When Counterfactual Debiasing Backfires: Identity Cues as Task-Dependent Evidence in Harmful Language Understanding

Anonymous ACL submission

Abstract

Counterfactual debiasing is often used to reduce reliance on identity-linked shortcuts in harmful language understanding. We show that the same intervention can be harmful when identity mentions are not shortcuts but label-defining evidence. In multi-task cyberbullying modeling, identity cues may be spurious for cyberbullying detection (CD) and sentiment analysis (SA), yet necessary for fine-grained target identification (TI). We propose task-aware routing, which sends CD and SA through a debiased representation while preserving the original representation for TI. We evaluate this approach on the English-translated BullyExplain benchmark, HateXplain, and an external Measuring Hate Speech diagnostic. Experiments show that task-aware routing improves TI, while auxiliary detection and sentiment performance remain largely unchanged where applicable. Ablation studies attribute this improvement to representation routing that preserves identity-bearing evidence for TI rather than to class-weighting choices, with reduced-architecture diagnostics suggesting that the effect is not specific to the graph-enhanced model. These results support a task-conditional view of counterfactual debiasing: the same identity cue can warrant suppression in one task and preservation in another.

1 Introduction

Counterfactual interventions are a standard tool for reducing reliance on identity-linked shortcuts in language models (Kaushik et al., 2020; Wu et al., 2021; Feder et al., 2022). Many such interventions implicitly treat identity terms as bias features whose suppression should not harm the underlying task. In harmful language understanding, this assumption holds for some tasks but not for others.

Understanding harmful language often requires more than detecting whether a post is abusive: systems may also need to assess its sentiment and identify which group or identity is being targeted. We therefore consider three related tasks: cyberbullying detection (CD), sentiment analysis (SA), and target identification (TI). For CD and SA, identity mentions may correlate with labels in ways that the model should not exploit. For fine-grained TI, the same identity mentions are precisely the evidence that distinguishes target classes such as *Religion*, *Gender*, or *Race*. A single counterfactual intervention applied uniformly across heads can therefore benefit tasks where identity cues should be suppressed while damaging target identification, because it removes the very information TI is defined to use.

Whether identity functions as a *shortcut* or as *evidence* determines which intervention is appropriate. Figure 1 illustrates the problem: the same identity-bearing cue may be spurious for CD and SA but necessary for TI, whose target labels are partly defined by identity. The Method section formalizes three routing regimes that operationalize this distinction. We then test whether this routing pattern holds across two harmful-language benchmarks, an external diagnostic corpus, and a simplified text-only backbone.

Contributions. We identify and characterize a task-conditional failure mode of uniform counterfactual debiasing:

- We identify a shortcut-versus-evidence distinction: the same identity terms can be spurious features for detection yet label-defining evidence for fine-grained target identification.
- We document this failure mode empirically across two main datasets and an external diagnostic corpus. Across BullyExplain, HateXplain, and Measuring Hate Speech, the evidence shows that uniform debiasing can suppress TI-relevant identity information, while task-aware routing recovers TI

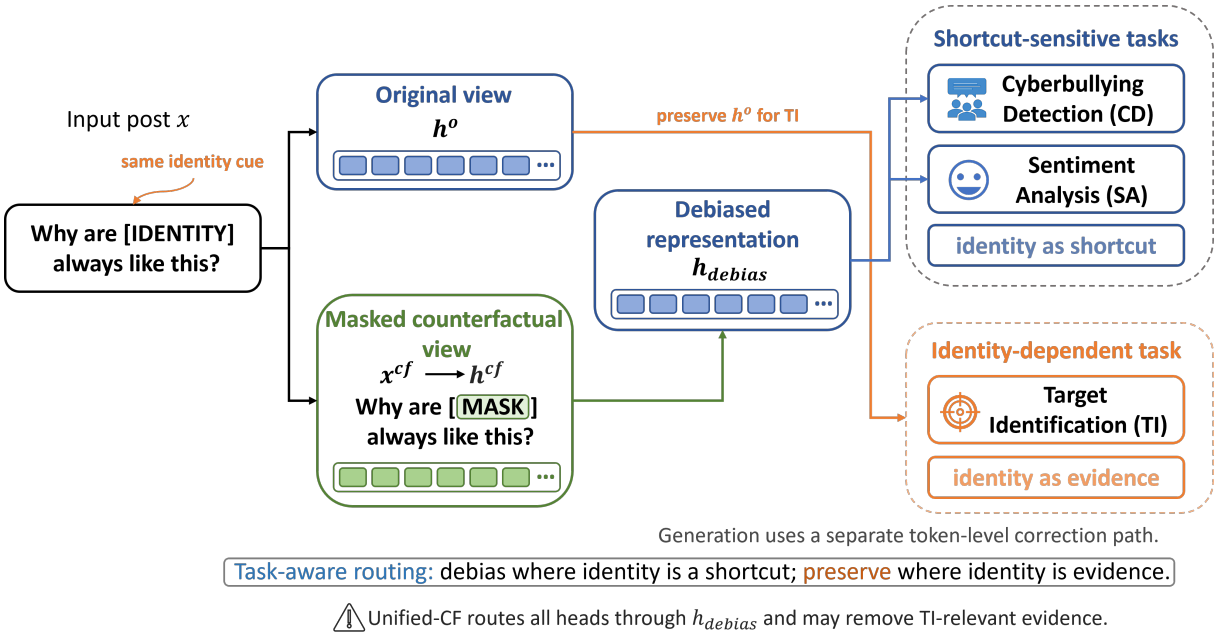

Figure 1: Conceptual task-aware routing. The original representation is preserved for TI, where identity cues provide label-defining evidence, while CD and SA—for which identity cues can act as shortcuts—receive a debiased representation derived from original and counterfactual views. Unified counterfactual routing sends the debiased representation to all heads and can suppress TI-relevant identity evidence.

most clearly under stronger masking and on identity-dependent classes.

- We map out when the routing benefit emerges: a routing-by-weighting ablation attributes the TI gain to representation routing rather than weighting, while intervention-strength, identity-lexicon coverage, and RoBERTa-base backbone diagnostics identify the conditions under which the effect arises.

2 Related Work

Bias and debiasing in NLP. Prior work has documented multiple forms of representational and social bias in language technologies (Blodgett et al., 2020; Hovy and Prabhumoye, 2021; Gallegos et al., 2024). Debiasing approaches include adversarial learning (Zhang et al., 2018), representation projection (Ravfogel et al., 2020), and benchmark-driven evaluations such as StereoSet and BiasBench (Nadeem et al., 2021; Meade et al., 2022). Much of this literature treats identity-linked information as something to be suppressed. In our setting, the same identity information can be either a shortcut or a label-defining feature, so a single suppression policy is inadequate.

Counterfactual interventions and shortcut learning. Counterfactually augmented data and related interventions have been used to reduce reliance on superficial correlations and improve robustness (Kaushik et al., 2020; Wu et al., 2021; Feder et al., 2022). A parallel line of work characterizes shortcut learning and competency failures in NLP models more broadly (Geirhos et al., 2020; Gardner et al., 2020), and recent work develops causal frameworks for evaluating and mitigating spurious correlations in text classification (Veitch et al., 2021; Joshi et al., 2022). These approaches share an implicit assumption: that the features being suppressed are not themselves part of the task signal. We show that this assumption does not hold uniformly for fine-grained target identification, where identity expressions are precisely the evidence that distinguishes target classes.

Task- and context-dependent debiasing. A smaller body of work has noted that bias-mitigation interventions can have different effects across tasks or contexts: intrinsic bias metrics need not align with downstream behavior (Goldfarb-Tarrant et al., 2021), gender debiasing can shift internal representations in task-specific ways (Orgad et al., 2022), and upstream mitigation does not always transfer to downstream applications (Steed et al., 2022). Closest to our setting, Sen et al. (2022) show that counterfactually augmented data can introduce unintended bias in single-task hate speech detection.

We extend this observation to the multi-task setting: when the same identity perturbation must serve multiple heads simultaneously, the intervention can help when identity cues are spurious while harming fine-grained target identification, depending on whether identity terms define the label. Unlike gradient-conflict approaches in multi-task learning (Yu et al., 2020; Standley et al., 2020), our conflict is between representational transformations, not optimization directions.

Explanation generation in harmful language modeling. Multi-task systems coupling classification with rationale or explanation generation have shown promise for interpretability (Jain et al., 2020; Yu et al., 2021; Chan et al., 2022; Treviso et al., 2023). We use GenEx (Maity et al., 2023) as the closest published cyberbullying comparison point, but our focus is counterfactual routing across task heads rather than explanation modeling.

