Who Let the Dogs Out? Auditing Dogwhistle Coverage in Hate Speech Benchmarks Across Target Groups

Reva Hirave
Princeton University
Princeton, NJ, USA
rhirave@princeton.edu

Ryan Oet
Princeton University
Princeton, NJ, USA

Abstract

Automated hate speech detection relies on benchmark datasets for both training and evaluation, making benchmark composition a fairness concern that precedes classifier behavior. We audit dogwhistle coverage in two widely-used benchmarks — HateXplain and Measuring Hate Speech — using an existing glossary of 340 dogwhistle entries mapped onto a four-level coding sophistication taxonomy, ranging from explicit slurs (L1) to opaque persona signals (L4). Across 59.6k posts in the benchmark union, annotation failure is the dominant failure mode: at the stereotype-based level (L2), every racial and religious group with stable match counts is labeled non-hateful at rates exceeding 60%. Collection failure, where entire dogwhistle categories are absent from the union, becomes co-dominant at higher coding levels, with twenty group×level cells showing categorical gaps that re-annotation cannot fix. These failures are unevenly distributed: *race:black* vs. *race:middle_eastern* fails the 4/5 rule on annotation DI at L2 (DI $= 0.661$), and *lgbtq:Trans/NB* posts one of the highest annotator failure rates (81.2%, second only to *origin:undocumented*) and the lowest presence rate of any comparably-sized group at every level. We distinguish collection failure from annotation failure as mechanisms with different remediation implications, and release our mapped glossary and pipeline publicly.

1 Introduction

As automated detection of overt hate speech has improved, individuals and organizations seeking to disseminate hateful, extremist messaging while evading platform moderation have increasingly shifted to obfuscated language, known as dogwhistles (van Ginkel et al., 2025; Hagman, 2024; Weimann and Am, 2022; Bhat and Klein, 2020). This lexicon of covert expressions, phrases, and emojis that act as shorthand for hateful messages is characterized by the ability to simultaneously convey one innocuous meaning to a broad audience, and another hateful message to a smaller receptive in-group. Social networks employ automated systems to detect hate speech, yet prior work shows that classifiers often fail on coded language (Mendelsohn et al., 2023; Kirk et al., 2022). We hypothesize that training dataset composition may be a key driver of this failure.

We build upon this body of work to examine the coverage gaps of HateXplain (Mathew et al., 2021) and Measuring Hate Speech (Kennedy et al., 2020), asking whether benchmark gaps originate in collection failure (terms absent from the corpus) or annotation failure (terms present but mislabeled). This is a distinction with different remediation implications that prior work has not drawn. We take no position on whether any given instance of coded language should be moderated; our concern is that existing measurement infrastructure fails to differentiate hate of varying opacity across target groups.

We ask the following three research questions. **(RQ1)** Do existing hate speech benchmarks have lower coverage of dogwhistle terms as coding sophistication increases? **(RQ2)** When dogwhistle terms do appear in benchmarks, are they correctly labeled as hateful, and does the correct labeling rate vary across groups by coding level? **(RQ3)** Do benchmarks satisfy equalized odds across target groups? That is, is the probability of a dogwhistle term being correctly labeled hateful equal across groups, conditional on coding level?

Our contributions are threefold. We contribute **(1)** a conceptual mapping of the Mendelsohn dogwhistle glossary onto a four-level coding sophistication taxonomy which we publicly release to support future benchmark audits[1], **(2)** a quantitative audit of benchmark coverage and annotation quality disaggregated by coding level and target

[1] https://github.com/Reva-h/Auditing-Dogwhistle-Coverage

group, and **(3)** evidence distinguishing collection failure from annotation failure as sources of benchmark gap, which has concrete implications for how future datasets should be constructed. Additionally, we contribute a union of the Measuring Hate Speech, HateXplain, and Implicit Hate datasets with annotated standardized group and subgroup target labels.

2 Related Work

Coded and Implicit Hate Speech. Coded language, often referred to as dogwhistles, consists of words, phrases, or symbols that appear innocuous to a general audience but convey a specific, hateful message to a receptive in-group (Bhat and Klein, 2020; Weimann and Am, 2022). This lexicon is highly dynamic, frequently utilizing intentional typos, letter substitutions, emojis, and numeric substitutions to bypass keyword filters (Weimann and Am, 2022). Mendelsohn et al. (2023) developed a typology and glossary of these terms, noting that the performance of large language models (LLMs) in recognizing such rhetoric is often poor and varies significantly across different target groups and dogwhistle types.

Automated Hate Speech Detection Systems. While automated systems are common on platforms like TikTok and Twitter, they frequently exhibit an "algorithmic tolerance" for human intolerance because they struggle with context, intent, and the nuance of coded language (Hagman, 2024; Davidson et al., 2017; van Ginkel et al., 2025). Alt-right actors intentionally exploit these limitations to radicalize audiences while avoiding regulation (Hagman, 2024; van Ginkel et al., 2025). Recent technical efforts have attempted to address this gap; for instance, Xu et al. (2022) developed a model using contextual information to determine if dogwhistles are used innocuously or hatefully, outperforming prior LSTM and ELMo architectures. Because automated systems often fail on this coded language, it is critical to examine whether the training dataset composition itself —– the foundation of these models —– is a driving factor behind these performance failures.

