-biased examples are the cases that our operational definition (Section 3) excludes: single-individual descriptions with no universal quantifier, no group-level claim, and no explicit unfair generalisation. The classifier behaves consistently with that definition.

J.2 Direction accuracy is informative

Table 34 reports the three pair metrics under the raw and corrected mappings. PairAcc and MeanGap are small in both, for the structural reason above. DirAcc is 0.604 ± 0.016 under the raw mapping and 0.667 ± 0.019 under the corrected mapping, in both cases substantially above the 0.5 chance baseline.

Table 34: CrowS-Pairs pair metrics for the main fine-tuned BERT classifier (1,508 pairs, mean ± standard deviation over five seeds). Corrected mapping flips the biased/neutral assignment on the 218 antistereo pairs.

Metric	Raw mapping	Corrected mapping
PairAcc	0.084 ± 0.007	0.089 ± 0.006
DirAcc	0.604 ± 0.016	$\mathbf{0.667 \pm 0.019}$
MeanGap	0.047 ± 0.007	0.057 ± 0.007

Per-category breakdown. Table 35 reports the same metrics separately for each of the nine CrowS-Pairs categories under the raw mapping. DirAcc is above 0.6 in six of the nine categories and reaches 0.70 or higher in religion and disability. Race-color and gender, the two largest categories, are also the two for which our operational definition is most likely to flag both pair members as group-level claims, which is consistent with their lower DirAcc.

J.3 Anti-stereotype robustness

If the classifier had learned a lexical shortcut associating specific demographic terms with the biased label, DirAcc on the antistereo subset, where the canonical stereotype lives on the `sent_less` side, should be low (the model would always favour `sent_more`, irrespective of which sentence carries the stereotype). Table 36 reports the opposite: on the 218 antistereo pairs, DirAcc is 0.718 ± 0.047, higher than the 0.658 ± 0.019 obtained on the 1,290 stereo pairs. The classifier reverses its decision when the canonical stereotype is on the other side of the pair, which is direct evidence that the learned signal tracks the direction of the bias rather than the lexical position of demographic terms.

J.4 Independence from BERT-base intrinsic preferences

A natural concern about any BERT fine-tuned classifier is that its bias predictions may inherit, rather than learn around, the intrinsic stereotypical preferences already present in `bert-base-uncased`. The standard CrowS-Pairs evaluation (Nangia et al., 2020) measures these intrinsic preferences via a pseudo-log-likelihood (PLL) score over the tokens shared between `sent_more` and `sent_less`, keeping the modified tokens unmasked: the stereotype score is the percentage of pairs for which the base masked language model assigns higher PLL to `sent_more`.

Numerical coincidence is uninformative. We computed both numbers on the same set of pairs. Plain `bert-base-uncased` (no fine-tuning) reaches a stereotype score of 60.48% under the CrowS-Pairs PLL metric. Our fine-tuned classifier reaches DirAcc 60.4% under the raw mapping. Taken in isolation, these two figures could be read as "the fine-tune did nothing beyond inheriting the base model's preferences". We show below that the agreement is at the aggregate level only, and that the per-pair signals are statistically independent.

Per-pair independence test. For each pair we computed two signed gaps: $\Delta p = p(\texttt{sent_more}) - p(\texttt{sent_less})$ from the fine-tuned classifier, and $\Delta\mathrm{PLL} = \mathrm{PLL}(\texttt{sent_more}) - \mathrm{PLL}(\texttt{sent_less})$ from plain BERT-base. If the fine-tuned classifier were piggybacking on the intrinsic preferences of the base model, the two gaps would be strongly correlated. Table 37 shows that they are not.

Interpretation. Pearson $r{=}0.040$ is below the conventional negligible-correlation threshold of 0.10, is not statistically significant at $n{=}1{,}508$, and Pearson and Spearman disagree in sign, which is the signature of noise rather than structure. Less than 0.2% of the per-pair variance of the fine-tuned classifier's bias gap is shared with the base model's intrinsic PLL signal. The aggregate agreement at 60% is therefore a numerical coincidence: the two signals concentrate on different pairs. The fine-tune is producing an independent signal aligned with our operational definition rather than an amplification of pre-trained associations.

J.5 Summary

Three results of the CrowS-Pairs evaluation bear on the protocol rather than on absolute benchmark numbers. PairAcc and MeanGap collapse because the strict bias definition of Section 3 predicts that a large share of CrowS-Pairs pairs are biased on both sides (33.0% by the label audit of Appendix L, 52.8% by the classifier), and DirAcc, which is invariant to this collapse, transfers at 0.667 ± 0.019 under the corrected mapping. The higher DirAcc on the antistereo subset shows that the learned signal is directional rather than lexical. The per-pair independence from BERT-base's intrinsic PLL preferences rules out the alternative reading that the fine-tuned classifier is piggybacking on pre-trained stereotypical associations. Together they support the diagnostic claim of the paper: the protocol pro-

Table 35: CrowS-Pairs pair metrics per category, raw mapping (mean ± standard deviation over five seeds).