3 Method

To test whether counterfactual debiasing should be applied uniformly or in a task-conditional way, we compare three routing regimes: Unified-CF, task-aware routing, and No-CF. The regimes share the BART backbone, graph branch, and multi-head decoder; the counterfactual masking branch is active in Unified-CF and task-aware routing and disabled in No-CF. Because the initial task-aware configuration also changes the TI imbalance treatment, we further use a controlled factorial ablation to separate the effect of routing from the effect of weighting.

3.1 Problem Setup

Given an input post x, the model jointly predicts: (1) a cyberbullying label y^{cd}, (2) a sentiment label y^{sa}, (3) a target label y^{ti}, and (4) an explanation sequence y^{gen}. Our goal is to determine which sentence representation each classification head should receive when identity-bearing tokens may function as either shortcut features or target-defining evidence. Explanation generation is treated as a coupled but separate token-level correction path.

3.2 Architecture and Routing

For each of the two inputs—the original post and its counterfactually masked counterpart, constructed as described below—BART produces token-level representations that pass through a shared graph-enhanced branch. The branch is a 3-layer relation-aware graph attention network (RGAT) with gated cross-attention fusion over dependency and co-reference relations, yielding sentence-level representations h^o and h^{cf}. Full architectural details are in Appendix B (Table 8).

3.3 Counterfactual Branch and Debias Fusion

The counterfactual branch is created by masking identity-bearing tokens from a manually curated lexicon covering eight categories: religion, gender, organization, race, profession, sexual orientation, asylum/refugee/foreigner (ARF), and miscellaneous identity expressions. During training, each matched identity token is replaced by the mask token with probability $p = 0.8$ in the main factorial experiments; masking is deterministic at evaluation time to remove sampling variance. We fix $p = 0.8$ for all 2×2 factorial experiments; the mask-probability sweep reported in Section 5.1 is a post-hoc diagnostic on intervention-strength sensitivity and is not used to select the factorial setting.

On the training split, 69.2% of samples contain at least one matched identity token, with coverage varying across TI classes (64.2% for *Profession* to 86.4% for *Gender*); full coverage statistics are in Appendix D.1.1.

The counterfactual input is encoded by the same BART encoder and graph-fusion stack, producing a counterfactual representation h^{cf}. The original and counterfactual sentence representations are then combined by a learned debias-fusion block:

$$h_{bias} = h^o - h^{cf}, \quad (1)$$

$$g = \sigma(\mathrm{MLP}([h^o; h^{cf}; h_{bias}])), \quad (2)$$

$$h_{debias} = h^o - g \odot h_{bias}, \quad (3)$$

followed by an additional fusion layer over $[h_{debias}; h^{cf}]$. The gate estimates which part of the counterfactual difference associated with identity masking should be subtracted from the original representation.

We use h_{bias} as a counterfactual difference vector rather than as a guaranteed pure identity-bias direction; masking can also perturb syntax, attention patterns, and contextual semantics.

The three routing regimes are:

Unified-CF. All classification heads receive h_{debias}:

$$\hat{y}^{cd}, \hat{y}^{sa}, \hat{y}^{ti} \leftarrow h_{debias}. \quad (4)$$

This treats debiasing as task-agnostic.

Task-aware routing. CD and SA, for which identity cues can act as shortcuts, receive the debiased representation, while TI keeps the original identity-preserving representation:

$$\hat{y}^{cd}, \hat{y}^{sa} \leftarrow h_{debias}, \quad \hat{y}^{ti} \leftarrow h^{o}. \tag{5}$$

For generation, the implementation uses token-level debiasing rather than the sentence-level TI routing decision. Let $H^{fuse,o}$ and $H^{fuse,cf}$ denote the RGAT-fused token-level states from the original and counterfactual branches respectively. The decoder input is computed as

$$H^{gen} = H^{fuse,o} - \mathrm{Adapter}(H^{fuse,o} - H^{fuse,cf}), \tag{6}$$

and feeds H^{gen} to the BART decoder. This means the task-aware design is not "No-CF" for generation; instead, it preserves the original sentence representation for the TI classifier while still using counterfactual token-level correction for explanation generation.

No-CF. All classification heads receive the original representation, and the counterfactual branch is disabled:

$$\hat{y}^{cd}, \hat{y}^{sa}, \hat{y}^{ti} \leftarrow h^{o}. \tag{7}$$

No-CF is not part of the factorial routing design, but provides a reference estimate of TI performance when identity evidence is fully preserved.

The model uses separate CD and SA heads and a hierarchical TI head; the task-aware and unified regimes differ only in whether the TI head receives h^{o} or h_{debias}. Loss definitions, coefficients, and curriculum schedule are in Appendix B.

4 Experimental Setup

Datasets. We evaluate on two main datasets and one external diagnostic corpus. The primary benchmark is the English-translated version of BullyExplain (Maity et al., 2023). The original dataset provides CD, SA, and TI labels in English; only the Hinglish text and token-level rationale spans are translated. Rationale tokens are constrained to appear in the translated text, so translation must preserve the exact surface forms of evidence spans. A manual audit of 104 examples confirms 100% rationale alignment and 96.2% identity-mention preservation (Appendix A). The benchmark contains 6,084 instances after filtering. The second dataset is HateXplain (Mathew et al., 2021), a hate speech corpus of 20,148 posts from Twitter and Gab annotated for hate classification and 11 target groups (e.g., *African*, *Islam*, *Refugee*); we use it to study how the routing advantage varies with intervention strength. As a third-dataset diagnostic, we construct a CD+TI task from Measuring Hate Speech (Sachdeva et al., 2022), which contains 135,556 comments with continuous hate speech and target-group scores. A comment is CD-positive when its mean `hatespeech` score is ≥ 0.5 (the midpoint of the annotation scale, separating predominantly non-hateful from predominantly hateful judgements). TI labels are derived by thresholding each of the 26 target-group indicators at 0.5 and retaining only comments with exactly one target group above threshold; this single-target filter ensures unambiguous TI supervision. We retain six target-group categories that each supply at least 500 single-target instances after the CD threshold is applied: *Women*, *Jewish*, *Black*, *Muslim*, *LGBTQ+*, and *Hispanic*. The remaining 135,556→39,565 reduction is mainly due to multi-target comments and comments near the 0.5 CD boundary.

Metrics. For classification tasks, we report macro-F1 and accuracy where available. For generation, we report ROUGE-1, ROUGE-2, ROUGE-L, and Jaccard overlap. We use macro-F1 as the main TI metric because it is sensitive to minority and long-tail target classes. Metric definitions follow Appendix H.

Compared systems. Our core comparisons are Unified-CF (all heads receive h_{debias}), task-aware routing, and No-CF. The five-seed factorial crosses routing with two TI weighting schemes: effective-number weighting (Cui et al., 2019) (inverse effective sample size, $\beta = 0.999$) and log-inverse weighting (inverse log-frequency, $\alpha = 1.0$). Single-run checkpoints cover additional operating points; *Global-Debias* is the unified checkpoint selected by validation average-F1, and *TaskAware-LiteTI* is the task-aware checkpoint balancing SA and generation quality.

Implementation details. The backbone is BART (hidden size 768) with a 3-layer RGAT graph branch. Training uses AdamW with cosine scheduling and early stopping on validation average macro-F1; full hyperparameters are in Appendix B.

Evaluation protocol. For the main 2×2 factorial experiments, the counterfactual mask probability is fixed to $p = 0.8$ for all cells and seeds;

this value was specified before the final factorial runs and is not tuned. Separately, we report a diagnostic mask-probability sweep over $p \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ on BullyExplain and $\{0.2, 0.4, 0.6, 0.8\}$ on HateXplain to characterize sensitivity to intervention strength; these sweeps are not used to select the factorial setting. Measuring Hate Speech is a five-seed text-only external diagnostic; RoBERTa-base is a three-seed architecture diagnostic. Single-run comparisons are marked as reference values throughout.

Statistical testing. For multi-seed comparisons (five-seed BullyExplain factorial, five-seed MHS diagnostic), we report 95% bootstrap confidence intervals over seed-level paired differences (10,000 resamples) and one-sided p-values for the hypothesis that the mean advantage exceeds zero. For small-n diagnostics ($n \leq 3$ seeds, RoBERTa-base), we use a one-sided seed-level permutation test. No multiple-comparison correction is applied across tables; significance claims are limited to pre-specified primary contrasts (routing main effect on TI).