Hate Speech Benchmarks. The efficacy of detection models is fundamentally limited by the quality of their training data in a "garbage in, garbage out" paradigm (Vidgen and Derczynski, 2020). Existing benchmarks vary widely in their taxonomic granularity and annotation methods. Davidson et al. (2017) highlighted the difficulty models face in distinguishing hate speech from general offensive language, particularly when explicit keywords are absent. HateXplain and the Measuring Hate Speech corpus provide large-scale, human-annotated data with target group labels, yet they do not intentionally include implicit forms of hate speech (Mathew et al., 2021; Sachdeva et al., 2022). Latent Hatred is one of the few benchmarks focused specifically on implicit hate, identifying categories such as "white grievance," "inferiority language," and "stereotypes" (ElSherief et al., 2021). Röttger et al. (2021) introduced functional tests for detection models in HateCheck, though it notably omits coded language from its evaluation suite. FETCH! identifies emergent dogwhistles using sentence embeddings (Sasse et al., 2025), but does not audit foundational benchmark coverage or annotation quality.

Additionally, prior audits have focused on classifier-level performance disparities (Mendelsohn et al., 2023), but have not decomposed benchmark gaps into collection failure (terms absent from corpora) and annotation failure (terms present but mislabeled) as distinct upstream mechanisms with different remediation implications.

Fairness Frameworks. Algorithmic fairness is a multi-faceted concept with various mathematical definitions (Barocas et al., 2023). This study utilizes the Equalized Odds framework, which enforces parity of true positive and false positive error rates across groups (Hardt et al., 2016). Equalized odds requires that the probability of a positive outcome (in this case, a correct "hateful" label) be equal across all protected groups, conditional on the actual presence of the behavior.

3 Data

3.1 Dogwhistle Glossary

Our audit operationalizes dogwhistle coverage through the glossary introduced by Mendelsohn et al., which comprises 340 entries and over 1,000 surface forms (variants) annotated with target group and dogwhistle type (Mendelsohn et al., 2023). This is the largest and most systematically annotated dogwhistle lexicon currently available, and provides a probe with which we measure benchmark coverage. We note, however, that the glossary's own coverage is uneven across target groups:

for example, transphobic and white supremacist entries are more densely represented than anti-Asian or anti-Latinx ones. To account for this, we normalize all group-level presence rates by the number of glossary entries available for that group, ensuring comparisons reflect benchmark gaps rather than glossary sparsity. As such, we treat this glossary as a probe into coded hate speech, not as a source of ground truth.

A related limitation concerns the glossary's own provenance: many Mendelsohn entries originate from political speech and advocacy discourse, a different register from the social-media posts our benchmark union draws from. Political-register dogwhistles may surface less naturally in social-media text regardless of true prevalence, understating gaps for register-mismatched terms in ways we cannot disentangle from genuine collection failure.

A further limitation is temporal. The Mendelsohn glossary is a snapshot of dogwhistle terminology as documented through 2023, and coded language evolves continuously as communities develop new terms in response to moderation (Sasse et al., 2025). Our audit therefore cannot claim to measure coverage of the full universe of dogwhistles in use. Crucially, however, this limitation operates in a consistent direction: dogwhistle terms coined after 2023 will be even less represented in benchmarks than those we audit for, meaning our coverage gap estimates are conservative lower bounds.

We classify glossary entries into three provenance tiers based on the editorial accountability of their source. Tier 1 sources are subject to formal editorial review by parties independent of the author: peer-reviewed academic publishers, established national and international journalism with professional editorial standards, and civil society organizations with institutional accountability for accuracy in their documented areas of expertise. Tier 2 sources have named authorship and organizational accountability but lack formal editorial review, or operate with an explicit advocacy mission that introduces directional bias risk without neutral editorial check. Tier 3 sources are community-edited with no identifiable accountable author (rationalwiki.org, en.wikipedia.org). Our primary analysis uses all three tiers; a separate analysis of tier 1 and 2 serves as a robustness check discussed further in Appendix A.

3.2 Benchmarking Datasets for Hate Speech

We construct our benchmark union from two large, human-annotated datasets that include target group labels: HateXplain (Mathew et al., 2021) and Measuring Hate Speech (Kennedy et al., 2020).

HateXplain consists of 20,148 posts drawn from Twitter and Gab, each annotated with a three-class hate label (hate, offensive, or normal), one or more target group labels (race, religion, gender, sexual orientation, or miscellaneous) (Mathew et al., 2021). Measuring Hate Speech contributes 39,565 posts from Twitter, Reddit, and YouTube, annotated with a continuous hate speech severity score and fine-grained target identity labels spanning race/ethnicity, religion, national origin, gender, sexual orientation, age, disability, and political ideology (Kennedy et al., 2020). Together, these datasets provide platform diversity and complementary annotation schemes, one categorical and one continuous.

Two limitations apply to both datasets in our union. First, both were collected using lexicon-based filtering rather than random sampling, meaning neither reflects the ambient distribution of hate speech on their respective platforms. This shared sampling bias reinforces our decision to avoid wild prevalence estimates as a metric, since the union cannot serve as a reliable background corpus. Second, converting Measuring Hate Speech's continuous severity scores to binary hate labels requires a threshold choice (see §4.1).

We conduct our primary audit on a union of HateXplain and Measuring Hate Speech, reflecting the datasets practitioners most commonly use for classifier training. We then replicate our audit with the Implicit Hate dataset (ElSherief et al., 2021) added to the union as a secondary analysis, reported in §5.4. By incorporating data explicitly curated to capture nuanced expressions, we verify whether our observed Disparate Impact (DI) is an artifact of dataset construction or a persistent failure of current annotation paradigms to reliably handle coded, non-explicit content.