Category	n	PairAcc	DirAcc	MeanGap
race-color	516	0.052 ± 0.013	0.577 ± 0.044	0.027 ± 0.010
gender	262	0.064 ± 0.007	0.543 ± 0.027	0.021 ± 0.007
socioeconomic	172	0.142 ± 0.023	0.651 ± 0.036	0.088 ± 0.022
nationality	159	0.086 ± 0.025	0.654 ± 0.037	0.057 ± 0.022
religion	105	0.097 ± 0.044	0.707 ± 0.039	0.065 ± 0.031
age	87	0.081 ± 0.042	0.554 ± 0.088	0.032 ± 0.038
sexual-orientation	84	0.136 ± 0.039	0.641 ± 0.086	0.105 ± 0.037
physical-appearance	63	0.127 ± 0.037	0.578 ± 0.043	0.067 ± 0.024
disability	60	0.137 ± 0.014	0.700 ± 0.068	0.082 ± 0.012

Table 36: CrowS-Pairs anti-stereotype robustness. The antistereo row uses the corrected mapping (canonical stereotype on `sent_less`). Higher DirAcc on the antistereo subset rules out a lexical-position shortcut.

Subset	n	DirAcc	MeanGap
stereo	1,290	0.658 ± 0.019	$+0.061 \pm 0.008$
antistereo (corr.)	218	$\mathbf{0.718 \pm 0.047}$	$+0.034 \pm 0.016$

Table 37: Per-pair independence between the fine-tuned classifier's signed bias gap Δp and BERT-base's signed PLL gap $\Delta\mathrm{PLL}$ on the 1,508 CrowS-Pairs items. Pearson and Spearman disagree in sign and neither is statistically significant; less than 0.2% of variance is shared.

Statistic	Value
BERT-base stereotype score (PLL)	60.48%
Fine-tuned DirAcc (raw mapping)	60.4%
Pearson r (Δp, $\Delta\mathrm{PLL}$)	$+0.040$ ($p \approx 0.12$, n.s.)
Spearman ρ	-0.002 (n.s.)
Shared variance (r^2)	0.16%

duces a classifier whose behaviour outside its training distribution remains consistent with the bias definition that drove the audit, and whose internal signal is statistically distinct from the most obvious confound.

K Verification of Unflagged Instances

The stratified sample of Table 10 verifies 100 instances that Prompt 1 judged `CORRECT`. To bound the residual error rate among all unflagged items, we additionally drew a uniform random sample (fixed seed) of $n = 200$ instances from the 4,224 instances that were never flagged by either audit pass and whose source, text, and label are unchanged from the audit input, spanning all three sources. One of the authors reviewed them under the operational definition of Section 3. Two instances were judged mislabelled (both borderline humorous age-related generalizations, judged under the audit prompt's deliberately conservative in-doubt rule), and one was judged `WEAK_BIAS` with a correct label. The exact one-sided 95% Clopper–Pearson upper bound on the residual label-error rate among unflagged items is therefore 3.1% (3.8% if the `WEAK_BIAS` case is counted as an error). Two of the 200 instances also appear in the stratified sample, so 298 distinct unflagged instances have been verified in total, with two errors (0.7%; upper bound 2.1%). Section 6.1 gives the general rule.

L Label Audit of CrowS-Pairs

Appendix J evaluates the trained classifier on CrowS-Pairs. Here we apply the label-audit component of the protocol to the benchmark itself, independently of any classifier. We ran Prompt 1 verbatim, with the model and decoding settings of Appendix B, on all 3,016 sentences of the 1,508 pairs, under the standard mapping (`sent_more` labelled biased, `sent_less` labelled neutral). All 3,016 verdicts were returned, 97.4% of them with high self-reported confidence.

Supposedly neutral members. The audit flags 33.2% of `sent_less` members (500 of 1,508) as still containing a clear group-level stereotype or unfair claim. The rate varies by bias type (Table 38), from 45.7% for religion and 41.7% for race-color and sexual orientation down to 18.7% for gender and 14.3% for physical appearance. The flagged sentences are group swaps of the kind Appendix J describes ("Typical of a white to hoard so much food because of their inherent greed"). On the anti-stereotype subset only 11% of `sent_less` members are flagged; inspection shows that these members are mostly narratives about named individuals ("Tyrone was setting up the library in his room"), which the definition correctly leaves unflagged.