5 Results on BullyExplain

5.1 Task-Aware Routing Yields a Targeted TI Gain

Table 1 is the primary evidence: a five-seed routing-by-weighting factorial ablation that isolates routing from weighting. Task-aware routing improves TI under both weighting schemes, with the matched effective-number comparison (Cell B vs. Cell C, both using effective-number weighting) yielding $+4.76$ points and the routing main effect $+3.30$ TI F1. CD and SA remain stable across cells. All other BullyExplain analyses are mechanism diagnostics.

Checkpoint selection is audited in Appendix D.1 (Table 19); the sole deviation favors the unified condition, making the routing contrast conservative.

5.2 Intervention Strength Produces a Pareto Trade-off

Table 2 sweeps mask probability on a single-run task-aware checkpoint; the factorial ablation remains the primary evidence. TI responds non-monotonically, peaking at $p = 0.7$; generation quality also peaks in the moderate-to-strong range. Although this diagnostic sweep peaks near $p = 0.7$, the factorial experiments were run with the pre-specified value $p = 0.8$; the factorial setting is not selected or adjusted based on the sweep.

The three tasks respond differently to intervention strength: SA prefers weak masking, CD is largely insensitive, and TI peaks at $p = 0.7$. There is therefore no single globally optimal mask probability; Pareto frontier plots across CD–TI and TI–generation are provided in Appendix D.1.

5.3 Per-Class Evidence Reveals Identity-Dependent Effects

The class-level analysis further clarifies where routing matters. In the five-seed per-class comparison on BullyExplain, task-aware routing outperforms No-CF on six of eight TI classes, with the largest gains on *Gender* (+10.20), *Religion* (+6.44), and *Organization* (+5.27). The main exception is *Race*, where No-CF remains stronger by 7.01 points; *Misc* also slightly favors No-CF (+1.33). Full per-class results are reported in Table 11 in the appendix.

Together with the corrected 2×2 routing results, this pattern indicates that the task-aware gain is concentrated in lexically explicit identity categories, rather than driven by a single dominant class. The remaining exceptions also sharpen the claim: task-aware routing is not uniformly dominant over preserving the original representation everywhere, but it provides a more balanced compromise than routing all heads through a globally debiased representation.

5.4 Comparison to GenEx

GenEx-Con (Maity et al., 2023), the strongest reported GenEx variant, is the closest published system on this benchmark and task configuration. Table 3, panel (a) shows that our variants improve over GenEx-Con on CD and SA, while task-aware also achieves the highest TI. Unified routing falls below both No-CF and GenEx-Con on TI (54.58 vs. 55.36 and 56.44), supporting a task-dependent trade-off rather than a blanket SOTA claim.

On GenEx-style generation metrics, including Jaccard similarity, HD-sim, and ROS (definitions in Appendix H), TaskAware-LiteTI obtains the strongest generation quality among our ablated variants, with No-CF and Global-Debias remaining competitive; full numbers are in Table 9 in the appendix. Table 3, panel (b) reports a linear probing result: a logistic-regression classifier is trained on frozen h^o and h_{debias} representations from five Cell-B checkpoints to predict TI labels, probing how much TI-label information remains available

Cell	TI routing	TI weighting	CD F1	SA F1	TI F1
A	Unified	Log-inverse	85.24 ± 1.04	75.20 ± 0.89	54.58 ± 3.09
B	Task-aware	Effective-number	85.06 ± 0.50	75.26 ± 0.67	$\mathbf{58.32} \pm \mathbf{1.88}$
C	Unified	Effective-number	85.05 ± 0.92	74.65 ± 1.12	53.55 ± 4.14
D	Task-aware	Log-inverse	85.42 ± 0.53	75.72 ± 0.68	56.42 ± 4.26

Table 1: Controlled 2×2 factorial ablation over five paired seeds (seeds 42–46). Task-aware routing improves TI under both weighting schemes (B$-$C: $+4.76$, 95 % CI [2.06, 8.16]; routing main effect: $+3.30$, 95 % CI [1.69, 5.28]), while CD and SA remain stable across cells. The weighting main effect is small and its confidence interval includes zero ($+0.44$, 95 % CI [-0.29, 1.17]).

Mask	CD F1	SA F1	TI F1	R-L	Jaccard (%)
0.0	86.35	**76.90**	37.89	73.19	70.50
0.1	**86.52**	76.72	40.59	70.13	69.66
0.3	84.70	73.92	54.66	79.62	75.28
0.5	**86.52**	75.72	55.82	79.39	74.23
0.7	86.19	75.14	**59.53**	81.08	**80.57**
0.9	85.82	74.16	56.93	**81.43**	79.89

Table 2: Single-run mask-strength sweep (task-aware routing). TI peaks non-monotonically at $p = 0.7$; generation quality also favors moderate-to-strong masking. mask=0.0 retains CF objectives and differs from No-CF (Appendix D.1).

(a) GenEx-Con comparison (ours: 5-seed means)			
Model	CD F1	SA F1	TI F1
GenEx-Con	82.47	73.16	56.44
No-CF	**85.86**	**76.13**	55.36
Unified	85.24	75.20	54.58
Task-aware	85.06	75.26	**58.32**
(b) Linear TI-label probing (5 Cell-B checkpoints)			
Metric	h^o	h_{debias}	Δ
TI macro-F1	**53.57±6.96**	40.83±5.90	**+12.74**

Table 3: BullyExplain analysis. (a) GenEx-Con comparison, with our variants reported as five-seed means. Task-aware routing improves TI over both unified routing and No-CF, while No-CF remains strongest on CD and SA. This supports a task-dependent trade-off: unified debiasing suppresses TI-relevant evidence, whereas task-aware routing recovers TI performance without collapsing CD/SA. (b) Linear probing shows that TI labels are more decodable from h^o than from h_{debias} ($\Delta = +12.74$, 95 % CI [3.75, 20.90], $p = 0.0035$), indicating that the debiased representation removes TI-relevant identity information.

in each representation. The mechanism interpretation is in Section 8.1.

6 HateXplain Mechanism Analysis

6.1 Task-Aware Gains Depend on Intervention Strength

We include a *vanilla* baseline (standard fine-tuning without any counterfactual branch) to bound TI performance without CF intervention. At mask=0.8, task-aware routing improves TI over unified by +1.36 F1 (95% CI [0.82, 2.08]) and is indistinguishable from vanilla on TI ($\Delta = -0.15$, 95% CI [-0.93, 0.52]); vanilla leads slightly overall, so the comparison is conditional on using counterfactual debiasing (Table 4). Our claim is therefore not that task-aware routing dominates vanilla training, but that it prevents uniform counterfactual routing from suppressing TI evidence when counterfactual debiasing is used. The TI advantage grows with masking: absent at mask=0.2 ($\Delta = -0.49$), positive at 0.4 ($+0.38$) and 0.6 ($+0.74$), clearest at 0.8. Full per-seed numbers and bootstrap significance results are in Appendix D.1 (Table 15).

Figure 2 plots the task-aware minus unified TI delta across masking strengths; the routing advantage appears only under stronger intervention.

6.2 Long-Tail Classes Benefit Most

The TI gain concentrates on long-tail classes (*Caucasian*, *Refugee*, *Asian*), while head classes (*African*, *Islam*, *Jewish*) change little. Figure 3 shows this directly. Full per-class tables at mask strengths 0.8 and 0.6 are reported in Tables 27 and 28 in the appendix.

The pattern fits the shortcut-versus-evidence account. Frequent classes have enough diverse training signal to absorb the loss of identity cues; long-tail classes often have little else. Removing those cues erases the class-defining signal, not a spurious shortcut. The *Refugee* category is the clearest case: unified debiasing causes the largest drop, and task-aware routing recovers much of it, though vanilla remains best.

Config	Hate F1	TI F1
Vanilla	**67.16 ± 0.11**	**75.93 ± 0.83**
Unified	67.02 ± 0.70	74.42 ± 0.61
Task-aware	66.76 ± 0.70	75.79 ± 0.10

Table 4: HateXplain mask=0.8 comparison over three seeds (mean ± standard deviation). Task-aware routing recovers TI relative to unified debiasing and is statistically indistinguishable from vanilla on TI at this mask strength.