4 Methods

4.1 Data Preprocessing

We first union HateXplain and Measuring Hate Speech into a single corpus, standardizing annotation format across datasets. Because Measuring Hate Speech uses a continuous severity score while HateXplain uses discrete labels, we follow

Kennedy et al. (2020)'s recommended threshold of 0.5 as our primary binarization cutoff. Our core claims concern the direction and pattern of annotation failures rather than their precise magnitude, and we expect them to be robust to neighboring threshold choices; we leave formal sensitivity analysis as future work. For posts that have multiple target labels, we use the label associated with the dogwhistle term. We then deduplicate the union by post ID where available, falling back to exact text matching where IDs have been stripped for privacy.

4.2 Label Harmonization

Two annotation tasks were required to bring all three datasets into a common schema. The ElSherief et al. (2021) corpus originally carried free-text target labels that are incompatible with the structured MHS taxonomy; we re-annotated these labels, mapping each free-text string to one or more fine-grained MHS identifiers (e.g., *"blacks," "black folks"* $\rightarrow$ *race_black*; *"trans people"* $\rightarrow$ *gender_transgender_unspecified*). Additionally, each entry in the Mendelsohn glossary required an inferred target label indicating which community the dogwhistle targets, assigned using the same schema. We refer to these as the **labels task** ($N = 587$ items) and the **glossary task** ($N = 340$ items) respectively.

We extended the MHS taxonomy with two additional values: *unknown_minority* for items whose target was a non-white minority community but could not be assigned to a specific group, and *self_referential_white_supremacist* for glossary entries signaling white supremacist in-group membership without an identifiable external target. Two annotators independently assigned labels to each item, with multi-label assignments permitted. Agreement was substantial across both tasks (Krippendorff's $\alpha = 0.721$ labels task, $\alpha = 0.744$ glossary task; full metrics and resolution rules in Appendix D). Disjoint disagreements, which consisted of 12.9% of labels-task cases and 36.1% of glossary-task cases, were resolved by four named rules prior to merging; remaining cases were adjudicated by the lead annotator. All other disagreements were resolved by superset union. Full details are included in Appendix D.

4.3 Proposed Taxonomy of Implicit Hate Speech

We organize the Mendelsohn glossary entries into a four-level coding sophistication taxonomy based on the degree of decoding required to connect a term to its target group, then map each Mendelsohn entry to a level based on its annotated dogwhistle type, as summarized in Table 1; §5-§6 report only L2–L4, since L1 (explicit slurs) serves here as a taxonomic anchor rather than an audited level. This four-level structure is not adopted from prior taxonomy work; we construct it specifically for this audit, organized around a single explicit criterion (degree of decoding effort required) rather than an a priori fixed count, and validate it empirically via the per-level inter-annotator agreement reported in Appendix H (Cohen's κ decreasing monotonically from L2 to L4, consistent with increasing genuine ambiguity at higher levels).

Level 1 covers explicit slurs and direct hate speech requiring no decoding. Level 2 covers stereotype-based labels and descriptors, where connecting the term to a target group requires cultural stereotype knowledge. Level 3 covers concept and policy dogwhistles, where the target group is never named even implicitly and ideological or political context is required to decode the reference. Level 4 covers persona signals and arbitrary labels, where no recoverable semantic path to the target group exists and decoding requires in-group membership.

We exclude entries labeled as humor or mockery from our primary analysis: these span multiple coding levels (some require only stereotype knowledge, others in-group knowledge), and cleanly assigning them would require a separate annotation pass. This exclusion may undercount coded language targeting Black communities specifically, since mock-AAVE entries disproportionately carry the humor label in the Mendelsohn glossary. We note this as a directional limitation on our group-level coverage estimates which merits further examination.

4.4 Audits and Metrics

4.4.1 Coverage Audit (RQ1)

For each surface form in the Mendelsohn glossary, we compute three metrics aggregated by coding level and target group. *Presence rate* is a binary indicator of whether a surface form appears anywhere in the benchmark union, aggregated to a proportion across all forms at a given level or within a given target group. Surface-form matching is case-insensitive, boundary-anchored exact-string matching against post text (regex alternation over glossary forms, anchored to whitespace or punctuation on both sides), with no edit-distance or

Table 1: Proposed taxonomy of implicit hate speech coding sophistication, mapped to Mendelsohn et al. dogwhistle types (Mendelsohn et al., 2023). Color intensity reflects explicitness: from direct slurs (L1) to maximally opaque coded language (L4).

Level	Description	Mendelsohn Type(s)	Examples
Level 1 Explicit slurs	Target group named directly; no decoding required. Baseline.	Direct slur / explicit label	[direct slurs]
Level 2 Stereotype-based	Requires cultural stereotype knowledge to connect term to group.	Stereotype-based target group label; stereotype-based descriptor; phonetic-based label	*"anchor babies"* (anti-Latinx), *"blueish"* (anti-semitic)
Level 3 Concept / policy	Requires ideological or political context; group never named, even implicitly.	Concept (policy); concept (values); concept (other); representative (Bogeyman)	*"All Lives Matter," "abolish birthright citizenship"*
Level 4 Persona signals	No recoverable semantic path to target group; requires in-group membership to decode.	Persona signal (symbol, shared culture, self-referential); arbitrary target group	*"alt-right," "biological woman"*

fuzzy-matching tolerance; this conservative choice means any misspelling, pluralization, or paraphrase of a dogwhistle is missed, so our coverage estimates should be read as a lower bound on true presence. Where surface forms overlap, the longest matching form is used (e.g., *"great replacement"* over *"replacement"*). *Type coverage* measures what proportion of Mendelsohn dogwhistle categories (rather than individual surface forms) appear at least once. *Token frequency* records how many times each matched surface form appears across the union.