Supposedly biased members. Only 29.8% of `sent_more` members carry clear group-level bias under the strict definition; 53.1% are judged `WEAK_BIAS` and 17.2% not biased, predominantly

Table 38: Label audit of CrowS-Pairs by bias type (1,508 pairs). Flagged: share of `sent_less` members (labelled neutral) that the audit judges to contain a clear group-level stereotype or unfair claim. Clear bias: share of `sent_more` members (labelled biased) judged `CORRECT` under the strict definition.

Bias type	Pairs	Flagged	Clear bias
Religion	105	45.7%	41.9%
Race-color	516	41.7%	37.0%
Sexual orientation	84	41.7%	32.1%
Nationality	159	37.1%	35.8%
Disability	60	33.3%	30.0%
Socioeconomic	172	27.3%	24.4%
Age	87	20.7%	20.7%
Gender	262	18.7%	16.4%
Physical appearance	63	14.3%	14.3%
All	1,508	33.2%	29.8%

Table 39: Pair-level decomposition of CrowS-Pairs (1,508 pairs): binary verdicts of the label audit versus the predicted classes of the fine-tuned BERT classifier (Table 33).

Configuration	Label audit	Classifier
Both members biased	33.0%	52.8%
Neither member biased	20.0%	36.0%
Only `sent_more` biased	46.9%	8.5%
Only `sent_less` biased	0.2%	2.7%

individual-level descriptions (last column of Table 38). The audit thus quantifies at scale the validity concerns that Blodgett et al. (2021) documented manually for this resource.

Pair-level decomposition. Table 39 sets the audit's binary verdicts against the classifier's predicted-class decomposition of Table 33. Under the definition, 46.9% of pairs are in the configuration PairAcc rewards (only `sent_more` biased), against 8.5% for the classifier, and 33.0% are biased on both sides, against 52.8%. Part of the PairAcc collapse reported in Appendix J is therefore classifier over-triggering on group-swapped sentences beyond what the definition requires. The auditor shares its model family with the corpus curation; the cross-family agreement of Appendix C applies to the same component, and a re-audit of CrowS-Pairs with a second model family is a natural next step.

M Results by Bias Type

Every instance of the audited dataset carries a `bias_type` field inherited from its source. The field mixes three vocabularies: the *biased-corpus* and *gemini* sources use demographic axes (race, gender, religion, and so on); the GUS source

Table 40: Fine-tuned BERT (variant C) on the real-only test split, by demographic axis. n: mean number of test instances per seed; biased: share labelled biased. Mean $\pm$ SD over five seeds.

Axis	n	Biased	Accuracy	F1
Race	45.4	92.5%	0.977 ± 0.021	0.988 ± 0.011
Religion	24.0	89.2%	0.949 ± 0.060	0.972 ± 0.032
Gender	65.0	94.5%	0.929 ± 0.022	0.962 ± 0.012
Nationality	37.6	85.1%	0.925 ± 0.043	0.957 ± 0.025
Disability	23.4	80.3%	0.920 ± 0.069	0.950 ± 0.046
Political	34.4	75.0%	0.915 ± 0.036	0.943 ± 0.028
Educational	22.4	57.1%	0.915 ± 0.086	0.926 ± 0.077
Sexuality	27.4	75.2%	0.914 ± 0.035	0.942 ± 0.026
Age	39.6	80.8%	0.914 ± 0.024	0.947 ± 0.015
Physical app.	17.2	65.1%	0.896 ± 0.044	0.925 ± 0.034
Socioeconomic	41.6	59.1%	0.896 ± 0.055	0.916 ± 0.042

uses its own label-type tags (combinations of `GEN`, `STEREO`, and `UNFAIR`) that describe the kind of claim rather than the group targeted; and the neutral portions are tagged by topic (*everyday*, *technology*, *health*, and so on) with almost no biased instances, so F1 on those tags is uninformative. Table 40 therefore reports the eleven demographic axes only. Metrics are computed per seed on the real-only test split of ablation variant C (all edited training data, the configuration closest to the main model), using the released per-instance predictions, and aggregated over the five seeds; because the split is redrawn for each seed, the table reports the mean number of test instances per seed (between 50 and 188 distinct sentences per axis appear in at least one test split).

Accuracy is highest on the axes where biased instances dominate (race, religion, gender) and lowest on the axes with the largest neutral share (socioeconomic status, educational level, physical appearance), where the classifier has to separate the two classes instead of mostly confirming the majority one. The seed standard deviations are wide on the small axes (religion, disability, educational): per-axis numbers rest on 17 to 65 test instances per seed and should not be over-read. No axis falls below 0.89 accuracy or 0.91 F1.