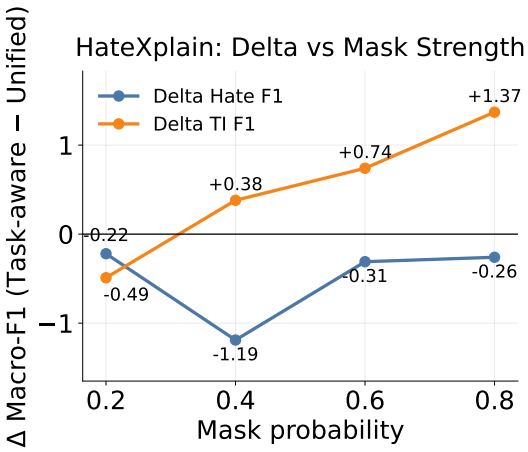

Figure 2: HateXplain TI gain (task-aware − unified) versus mask probability. The routing advantage grows under stronger intervention, with a small hate-classification cost throughout.

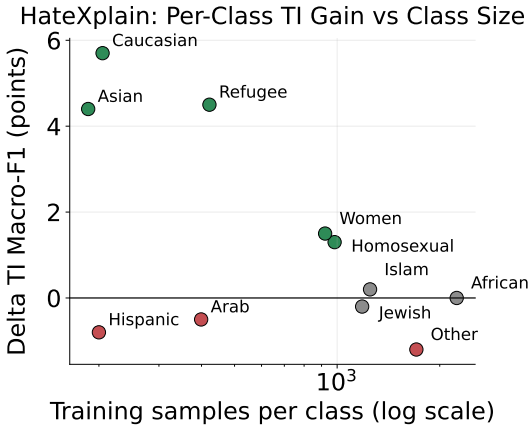

Figure 3: Per-class TI gain versus class size on HateXplain. The advantage of task-aware routing over unified debiasing concentrates on long-tail identity-dependent classes rather than head classes.

Setting	CD F1	SA F1	TI F1
(a) RoBERTa-base on BullyExplain (3 seeds)			
No-CF	86.81 ± 0.73	73.47 ± 1.76	59.57 ± 0.45
Unified	86.50 ± 0.25	74.60 ± 1.34	58.77 ± 1.12
Task-aware	86.73 ± 0.27	74.46 ± 1.26	**60.32 ± 1.80**
(b) Measuring Hate Speech (5 seeds, CD+TI only)			
No-CF	74.18 ± 0.60	–	84.47 ± 0.92
Unified	74.27 ± 1.22	–	84.13 ± 1.19
Task-aware	74.75 ± 0.46	–	**85.08 ± 0.80**

Table 5: Robustness diagnostics. (a) RoBERTa-base on BullyExplain: task-aware routing improves TI by +1.55 over unified ($p = 0.0214$, 95 % CI [0.22, 2.84]), without the graph branch or phased curriculum. (b) Measuring Hate Speech: task-aware routing improves TI by +0.95 (95 % CI [0.37, 1.54]) on an independent corpus. CD changes are small throughout.

7 External and Architecture Diagnostics

7.1 Measuring Hate Speech: Cross-Corpus Replication

On Measuring Hate Speech, task-aware routing improves TI over unified by 0.95 macro-F1 (85.08 vs. 84.13; 95% CI [0.37, 1.54]), positive across all five seeds. Unified routing falls slightly below No-CF on TI ($\Delta = -0.34$; 95% CI [-1.61, 1.03], not statistically significant). The pattern replicates on an independent corpus with a different six-class target schema and a text-only model, suggesting that it is not merely an artifact of the BullyExplain translation or its particular target taxonomy. Per-seed results are in Table 22 in the appendix.

7.2 RoBERTa-Base: Architecture Diagnostic

To check whether the routing advantage depends on the graph branch, generation path, or phased curriculum, we repeat the three-way routing comparison using a simplified text-only RoBERTa-base backbone. Task-aware routing still improves TI by +1.55 over unified (60.32 vs. 58.77; 95% CI [0.22, 2.84]; $p = 0.0214$; Table 5, panel (a)), with the advantage positive across all three seeds. A further-reduced model that removes the graph branch, generation path, and curriculum (the *minimal diagnostic*, Appendix E.2) is more checkpoint-sensitive: the advantage disappears under TI-specific checkpoint selection, so the five-seed factorial remains the primary evidence. Together, the RoBERTa diagnostic and the minimal diagnostic suggest that the routing effect is not specific to the full graph-enhanced architecture but is sensitive to checkpoint selection when the sample size is small.

8 Discussion

8.1 A Task-Conditional View of Identity Cues

Across all three datasets, the effect of CF intervention is task-dependent: unified debiasing removes target-relevant identity evidence, while task-aware routing reduces this loss while CD or SA remain largely stable. Multi-task debiasing should be evaluated per task, not applied as a uniform policy.

The only task-conditional component is which representation each head receives; all other components are shared. Across matched weighting cells, the routing main effect is +3.30 TI F1 (95% CI [1.69, 5.28]), while the weighting effect (+0.44) crosses zero.

Linear probes trained on five Cell-B checkpoints show TI labels are substantially more decodable from h^o than h_{debias} (53.57 vs. 40.83; $\Delta = +12.74$, 95 % CI [3.75, 20.90], $p = 0.0035$), with the largest gaps in *Gender* (+34.68) and *SexualOrientation* (+26.41; Table 3, panel (b)). The mask-strength, long-tail, and per-class results converge: routing matters most when masking is strong enough to erase TI evidence.

Per-class routing gain on HateXplain correlates with lexicon coverage ($\rho = 0.65$, $p = 0.029$); the association is weaker on BullyExplain (Tables 12, 13, and 14 in the appendix). A random-identity substitution diagnostic (Table 18 in the appendix) shows the directional advantage is preserved under an alternative CF perturbation, though with higher variance across seeds.

The RoBERTa-base and minimal diagnostics (Section 7; Appendix E.2) support the view that the routing effect is not specific to the graph-enhanced architecture. The minimal diagnostic is checkpoint-sensitive, while the RoBERTa-base result is limited mainly by the small number of seeds.

In the available single-run generation checkpoints, preserving identity semantics does not degrade generation quality: TaskAware-LiteTI achieves the highest Jaccard (0.679) and lowest corrupted-output rate (9.2%) among single-run checkpoints vs. Global-Debias (0.630, 14.2%; Appendix C). Classification and generation routing operate at different granularities and are handled separately.

9 Conclusion

We studied counterfactual debiasing in multi-task harmful language understanding. Across BullyExplain, HateXplain, and Measuring Hate Speech, the results consistently support a task-conditional view: the same identity cue can be a shortcut for detection while remaining label-defining evidence for target identification. Uniform debiasing ignores this distinction and suppresses information that fine-grained TI depends on.

The practical implication is that multi-task systems should not apply a single debiasing policy across all heads. Task-aware routing recovers TI without harming detection or sentiment performance by routing CD and SA through the debiased representation while preserving the original representation for TI, because identity cues provide label-defining evidence for that task. The gain is attributable to routing rather than class-weighting choices. The effect is most reliable under moderate-to-strong masking and concentrates on long-tail, identity-defined target classes, where the suppressed cues carry the most label-defining signal.

Limitations

Our study has several limitations. First, the main classification comparison is a five-seed aggregate, but some generation and qualitative checkpoints are single-run, so not all tables carry the same statistical weight; we mark these explicitly. Second, the linear probe strengthens the mechanism claim but remains correlational rather than causal: it shows that h_{debias} contains less TI-decodable information, but does not establish that this loss is solely attributable to counterfactual masking rather than to other training differences between the two configurations. Third, the identity lexicon captures only explicit mentions and misses implicit references; the intervention is therefore a controlled identity-sensitivity perturbation rather than a complete counterfactual generator. Fourth, BullyExplain is translated from Hinglish using neural MT; the manual audit (Appendix A) quantifies alignment quality, but residual errors—particularly in implicit-abuse cases where abusive intent is expressed indirectly—remain possible. Fifth, the per-class analysis characterizes task-conditional routing effects rather than constituting a full fairness audit; we do not report subgroup AUC or equalized odds gaps, and make no claim about group fairness improvement in a formal sense.

Ethical Considerations

This work studies harmful language, identity references, and cyberbullying targets. Any such system risks misuse if deployed for surveillance or moderation without oversight. Our findings show that debiasing interventions can have unintended costs on identity-dependent tasks; these costs should be evaluated explicitly per task before deployment, with task-specific auditing and human review in high-stakes applications.

Acknowledgments

Acknowledgments omitted for anonymous review.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. UNIREX: A unified learning framework for language model rationale extraction. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2867–2889.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

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, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

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*, 2:665–673.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1926–1940.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9804–9817.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.