4.4.2 Annotation Quality Audit (RQ2)

For surface forms that do appear in the benchmark union, we examine their assigned labels. We compute the *correct labeling rate* (the proportion of dogwhistle-containing posts labeled hateful) disaggregated by coding level and target group. This is our core annotation failure metric and requires no external resources. We further decompose matched posts into three cases: Case A (present and labeled hateful $\rightarrow$ correct), Case B (present but labeled non-hateful $\rightarrow$ annotator failure), and Case C (absent from the union entirely $\rightarrow$ collection failure). We report the Case B / (Case A + Case B) ratio as a direct measure of annotator failure, independent of collection gaps.

4.4.3 Disparity Analysis (RQ3)

To assess whether the probability of a correct "hateful" label is distributed evenly across target groups conditional on coding level, we compute *annotation Disparate Impact (DI)* (labeling rate disparity) under an equalized odds framework (Hardt et al., 2016).

We also assess structural collection gaps via coverage DI between groups at the same level. For this secondary analysis, we report the worst (minimum) of two sub-metrics—presence rate and type coverage—since a pair should only be considered non-disparate if it clears the threshold on both. For example, we compute whether the Level 3 presence rate for antisemitic terms differs significantly from the Level 3 presence rate for anti-Latinx terms.

We treat the interaction between these two metrics as our core fairness finding: groups for whom both metrics are lower at higher coding levels bear a *compounded disadvantage*, as their coded language is structurally less likely to be collected and systematically mislabeled when it is. Group pairs for which either member has fewer than 30 total matches at a given coding level are excluded from pairwise DI computation and flagged as statistically unstable.

Across both metrics, we adopt the four-fifths rule as a heuristic threshold for disparate impact, following its established use in the algorithmic fairness literature (Equal Employment Opportunity Commission, 1978): a group's rate falling below 80% of the highest-rate group's is treated as evidence of disparity. We use it descriptively rather than as a legal or causal claim. Finally, we note that our annotation DI is a proxy for, rather than a direct measurement of, equalized odds. True equalized odds requires equal true/false positive rates conditional on ground-truth hatefulness; we instead condition on surface-form match, i.e., $P(\textit{labeled hateful} | \textit{surface form matched})$. Since our FPR estimates (Appendix H) indicate surface-form matching is only 65–70% accurate at identifying genuine dogwhistle use, this substitution is more likely to attenuate observed disparities than inflate them. As such, our annotation DI values

should be read as conservative lower bounds on the true violation.

5 Results

5.1 Coverage Audit (RQ1)

Coverage is superficially adequate for most groups at L2 but masks two distinct failure modes at higher coding levels (full heatmap in Appendix B, Figure 5). First, several group×level cells show complete collection failure: matched surface forms are entirely absent despite glossary entries existing to search for. *gender:men* and *gender:women* have no matched L3 entries despite one and two glossary entries respectively, *race:asian* has no matched L2 entries despite one glossary entry existing, and *politics:communist*, *politics:democrat*, and *politics:libertarian* each have a single glossary entry that goes entirely unmatched at the one level (L4) where it exists; the political cases rest on one entry apiece and carry limited inferential weight.

Second, and more consequential at scale, twenty group×level cells exhibit type coverage below 100%, meaning entire dogwhistle *categories* (not merely individual surface forms) are absent from the union. These cells span every demographic category we code (four in race, three in religion, one in lgbtq, six in gender, one in origin, four in politics, one in disability), with gender disproportionately affected: four of its six flagged cells involve transgender subgroups at L3 or L4 (*transgender_women* at 20% and 40%, *transgender_unspecified* at 27%, and *transgender_men* at 0%), indicating the categorical gap for gender-nonconforming targets persists across multiple levels rather than resolving at any single one.

Among racial and religious groups with more than one glossary entry at the relevant level, the lowest-presence flagged cells are *race:black* at L3 (42%) and *religion:muslim* at L4 (33%); *race:middle_eastern* is flagged at both L3 (88%) and L4 (50%), and *religion:muslim* is additionally flagged at L3 (89%).

Type coverage is 100% for the large majority of remaining cells with any glossary entries, so most of the scatter in presence rates below 100% reflects surface-form undersampling within an already-covered category. This is a meaningful gap, but one that additional data collection could in principle close, unlike the categorical absences above.

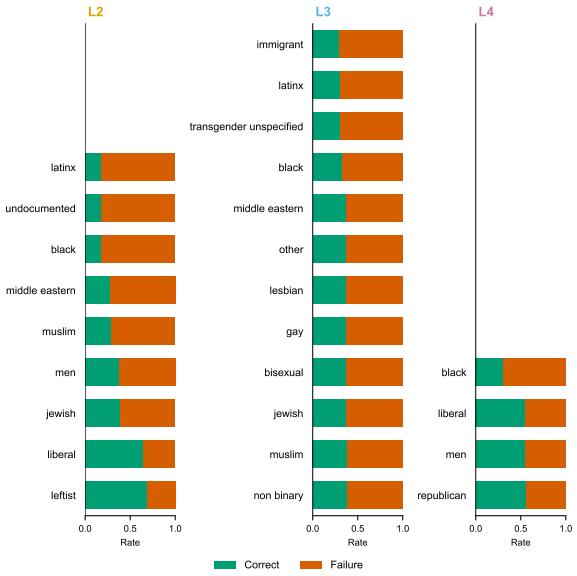


Figure 1: Correct labeling rate (Case A; teal) vs. annotator failure rate (Case B; vermillion) by target group and coding level, for groups with stable match counts ($n \geq 30$). Cells whose matches derive partly from glossary entries dual-tagged as self-referential (religion:jewish and transgender subgroups at L4) are excluded here; pooled statistics elsewhere retain them.

5.2 Annotation Quality Audit (RQ2)

Annotation failure is the dominant failure mode across all coding levels. Figure 1 shows correct labeling rate (teal) versus annotator failure rate (vermilion) disaggregated by group and coding level. Failure rate exceeds correct rate in the large majority of stable cells across all three levels.

At L2, *race:black* posts a failure rate of 82%, which is among the highest of racial and religious groups at this level, and the most consequential by absolute count. Figure 9 shows the volume asymmetry: *race:black* at L2 accounts for over 1,100 matched posts, roughly four times the next-largest group (*religion:jewish*), with Case B (non-hateful) matches dwarfing Case A (hateful) by approximately 4.5:1. This means that Black-targeted L2 dogwhistles drive the majority of all annotation failures in the union by count.

At L3, failure rates remain elevated across racial and religious groups, with *origin:immigrant* posting the highest failure rate (71%, $n = 31$) among L3 cells, narrowly ahead of *race:latinx* (70%) and *transgender_unspecified* (69%). At L4, political groups (*liberal*, *republican*) show near-parity between correct and failure rates. However, this reflects the high surface-form frequency of political terms in non-hateful contexts, and should not be in-

terpreted as evidence of superior annotation quality. The worst stable L4 cell belongs to lgbtq:Trans/NB, whose failure rate reaches 89.3% ($n = 84$; collapsed reporting group), extending the pattern documented at L3. politics:conservative (86%, $n = 7$) and race:latinx (100%, $n = 4$) post high but unstable L4 failure rates (n < 20).

To illustrate this mechanism directly: one L4 post matched to the dogwhistle *"identify as"* — a transphobic snowclone, a reusable mockery template rather than a single fixed phrase, that dismisses gender identity as something invoked casually to "cheat" a system, e.g. "I'll just identify as a woman and problem solved" — was scored non-hateful by the source benchmark (-2.08 on its continuous severity scale, well below our 0.5 binarization threshold, aggregated from three original crowdworkers with no dogwhistle glossary or instruction to look for coded language). Both of our FPR annotators, working from the post text and matched surface form alone, independently judged it *Hateful*. This is not a disagreement at the margin of interpretation: it is a direct demonstration of what term-blind annotation misses and what a targeted annotation protocol catches.

Across collapsed reporting groups (groups formed by merging related fine-grained subcategories (e.g. lgbtq:LGB, lgbtq:Trans/NB) where individual subgroups are sparse), six stable cells ($n \geq 30$) show correct labeling exceeding failure: politics:leftist (69%) and politics:liberal (65%) at L2; and gender:men (55%), politics:liberal (55%), politics:republican (57%), and religion:jewish (56%) at L4 (jewish is excluded from Figure 1; see caption). Two additional cells reach majority-correct labeling on unstable match counts: lgbtq:LGB (73%, $n = 22$) at L2 and gender:women (100%, $n = 2$) at L4.

Absolute Case A and Case B match counts appear in Figure 9 (Appendix C), alongside per-level rates for unstable groups (n < 20).

5.3 Fairness Audit (RQ3)

Equalized odds requires that the probability of a correct hateful label be equal across groups, conditional on coding level. We assess this through pairwise annotation DI (the ratio of correct-labeling rates between group pairs at the same level) and through coverage DI as a secondary lens on structural collection disparities.

Annotation DI violations are concentrated at L2 (Figure 2); no stable L3 pair fails. At L2, three pairs fail: *race:black vs. race:middle_eastern* (DI = 0.661), *race:latinx vs. race:middle_eastern* (DI = 0.651), and *religion:jewish vs. religion:muslim* (DI = 0.747). *race:black vs. race:latinx*, by contrast, passes (DI = 0.985): Black- and Latinx-targeted dogwhistles are correctly labeled at comparable rates at L2, while Middle-Eastern-targeted dogwhistles (relative to both Black- and Latinx-targeted) and Muslim-targeted dogwhistles (relative to Jewish-targeted) are not. At L3, all four stable racial and religious pairs pass, including *race:latinx vs. race:middle_eastern* (DI = 0.812) and *race:black vs. race:latinx* (DI = 0.923). At L4, only *liberal vs. republican* is stable and it passes (DI = 0.970), though near-parity here reflects high surface-form frequency in non-hateful contexts rather than genuine annotation quality.

Coverage DI (disparity in the quantity of surface forms collected per group) is a distinct question from equalized odds, and we report it as evidence of structural collection gaps rather than as a fairness metric. Figure 3 shows the distribution of pairwise presence DI ratios at each level. At L2, presence-DI failures are sparse and type-coverage DI is 1.0 for every stable pair; at L3, the distribution shifts markedly toward disparity; at L4, only one pair is stable enough to report at all, and its near-absence reflects collection failure rather than genuinely equalized coverage.

The *LGB vs. Trans/NB* pair, computed across collapsed reporting groups, posts the worst coverage DI in the audit (pooled = 0.563; worst per-level = 0.56 at L3, the only level stable enough to report), reflecting Trans/NB's substantially smaller glossary footprint relative to LGB. Because this difference reflects categorical glossary absence (§5.1) rather than ambient prevalence in the wild, we treat it as a structural collection gap rather than an expected sampling artifact.