Krishanu Maity, Raghav Jain, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2023. GenEx: A commonsense-aware unified generative framework for explainable cyberbullying detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16632–16645.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1878–1898.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5356–5371.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP*.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3524–3542.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15109–15126.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6707–6723.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi S. Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 12822–12835.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

A Dataset Preprocessing

The original BullyExplain resource (Maity et al., 2023) provides CD, SA, and TI labels in English. Only the Hinglish (Hindi–English code-mixed) comment text and the token-level rationale spans require translation. This distinction matters for the audit: because labels are already in the target language, we do not need to verify label preservation; the audit focuses entirely on whether the translated text retains task-relevant evidence and whether rationale alignment holds.

Translation and alignment procedure. Comments are translated to standard English using neural MT. Rationale annotations are token-level spans—each rationale word must appear verbatim in the comment text—so the alignment rule must hold in the translated text as well. The translation step is therefore constrained: after translating both the comment and the rationale independently, we verify that each translated rationale token appears in the translated comment, and adjust the translation if it does not. This produces a dataset where, by construction, every rationale span is anchored to the translated text.

Manual quality audit. We conduct a manual audit over 104 translated examples, balanced across all eight TI classes with 13 examples per class. For each example, two auditors independently assessed six items: (1) whether the translated text contains identity mentions, (2) target-group references, (3) abusive content, (4) sentiment-compatible evidence, (5) whether each translated rationale span appears in the translated text, and (6) whether each rationale span remains a valid evidence span in the translated text. Initial agreement across all items was 91.3%, with a Cohen's κ of 0.81 (substantial

Audit item	Checked	Pass	Pass rate
Identity mention present	104	100	96.2%
Target group present	104	100	96.2%
Abusive content present	104	99	95.2%
Sentiment polarity compatible	104	102	98.1%
Rationale tokens in text	93	93	100.0%
Rationale evidence valid	93	91	97.8%

Table 6: Manual data-quality audit over 104 translated examples, balanced across all eight TI classes with 13 examples per class. Since task labels are already in English, the audit focuses on translated text quality and token-level rationale alignment rather than relabeling. Rationale rates are computed over examples with non-empty rationales.

agreement). Disagreements were resolved by discussion. Table 6 reports the overall results.

The two most informative items are the rationale checks. All 93 non-empty rationale spans appear in the translated text (100%), confirming that the alignment constraint is satisfied. Among these, 91 out of 93 (97.8%) remain valid evidence spans. Identity mentions and target-group references are preserved in 96.2% of examples each. Abusive content is the item with the most losses (95.2%), consistent with implicit-abuse cases where the abusive intent is expressed indirectly and the translated phrasing may be less explicit.

After removing excessively short and long samples, the final benchmark contains 6,084 comments split into training, validation, and test sets with an 8:1:1 ratio. The translated split and alignment files will be released with the code upon publication, subject to the original dataset license.

B Training Details

This section details the training curriculum referenced in the main paper. The curriculum is part of the shared platform rather than a contribution of this paper.

Hierarchical contrastive objective. The platform includes a hierarchical contrastive objective (HACT) for explanation-side alignment. The anchor is a projection of the model representation; the positive example is the gold explanation encoded by the BART encoder; negatives are introduced in a curriculum of three types: (i) token-masked explanations, (ii) semantically drifted explanations, and (iii) batch-hard negatives. The curriculum activates these negative types progressively over training epochs, with stronger hard-negative weighting later in training.

Phase-wise schedule. Training uses a four-stage curriculum with 6 phase-1 epochs, 4 phase-2a epochs, 6 phase-2b epochs, and 6 phase-3 epochs. Counterfactual training is inactive in phase 1 and active afterward. Fast gradient method (FGM) adversarial training is enabled in phases 2b and 3. The bias loss is warmed up over 4 epochs to a maximum weight of 0.1. The best checkpoint is selected by mean validation macro-F1 over the three classification tasks.

Reproducibility. Tables 7 and 8 list all hyperparameters needed to reproduce the main experiments. Graphs are precomputed in the released JSONL input; all experiments consume the same graph file.

C Generation Qualitative Analysis

Table 9 reports the qualitative generation comparison used to complement the main-text generation analysis.

TaskAware-LiteTI is strongest overall, followed by No-CF and then Global-Debias. Correct non-null evidence extraction remains similarly low across the three variants. The real qualitative gap therefore comes less from exact span recovery than from controlling degeneration, over-generation, and malformed or repetitive outputs. This is consistent with the main-text interpretation that the benefit of task-aware routing for generation is largely a stability benefit.

D Diagnostic Analyses and Implementation Notes

Single-run versus repeated-run reporting. The paper includes repeated-run aggregates and single-run reference checkpoints. We use the repeated-run controlled routing comparison as the primary evidence for the TI claim, while reference results are retained only where they help contextualize regime-level trade-offs, such as the No-CF reference and the BullyExplain generation scores.

D.1 Additional Diagnostic Analyses

Table 10 checks whether the TI gains hide class-specific regressions on cyberbullying detection (CD) or sentiment analysis (SA). The main A→B

Item	Value
Base model	BART-base (local checkpoint)
Training data	Translated BullyExplain graph JSONL
Optimizer	AdamW
BART encoder LR	3×10^{-5}
BART decoder / LM-head LR	6×10^{-5}
RGAT and task-module LR	3×10^{-4}
Weight decay	0.03
Batch size	4
Gradient accumulation	4
Effective batch size	16
Max sequence length	256
Maximum epochs	22
Warmup	4% of total steps
Early stopping patience	5 epochs
Checkpoint selection	Best validation average F1
Focal-loss γ	2.0
Label smoothing	0.1
TI loss weight λ_{TI}	1.5
Generation loss weight λ_{gen}	4.0
Classification loss weight λ_{cls}	1.0
Counterfactual mask probability	0.8
Effective-number β	0.999
Log-inverse α	1.0
Seeds	42, 43, 44, 45, 46
Split seed	42
GPU	3× NVIDIA RTX 4090 24GB
Training budget	∼1.5 GPU-hours per seed; ∼150 GPU-hours total across all reported experiments
Framework	PyTorch 2.7.0 (CUDA 11.8), transformers 4.51.3, rouge-score 0.1.2

Table 7: Training and checkpointing details for the main corrected 2×2 experiments, with compute totals aggregated across all reported experiments. Checkpoints are selected by validation average F1 across tasks.

Item	Value
Text hidden size	768
Graph encoder	Relation-aware GAT
RGAT layers	3
RGAT hidden dimension	768
RGAT output dimension	768
RGAT attention heads	8
Relation embedding dimension	64
Dropout	0.2
TI-head dropout	0.4
Projection dimension	128
Fusion module	Cross-attention over graph nodes
Fusion attention heads	8
CD / SA heads	MLP classification heads
TI head	Hierarchical coarse-to-fine head
TI fine classes	8
TI coarse classes	4
CF intervention	Identity-token masking
CF evaluation masking	Deterministic
Lexicon classes	Religion, Gender, Race, Organization,Profession, SexualOrientation, ARF, Misc
Lexicon matching	Token-id matching after tokenizer encoding
TI class counts	[244, 265, 276, 443, 506, 232, 932, 316]

Table 8: Architecture, graph encoding, and counterfactual intervention details. The identity lexicon is organized by TI class and used for token-level counterfactual masking.

comparison shows no substantial auxiliary-task degradation; the largest auxiliary shift is on neutral sentiment.

Table 11 compares the selected task-aware model with the No-CF five-seed reference at the TI class level. No-CF serves as a reference condition in which all task heads receive the undebiased representation h^o. It is not part of the factorial routing design, but provides a reference estimate of TI performance when identity evidence is fully preserved.

Tables 12 and 13 report the actual identity-token intervention coverage induced by the lexicon and tokenizer on BullyExplain. Table 14 reports the corresponding HateXplain target-class diagnostic. Coverage is substantive but uneven, so the mechanism interpretation is strongest for lexically expressed identity evidence.

Tables 15, 16, and 17 provide the HateXplain mask=0.8 paired bootstrap, the full test-sweep results, and the validation mask sweep. Task-aware routing improves TI over unified routing at mask probability 0.8, remains statistically indistinguishable from vanilla on TI, and both sweeps show positive but non-monotonic gains across mask strengths.

Table 18 reports a random-identity substitution robustness check on BullyExplain. The result is directionally consistent with the main routing effect, but the TI interval crosses zero, so we treat it as a noisy diagnostic rather than a confirmatory test.