5.4 Comparison with Implicit Hate Alone

Figure 4b shows that adding HateXplain and MHS to the Implicit Hate corpus universally extends coverage: all groups show non-negative presence rate deltas. *race:white* gains the full 100 percentage points possible, as Implicit Hate alone contains no matches for this group. However, *gender:women* (+0.57), *politics:republican* and *origin:specific_country* (each ≈+0.50), and *politics:conservative* (+0.33) see substantial but smaller gains, together reflecting Implicit Hate's

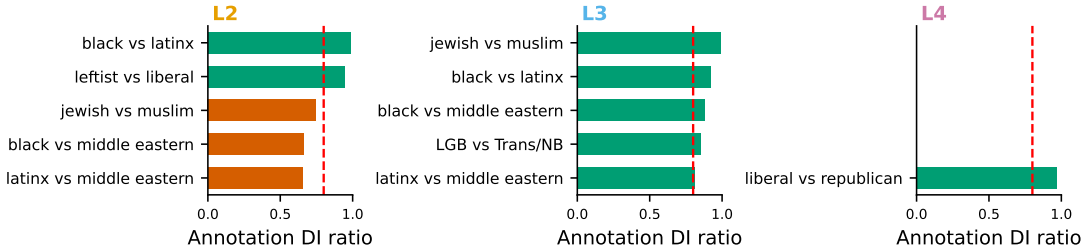


Figure 2: Pairwise annotation DI ratio by fine-grained target pair and coding level, for stable pairs only ($n \geq 30$ in both groups at that level). Annotation DI is computed as $\min(\hat{p}_A, \hat{p}_B)/\max(\hat{p}_A, \hat{p}_B)$, where $\hat{p}$ is the correct-labeling rate (Case A / (Case A + Case B)) for each group. Teal bars pass the 4/5 threshold (DI ≥ 0.80); vermilion bars fail.

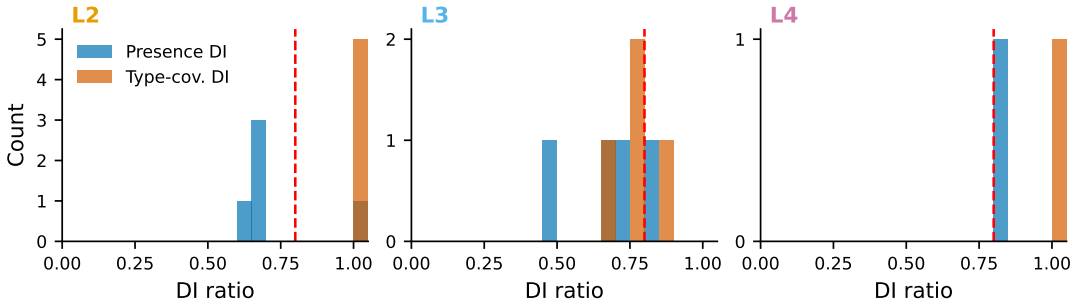


Figure 3: Distributions of pairwise coverage DI ratios across stable group pairs ($n \geq 30$ in both groups) at each coding level. Presence DI (blue) measures surface-form coverage disparity; type-coverage DI (orange) measures dogwhistle category coverage disparity. The dashed red line marks the 4/5 rule threshold (DI $= 0.80$).

collection priorities outside these groups.

Figure 4a tells the opposite story for annotation quality. The union degrades correct labeling for the groups most disadvantaged in the primary audit. The largest declines affect *origin:undocumented* ($\Delta \approx -0.45$), *race:latinx* (-0.35), *race:black* (-0.28), and *origin:immigrant* (-0.20). The mechanism is volume dilution: HateXplain and MHS contribute large counts of mislabeled matches for these groups (1,408 for *race:black* versus Implicit Hate's 112) overwhelming Implicit Hate's superior annotation quality on the few posts it does contain.

The coverage-quality trade-off is group-specific and asymmetric. The dataset that achieves better coverage for a given community is not necessarily the dataset that annotates that community's dogwhistles more accurately, and no single dataset achieves both properties across all target groups.

6 Discussion

6.1 Collection Failure vs. Annotation Failure

Standard benchmarks conflate two distinct phenomena: collection failure and annotation failure. At the surface-form level, most targeted communities' dogwhistles do appear in HateXplain and MHS; the gap is in labeling, not collection, and requires re-annotation rather than additional data.

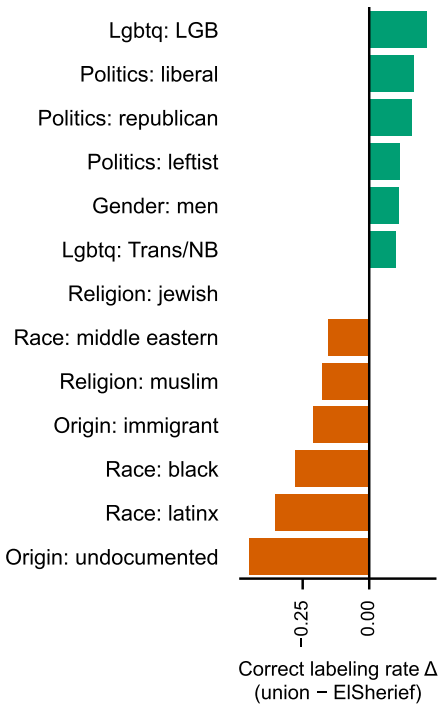
Categorical absence (type coverage below 100%, §5.1) is a qualitatively different failure that *cannot* be addressed by re-annotation: the benchmark contains no examples from which a classifier could learn these patterns regardless of label quality. Closing these gaps requires targeted collection of L3/L4 coded language, which by definition requires in-group or expert knowledge to identify at collection time.

Independent validation comes from our FPR annotation sample: human annotators and benchmark labels agreed on only 46.2% of L2 matched posts, rising to 63.8% at L3 and 69.1% at L4, driven by posts annotators judged hateful but benchmarks labeled non-hateful. This confirms the mislabeling is not an artifact of our matching procedure (Appendix H).

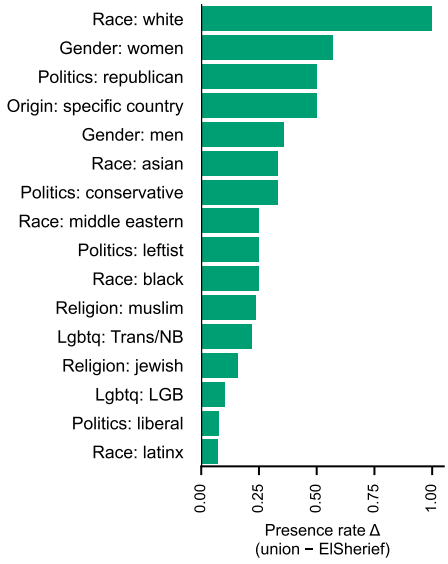
6.2 Compounded Disadvantage

Our core fairness finding is the interaction between coverage and annotation quality: the same groups that suffer coverage gaps also suffer annotation failures, and the pattern is not random.

race:black's largest-volume matched-post count (§5.2) coincides with the third-highest pooled fail-



(a) Correct labeling rate Δ (union − Implicit Hate) for groups with $n \geq 30$ following the union. Teal bars indicate groups for which the union improves annotation quality; vermilion bars indicate groups for which Implicit Hate alone achieves higher correct labeling.



(b) Presence rate Δ (union − Implicit Hate).

Figure 4: Effect of adding HateXplain and MHS to the Implicit Hate corpus.

ure rate among stable groups (80.0%), behind *origin:undocumented* (82.4%) and *lgbtq:Trans/NB* (81.2%); *race:latinx* follows closely at 79.9%, leaving four groups clustered within 2.5 points of one another at the top of the pooled failure-rate distribution. Its L2 coverage DI (black vs. middle eastern, $DI = 0.696$, failing the 4/5 threshold) means Black-targeted terms are simultaneously underrepresented relative to other racial groups and mislabeled at a higher rate when present.

lgbtq:Trans/NB presents a different but equally severe pattern: the lowest presence rate in the audit (33.8% overall, pooled) and one of the highest annotator failure rates (81.2%, second only to *origin:undocumented*); this is both the least collected and the most consistently mislabeled (worst pairwise coverage DI: LGB vs. Trans/NB, §5.3). Annotation quality for both groups is similarly poor at L3, the only level where both are stable, with correct-labeling rates of 37.50% (LGB) and 32.08% (Trans/NB), contributing to an annotation DI ratio of 0.855 that passes the 4/5 threshold. Neither community can expect classifiers trained on standard benchmarks to detect their coded hate speech. This is not because classifiers are flawed, but because the training signal is absent or inverted for both groups.

7 Conclusion

This audit shows that benchmark gaps in dogwhistle coverage are not one problem but two: collection failure, where coded language is structurally absent from training data, and annotation failure, where it is present but mislabeled. The two mechanisms compound for specific communities — Black-targeted and Trans/NB-targeted dogwhistles are simultaneously undercollected and systematically mislabeled — so a classifier trained on these benchmarks inherits a predictable blind spot rather than a random one. Closing the annotation gap requires re-labeling; closing the collection gap requires targeted, in-group-informed data collection that re-annotation alone cannot substitute for.

Two extensions follow directly. First, our findings describe benchmark composition, not classifier behavior; auditing deployed detectors (e.g., HateBERT, the Perspective API) on our benchmark union, with attention to group-level false negative rates, would connect these dataset-level gaps to downstream model failures. Second, the Mendelsohn glossary is a 2023 snapshot of an actively evolving lexicon; re-running this audit against an updated lexicon would likely reveal deeper gaps, particularly at L3 and L4.

8 Limitations

Presence rate is sensitive to surface-form count: groups with more documented variants per glossary entry face a larger denominator, and rare or evolving variants are less likely to appear in any fixed corpus. This potentially causes us to underestimate presence for well-documented groups independent of true coverage. Type coverage, which only requires one matched form per category, is invariant to this and should be weighted more heavily in cross-group comparisons.

Additionally, surface-form matching produces non-negligible false positives across all coding levels. Per-level FPR estimates (stratified sample of 75 posts per level, two annotators; opposing judgments excluded, resolution rules in Appendix H; Wilson 95% CIs) are: L2: 0.308 [0.209, 0.428]; L3: 0.362 [0.251, 0.491]; L4: 0.345 [0.234, 0.477]. Confidence intervals overlap substantially across levels and the monotonic increase we hypothesized does not hold, precluding level-to-level comparisons. Roughly 30–36% of matched posts may not represent genuine dogwhistle deployment (Appendix H). This cuts two ways: since false positives can only inflate apparent presence, coverage gaps (RQ1) remain conservative lower bounds on true collection failure; since some are false matches correctly labeled non-hateful, raw annotation failure rates (RQ2) are conservative upper bounds on genuine annotator error. Furthermore, our FPR estimates are pooled rather than stratified by target group. If some groups' surface forms appear disproportionately in counter-speech or quotation contexts, our annotation DI partly reflects this context-of-use disparity rather than pure annotator failure; we cannot rule this out with the current sample and leave group-stratified FPR to future work (see Appendix H for full details).

Finally, the four-fifths rule is a fixed-threshold heuristic that does not account for sampling uncertainty in correct-labeling rates; significance testing of pairwise rate differences (e.g., Fisher's exact tests with multiple-comparison control) is a complementary lens we leave to future work, noting that threshold-based and test-based flags need not coincide for cells near our stability minimum.