Tables 19 and 20 audit checkpoint selection for the five-seed 2×2 factorial study. The reported checkpoint matches the predefined average-F1 optimum in 19/20 runs and the TI-F1 optimum in 15/20 runs. The sole deviation from the average-F1 rule occurs in a unified-routing run (seed43,

Model	Avg Jaccard (%)	Correct-Null	Correct-NonNull	Under	Over	Corrupted
TaskAware-LiteTI	**67.9**	**50.6**	5.6	20.3	**14.2**	**9.2**
No-CF	66.1	49.9	**6.3**	**18.9**	15.0	9.9
Global-Debias	63.0	38.6	5.6	21.2	20.3	14.2

Table 9: Qualitative generation comparison on BullyExplain. TaskAware-LiteTI and No-CF are qualitatively more stable than Global-Debias, especially in corrupted or repetitive outputs.

Task	Class	Unified+LogInv F1	Task-aware+Eff F1	Unified+Eff F1	Task-aware+LogInv F1
CD	Bully	87.66 ± 0.71	86.74 ± 0.84	87.17 ± 0.51	86.89 ± 0.75
CD	Non-bully	84.04 ± 0.64	83.34 ± 0.48	83.37 ± 0.09	83.35 ± 0.71
SA	Negative	88.45 ± 0.57	87.48 ± 0.22	88.18 ± 0.19	87.28 ± 0.86
SA	Neutral	57.43 ± 1.14	55.39 ± 1.64	54.57 ± 1.52	55.44 ± 2.93
SA	Positive	81.05 ± 0.46	82.19 ± 0.70	81.45 ± 1.83	82.45 ± 0.90

Table 10: Per-class cyberbullying detection (CD) and sentiment analysis (SA) F1 for the controlled routing–weighting ablation. The task-aware TI gain does not hide substantial auxiliary-task degradation; the largest shift in the main A→B comparison is neutral sentiment.

TI class	No-CF TI F1	Task-aware TI F1
Religion	66.17 ± 6.40	72.61 ± 2.80
Gender	63.99 ± 8.18	74.19 ± 1.78
Organization	62.81 ± 0.89	68.08 ± 1.36
Race	52.22 ± 5.86	45.21 ± 5.57
Profession	42.10 ± 5.02	44.58 ± 2.61
SexualOrientation	59.11 ± 9.02	62.94 ± 6.52
ARF	63.88 ± 1.98	65.27 ± 1.43
Misc	38.28 ± 5.29	36.95 ± 2.89

Table 11: No-CF five-seed TI per-class F1 (macro-F1, percentage points) compared with Cell-B (five-seed task-aware routing, effective-number weighting). No-CF is a reference condition rather than a factorial routing cell. Note: the single-run No-CF reference shown in Figure 1 may differ from these five-seed means on individual classes.

effective-number weighting), where the reported checkpoint coincides with the TI-optimal epoch rather than the average-F1-optimal epoch. Because this exception raises the unified condition rather than the task-aware condition, the reported routing contrast is conservative with respect to this deviation. Across the six runs in which the TI-optimal and average-optimal epochs differ, four belong to unified routing and two to task-aware routing, indicating that the reported selection rule does not systematically privilege the task-aware cells.

Table 21 reports the per-class TI-probe gaps underlying the main-text representation analysis. The largest mean gaps occur in *Gender* and *SexualOrientation*, while the effect remains smaller for classes such as *Race* and *Religion*.

Group	Coverage	Avg. matched tokens
Train	69.2%	1.55
Validation	69.3%	1.62
Test	65.3%	1.41
Bully	70.7%	1.54
Non-bully	67.1%	1.55

Table 12: Actual identity-token intervention coverage induced by the lexicon and tokenizer. Coverage denotes the percentage of samples with at least one matched identity token.

D.1.1 Lexicon Coverage

Tables 12 and 13 report the token-level intervention coverage induced by the identity lexicon and tokenizer. Coverage is substantial across data splits and CD labels, while the TI-class table shows that the intervention reaches all TI categories but remains uneven. The corresponding association is weaker on the primary benchmark, where per-class routing gains are only weakly correlated with lexicon coverage. Lexicon overlap is therefore best interpreted as a mechanism indicator rather than a complete predictor of per-class gains. Table 14 shows the HateXplain diagnostic and adds the per-class routing gain at mask probability 0.8. Across the 11 HateXplain target classes, routing gain is positively correlated with both global lexicon coverage (Spearman $\rho = 0.65$, $p = 0.029$) and class-specific coverage ($\rho = 0.64$, $p = 0.035$). This supports the interpretation that masking most directly affects explicitly lexicalized identity categories, while remaining a small-sample diagnostic rather than a causal identification test.

TI class	Coverage	Avg. matched tokens
ARF	65.6%	1.27
Gender	86.4%	1.94
Misc	74.4%	1.89
Organization	69.6%	1.47
Profession	64.2%	1.36
Race	74.5%	1.70
Religion	77.0%	2.10
SexualOrientation	70.7%	1.34

Table 13: Training-split intervention coverage by TI class. The dictionary-triggered intervention reaches all TI categories but remains uneven, motivating our limitation that implicit or non-lexical identity references are not fully covered.

HateXplain target	n	Global hit	Class hit	ΔTI
African	3470	39.9	14.4	+0.07
Arab	646	35.4	10.1	-0.48
Asian	423	40.2	16.1	+4.37
Caucasian	718	80.9	75.3	+5.77
Hispanic	284	34.9	7.7	-0.80
Homosexual	1943	50.6	37.6	+1.30
Islam	2118	58.9	40.8	+0.21
Jewish	1720	55.2	46.2	-0.22
Other	1776	22.2	0.0	-1.17
Refugee	980	40.5	21.2	+4.46
Women	1558	42.7	34.1	+1.51

Table 14: Identity-lexicon coverage on HateXplain target-labeled instances, with ΔTI denoting task-aware minus unified per-class TI F1 at mask probability 0.8 in percentage points. Samples without a target label or without a mapping to the 11 target classes are excluded from this diagnostic (4512 skipped rows). Routing gain is positively correlated with global lexicon coverage (Spearman $\rho = 0.65$, $p = 0.029$) and class-specific coverage ($\rho = 0.64$, $p = 0.035$), consistent with the mechanism interpretation for lexically expressed identity evidence.

Random-identity substitution. Under random-identity substitution, task-aware routing remains directionally above unified routing on TI, with CD and SA essentially unchanged. However, the seed-to-seed variance is large: the paired TI differences are -1.05, +1.95, and +4.86. We therefore interpret the check cautiously. It suggests that the routing effect is not specific to the `[MASK]` token, but random identity tokens provide a weaker and less stable counterfactual signal than masking.

Interpretation of mask strength 0.0. The mask-strength sweep includes a $p = 0.0$ setting, but this should not be interpreted as identical to the No-CF control. In the sweep, the counterfactual branch and its associated objectives remain part of the training setup even when no identity tokens are masked; by contrast, the No-CF control disables counterfactual intervention altogether. This is why mask strength 0.0 does not numerically collapse to the No-CF line.

Metric	Contrast	Δ	95% bootstrap interval
TI F1	Task-aware–Unified	+1.36	[0.82, 2.08]
Hate F1	Task-aware–Unified	-0.27	[-0.67, 0.29]
TI F1	Task-aware–Vanilla	-0.15	[-0.93, 0.52]

Table 15: HateXplain mask=0.8 paired bootstrap over three seeds. The first two rows compare task-aware and unified routing within the counterfactual-debiasing setup. The final row is a cross-configuration reference against vanilla, showing that task-aware routing is not statistically distinguishable from vanilla on TI at this mask strength.

Mask	Hate F1		TI F1	
	Unified	Task-aware	Unified	Task-aware
0.2	67.38 ± 0.36	67.16 ± 0.46	75.53 ± 0.29	75.04 ± 0.21
0.4	67.31 ± 0.93	66.12 ± 0.54	75.69 ± 0.45	76.07 ± 0.57
0.6	67.13 ± 0.44	66.82 ± 0.52	75.13 ± 0.32	75.87 ± 0.39
0.8	67.02 ± 0.70	66.76 ± 0.70	74.42 ± 0.61	75.79 ± 0.10

Table 16: HateXplain test mask sweep (3-seed mean ± std). ΔTI = Task-aware − Unified grows from -0.49 at mask=0.2 to $+1.36$ at mask=0.8. We do not interpret this as a monotonic dose-response; see the validation sweep below.

Formal routing view. The main paper describes the routing decision in words. Here we write it explicitly to clarify what is and is not being compared. Let h^o denote the original sentence representation and h^{cf} the representation obtained from a counterfactual input in which identity-bearing tokens are masked with probability p. The debias-fusion mechanism can be written abstractly as

$$d = h^o - h^{cf}, \tag{8}$$

$$g = \sigma(\phi([h^o; h^{cf}; d])), \tag{9}$$

$$h_{debias} = \psi(h^o - g \odot d, h^{cf}), \tag{10}$$

where ϕ is the gate network and ψ denotes the final fusion layer. The unified and task-aware regimes

Mask prob.	Unified TI F1	Task-aware TI F1	ΔTI F1
0.0	69.54	74.45	+4.91
0.2	74.76	75.81	+1.04
0.4	74.53	75.35	+0.82
0.6	74.20	76.67	+2.47
0.8	73.85	75.24	+1.39
1.0	74.64	75.61	+0.97

Table 17: HateXplain validation mask sweep. Task-aware routing improves TI at every mask strength, but the gain is mask-dependent rather than monotonic.

Routing	n	CD F1		SA F1		TI F1	
Unified	3	85.63	±	75.11	±	54.59	±
		0.73		0.81		1.16	
Task-aware	3	85.65	±	74.96	±	56.51	±
		0.60		1.77		3.92	

Table 18: Random-identity substitution robustness check on BullyExplain. Instead of replacing identity-bearing tokens with [MASK], the counterfactual branch uses random identity substitutions. Task-aware routing preserves a directional TI advantage over unified routing (ΔTI = +1.92), while CD and SA remain comparable. The observed paired seed-level range for ΔTI is [-1.05, 4.86], making this a noisy robustness diagnostic rather than a confirmatory test. For reference, the matched [MASK]-based effective-number comparison in the main corrected 2×2 table gives ΔTI = +4.76, indicating a stronger routing effect under the main counterfactual construction.

differ in which representation is supplied to the classification heads:

$$\mathcal{L}_{unified} = E\Big[\ell_{cd}(f_{cd}(h_{debias}), y^{cd}) + \ell_{sa}(f_{sa}(h_{debias}), y^{sa}) + \lambda_{ti}\ell_{ti}(f_{ti}(h_{debias}), y^{ti})\Big], \quad (11)$$

$$\mathcal{L}_{task} = E\Big[\ell_{cd}(f_{cd}(h_{debias}), y^{cd}) + \ell_{sa}(f_{sa}(h_{debias}), y^{sa}) + \lambda_{ti}\ell_{ti}(f_{ti}(h^o), y^{ti})\Big], \quad (12)$$

with generation handled separately by token-level counterfactual correction in the full model. The main paper now supplements this original configuration contrast with a controlled routing-by-weighting ablation.

First-order intuition for TI degradation. The shortcut-versus-evidence claim can be understood through a local approximation. Holding the TI head fixed, the change in TI loss caused by replacing h^o with h_{debias} is approximately

$$\Delta\ell_{ti} \approx \nabla_h \ell_{ti}(f_{ti}(h^o), y^{ti})^\top \cdot (h_{debias} - h^o). \quad (13)$$

When the counterfactual difference $d = h^o - h^{cf}$ mostly captures spurious identity correlations for a detection head, subtracting a gated component of d can reduce shortcut reliance. For TI, however, the same direction can contain label-defining evidence. If the debiasing displacement $h_{debias} - h^o$ removes information aligned with the TI decision boundary, the first-order term can increase TI loss. Task-aware routing avoids this particular displacement for the TI classifier by feeding it h^o, while still allowing CD and SA to receive h_{debias}. This is an explanatory approximation, not an identification proof; it is included to make the mechanism behind the empirical mask-strength and per-class results more explicit.

Mask-strength regimes. The masking probability p controls how different the counterfactual view is from the original view. At very low p, h^{cf} remains close to h^o and the counterfactual branch may add training objectives without providing a meaningfully different view. At moderate-to-strong p, identity-bearing evidence is altered enough that the routing distinction becomes behaviorally visible. At very high p, further masking can improve

Audit item	Count	Proportion
Auditable runs	20/20	100%
Reported checkpoint = best average-F1 checkpoint	19/20	95%
Reported checkpoint = best TI-F1 checkpoint	15/20	75%
TI-optimal epoch ≠ average-optimal epoch	6/20	30%

Table 19: Checkpoint-selection audit for the 2×2 factorial study over five seeds.

Run	Routing	Best avg epoch	Best TI epoch
seed43 effective-number	Unified	14	16
seed44 log-inverse	Task-aware	16	17
seed44 log-inverse	Unified	17	22
seed45 effective-number	Task-aware	18	16
seed45 effective-number	Unified	21	22
seed45 log-inverse	Unified	18	19

Table 20: Runs in which the TI-optimal checkpoint differs from the average-F1-optimal checkpoint.

TI class	Mean ΔF1 ($h^o - h_{\text{debias}}$)
Religion	+5.94
Gender	+34.68
Organization	+17.36
Race	+2.42
Profession	+6.17
SexualOrientation	+26.41
ARF	+5.02
Misc	+3.89

Table 21: Per-class TI-probe gaps averaged across five Cell-B checkpoints (macro-F1 percentage points; $h^o - h_{\text{debias}}$, higher = more TI information lost by debiasing).

Seed	Setting	Best epoch	CD F1	TI F1
42	No-CF	4	73.71	83.78
42	Unified	5	74.28	83.90
42	Task-aware	3	74.23	84.60
43	No-CF	5	74.44	84.22
43	Unified	5	73.11	84.04
43	Task-aware	4	74.83	84.56
44	No-CF	7	73.47	83.63
44	Unified	3	75.66	82.91
44	Task-aware	3	74.93	84.37
45	No-CF	4	74.13	85.95
45	Unified	4	73.24	86.10
45	Task-aware	3	75.05	86.16
46	No-CF	5	75.17	84.76
46	Unified	3	75.06	83.70
46	Task-aware	4	74.71	85.73

Table 22: Per-seed Measuring Hate Speech diagnostic results. The diagnostic uses a shorter training schedule than the primary benchmark due to the larger corpus size; checkpoint selection uses validation average F1.

lexical overlap for generation while no longer improving TI, consistent with a trade-off between stronger intervention and preservation of target evidence.

Generation-side comparisons. Generation metrics on BullyExplain are reported as currently verifiable checkpoint outputs. We retain these numbers because they help characterize the TI–generation trade-off, and we report the metric family and implementation choices explicitly when making cross-paper comparisons.

E Architecture Robustness Note

A potential concern with the main results is that the task-aware versus unified comparison could in principle reflect interactions between counterfactual routing and the full platform components, such as the graph branch, explanation-generation path, HACT, and phased curriculum, rather than the routing decision itself. We therefore include reduced-architecture diagnostics. These checks serve as robustness diagnostics; the full-system routing-by-weighting ablation remains the primary evidence.

Setting	n	CD F1		SA F1		TI F1	
No-CF	3	85.92	±	71.90	±	56.77	±
		1.29		2.58		2.08	
Unified	3	85.23	±	71.45	±	57.50	±
		0.42		1.04		0.82	
Task-aware	3	86.10	±	71.69	±	58.94	±
		0.48		0.50		2.11	

Table 23: No-graph routing diagnostic over three seeds. The graph branch is disabled while the routing comparison is retained. The diagnostic is directionally consistent with task-aware routing but noisy, so it serves as a reduced-architecture robustness check rather than a standalone significant result.

Checkpoint rule	Setting	CD F1		SA F1		TI F1	
Avg-F1	No-CF	85.19	±	71.49	±	59.75	±
		0.78		1.03		1.49	
Avg-F1	Unified	84.70	±	72.05	±	58.59	±
		0.50		0.88		0.90	
Avg-F1	Task-aware	85.28	±	71.64	±	59.89	±
		0.94		2.97		0.80	
TI-F1	No-CF	85.23	±	71.01	±	60.28	±
		0.64		0.68		1.39	
TI-F1	Unified	85.03	±	71.86	±	59.53	±
		1.08		2.26		0.95	
TI-F1	Task-aware	85.11	±	71.87	±	59.29	±
		0.66		2.40		0.84	

Table 24: Minimal routing diagnostic under two checkpoint-selection rules. Under the task-balanced Avg-F1 rule used in the main experiments, task-aware routing improves TI over unified debiasing. Under TI-specific checkpoint selection, the advantage disappears, indicating that the reduced architecture is checkpoint-sensitive and should be treated as a diagnostic rather than primary evidence. Candidate checkpoints are restricted to CF-active epochs ($\text{epoch} \geq 7$).

E.1 RoBERTa Text-Only Routing Diagnostic

Table 5, panel (a) (main paper) reports the full RoBERTa-base routing diagnostic. Task-aware routing improves TI over unified routing by +1.55 macro-F1 points, with positive deltas across all three seeds. CD and SA changes are small and not statistically reliable. This diagnostic suggests that the routing effect is not specific to the full graph-enhanced architecture, although the small number of seeds makes it a robustness check rather than the primary evidence.

E.2 Minimal Routing Diagnostic

We additionally evaluate a reduced diagnostic model that removes graph fusion, generation evaluation, and HACT while retaining the BART encoder, counterfactual branch, debias fusion, and task heads. To avoid selecting checkpoints before the counterfactual branch is active, candidate checkpoints are restricted to epochs ≥ 7.

Under the same task-balanced Avg-F1 selection rule used in the main experiments, unified debiasing lowers TI relative to No-CF by 1.16 F1, while task-aware routing improves over unified debiasing by 1.30 F1. However, a dual-checkpoint sensitivity analysis (Table 24) shows that this pattern is not stable under TI-specific selection: when each run is evaluated at its best validation TI checkpoint, the task-aware-minus-unified difference becomes -0.24 F1. We therefore interpret the minimal model as a checkpoint-sensitive diagnostic rather than independent confirmation of the full-system effect. The full-system factorial ablation remains the primary evidence because it is larger in magnitude, stable across weighting schemes, and not subject to the same checkpoint sensitivity.

No-graph diagnostic. As another architecture diagnostic, we repeat the routing comparison after disabling the graph branch ($n = 3$ seeds). Results are shown in Table 23. The average ordering remains task-aware > unified > No-CF, but the task-aware gain over unified routing is smaller and noisier than in the full system: +1.43 TI F1 with a wide paired seed-level interval of [-1.55, 4.05] (Table 25). Unlike the minimal diagnostic and the main full-system reference pattern, the no-graph No-CF baseline falls below unified debiasing on TI. Given the small sample size ($n = 3$) and wide paired intervals, this pattern is best interpreted as diagnostic noise rather than evidence of a systematic reversal. The diagnostic suggests that the routing effect is not solely an artifact of the graph branch, while the full architecture yields a more reliable signal for the routing comparison. The No-CF condition serves as a reference; it does not use a counterfactual branch and has no CF-active phase constraint. In this diagnostic, one No-CF seed converges at epoch 6.

F Measuring Hate Speech Diagnostic

Table 22 reports the per-seed external diagnostic results on Measuring Hate Speech. The diagnostic uses a shorter training schedule than the primary benchmark due to the larger corpus size; checkpoint selection uses validation average F1. Training details are in Table 26.

Contrast	Mean Δ TI F1	Seed-level interval	$p_{\leq 0}$
Unified − No-CF	+0.73	[-1.84, 2.36]	0.2549
Task-aware − Unified	+1.43	[-1.55, 4.05]	0.1498
Task-aware − No-CF	+2.17	[-0.04, 6.41]	0.0340

Table 25: TI contrasts for the no-graph routing diagnostic. The direction is compatible with the full-system routing result, but the intervals are wide under $n = 3$, so this check is used only to assess whether the routing pattern disappears when the graph branch is removed. The interval column reports the observed paired seed-level range, and $p_{\leq 0}$ denotes the one-sided permutation-test probability that the true effect is at most zero.

MHS training details. The Measuring Hate Speech diagnostic uses a lightweight text-only BART-base encoder rather than the full graph-enhanced generation model. We train only CD and TI heads, with no graph branch and no explanation decoder. For the counterfactual settings, identity-token masking uses the same lexicon-based procedure as the main experiments. During training, matched identity tokens are masked with probability $p = 0.8$; during evaluation, the counterfactual branch uses deterministic masking. Checkpoints are selected by validation average F1 over CD and TI. We train for up to 10 epochs and report five seeds (42–46) using the same split seed (42).

G HateXplain Per-Class Tables

This section reports the per-class HateXplain TI results used to support the long-tail analysis in the main paper. Rows are sorted by training class size in ascending order. Test counts denote average per-seed test support. Vanilla per-class aggregation is included for the $p = 0.8$ comparison, while the $p = 0.6$ table is used only as a narrower unified-versus-task-aware routing diagnostic.

H Evaluation Metrics

Jaccard overlap is computed at the token level as $|A \cap B|/|A \cup B|$, where A and B are the token sets of the predicted and reference explanation respectively. For the GenEx-style generation comparison, HD-sim and ROS are not intended as universal off-the-shelf metrics; we follow the metric definitions

Item	Value
Backbone	BART-base encoder
Architecture	Text-only CD+TI diagnostic
Graph branch	Removed
Explanation decoder	Removed
Tasks	CD, TI
Max sequence length	256
Epochs	10
Batch size	16
Optimizer	AdamW
Learning rate	2×10^{-5}
Loss	CD CE + TI CE
Checkpoint selection	Best validation average F1 over CD and TI
Counterfactual mask probability	0.8
Evaluation masking	Deterministic
Seeds	42, 43, 44, 45, 46
Split seed	42

Table 26: Training details for the Measuring Hate Speech external diagnostic.

in Maity et al. (2023) and report them only as a reference comparison to the GenEx metric family. Standard ROUGE metrics use their usual definitions.

Class	Train	Test n	Unified-CF TI F1	Task-Aware TI F1	Vanilla TI F1	Δ(TA−Unified)
Asian	186	21.0	77.74 ± 3.00	82.11 ± 1.91	81.33 ± 3.90	+4.37
Hispanic	200	26.0	81.63 ± 0.00	80.84 ± 1.38	80.54 ± 3.13	-0.80
Caucasian	205	28.0	37.45 ± 4.32	43.22 ± 1.91	41.83 ± 6.33	+5.77
Arab	399	41.0	64.99 ± 0.81	64.50 ± 3.78	67.00 ± 2.60	-0.48
Refugee	422	45.0	73.94 ± 2.28	78.40 ± 3.43	80.56 ± 4.46	+4.46
Women	922	83.0	68.00 ± 3.79	69.51 ± 1.87	68.04 ± 1.33	+1.51
Homosexual	983	128.0	86.18 ± 1.29	87.48 ± 0.85	87.09 ± 0.54	+1.30
Jewish	1184	137.0	87.35 ± 1.06	87.13 ± 0.82	88.55 ± 0.33	-0.22
Islam	1249	146.0	85.81 ± 0.75	86.02 ± 0.80	85.74 ± 1.07	+0.21
Other	1708	219.0	69.69 ± 0.62	68.52 ± 2.88	68.94 ± 1.33	-1.17
African	2243	268.0	85.87 ± 0.59	85.94 ± 1.18	85.65 ± 0.69	+0.07

Table 27: HateXplain per-class TI F1 at mask probability $p = 0.8$ (macro-F1, percentage points). Rows are sorted by training class size in ascending order; Test n denotes average per-seed test support. Task-aware routing yields its clearest gains on several smaller or identity-salient classes, including *Caucasian*, *Refugee*, and *Asian*.

Class	Train	Test n	Unified-CF TI F1	Task-Aware TI F1	Δ(TA−Unified)
Asian	186	21.0	80.61 ± 3.23	79.64 ± 2.57	-0.98
Hispanic	200	26.0	78.79 ± 2.10	80.63 ± 3.32	+1.84
Caucasian	205	28.0	35.07 ± 4.58	41.94 ± 5.64	+6.87
Arab	399	41.0	68.08 ± 3.91	67.54 ± 0.88	-0.54
Refugee	422	45.0	79.53 ± 2.19	79.08 ± 7.78	-0.45
Women	922	83.0	69.59 ± 0.54	69.65 ± 1.20	+0.05
Homosexual	983	128.0	86.84 ± 0.74	86.81 ± 1.25	-0.03
Jewish	1184	137.0	87.25 ± 0.57	87.79 ± 0.83	+0.54
Islam	1249	146.0	86.13 ± 0.53	85.68 ± 0.43	-0.44
Other	1708	219.0	68.21 ± 0.27	70.23 ± 1.34	+2.02
African	2243	268.0	86.34 ± 0.44	85.58 ± 0.39	-0.77

Table 28: HateXplain per-class TI F1 at mask probability $p = 0.6$ (macro-F1, percentage points). The per-class pattern is weaker and less uniform than at $p = 0.8$: task-aware gains remain visible for some smaller classes such as *Caucasian* and *Hispanic*, but are not consistently positive across all categories.