Level	N	FPR	95% CI	Cohen's κ
L2	65	0.308	[0.209, 0.428]	0.510
L3	58	0.362	[0.251, 0.491]	0.454
L4	55	0.345	[0.234, 0.477]	0.349

Table 2: False positive rate estimates by coding level, from stratified manual annotation (75 posts per level, two annotators). Items where both annotators agreed on a definitive judgment retain that label; items where exactly one annotator marked Unclear adopt the other annotator's definitive label; items where both marked Unclear (5) or gave opposing definitive judgments (42) are excluded, yielding N = 65, 58, and 55 per level. FPR is the proportion of retained items resolved as Not Hateful. Wilson 95% CIs. Cohen's κ is per-level inter-annotator agreement on the raw three-way judgments. Overlapping CIs preclude level-to-level comparisons.

9 Acknowledgments

We thank Lydia Liu and Kara Schechtman for guidance and feedback; Manoel Horta Ribeiro and Blossom Metevier for review; and Phillip Miavelstuck and Jing Liu for FPR annotation work.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Prashanth Bhat and Ofra Klein. 2020. Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In Gwen Bouvier and Judith E. Rosenbaum, editors, *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, pages 151–172. Springer International Publishing, Cham.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Equal Employment Opportunity Commission. 1978. 29 C.F.R. 1607.4 – Information on impact (Uniform Guidelines on Employee Selection Procedures). Electronic Code of Federal Regulations (eCFR). Accessed: 2026-07-03.

Malvin Hagman. 2024. *Algorithmic tolerance of human intolerance : An explorative study on the proliferation of the alt-right's abuse of TikTok's algorithmic content moderation system*.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In

Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott A. Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

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.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

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 @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Kuleen Sasse, Carlos Alejandro Aguirre, Isabel Cachola, Sharon Levy, and Mark Dredze. 2025. Making FETCH! happen: Finding emergent dog whistles through common habitats. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5709, Vienna, Austria. Association for Computational Linguistics.

Bibi van Ginkel, Yael Boerma, and Tanya Mehra. 2025. Reading Between the Lines: The Importance of Human Moderators for Online Implicit Extremist Content Moderation.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Gabriel Weimann and Ari Ben Am. 2022. Digital Dog Whistles: The New Online Language of Extremism. *Digital Dog Whistles*, 2(1).

Depeng Xu, Shuhan Yuan, Yueyang Wang, Angela Uchechukwu Nwude, Lu Zhang, Anna Zajicek, and Xintao Wu. 2022. Coded Hate Speech Detection via Contextual Information. In *Advances in Knowledge Discovery and Data Mining*, pages 93–105, Cham. Springer International Publishing.

A Glossary Provenance and Robustness Check

As a reminder, glossary entries are classified into three provenance tiers by editorial accountability: Tier 1 (peer-reviewed publishers, established journalism, and civil-society organizations with institutional accountability), Tier 2 (named authorship without formal editorial review, including advocacy sources), and Tier 3 (community-edited sources with no identifiable accountable author, e.g., `rationalwiki.org`, Wikipedia). These three tiers differ substantially in size: Tier 3 sources (primarily `rationalwiki.org`) account for 179 of 340 entries (52.6%), concentrated in LGBTQ+ and white supremacist categories. Our primary analysis uses all three tiers to maximize glossary coverage; this appendix reports a robustness check restricted to Tier 1 and Tier 2 entries only, summarized in Table 3.

Two of the three L2 annotation-DI violations are robust to tier restriction. *race:black vs. race:middle_eastern* annotation DI shifts from 0.661 (all tiers) to 0.622 (Tier 1+2), and *race:latinx vs. race:middle_eastern* shifts from 0.651 to 0.737 — both pairs fail the four-fifths rule under both tier sets. The third L2 violation, *religion:jewish vs. religion:muslim* (DI = 0.747 under the full glossary), does not survive restriction: it passes under Tier 1+2 (DI = 0.829). *race:black vs. race:latinx* passes under the full glossary already (DI = 0.985) and remains passing under Tier 1+2 (DI = 0.844): Black- and Latinx-targeted dogwhistles are annotated at comparably poor rates at L2 (82% and 82% failure, respectively), which is precisely why both pairs fail when each is compared against the

Pair	Level	Metric	Full	Tier 1+2	4/5 rule (full → T1+2)
race:black vs. race:latinx	L2	Annotation	0.985	0.844	pass → pass
race:black vs. race:middle_eastern	L2	Annotation	0.661	0.622	fail → fail
race:latinx vs. race:middle_eastern	L2	Annotation	0.651	0.737	fail → fail
religion:jewish vs. religion:muslim	L2	Annotation	0.747	0.829	**fail → pass**
race:black vs. race:middle_eastern	L2	Coverage	0.696	0.600	fail → fail
race:black vs. race:latinx	L3	Annotation	0.923	0.979	pass → pass
race:black vs. race:middle_eastern	L3	Annotation	0.880	0.960	pass → pass
race:latinx vs. race:middle_eastern	L3	Annotation	0.812	0.939	pass → pass
religion:jewish vs. religion:muslim	L3	Annotation	0.990	0.963	pass → pass
lgbtq:LGB vs. lgbtq:Trans/NB	L3	Annotation	0.855	0.966	pass → pass
race:black vs. race:latinx	L3	Coverage	0.705	0.667	fail → fail
race:black vs. race:middle_eastern	L3	Coverage	0.484	0.480	fail → fail
race:latinx vs. race:middle_eastern	L3	Coverage	0.667	0.667	fail → fail
religion:jewish vs. religion:muslim	L3	Coverage	0.750	0.750	fail → fail
lgbtq:LGB vs. lgbtq:Trans/NB	L3	Coverage	0.556	0.500	fail → fail
politics:liberal vs. politics:republican	L4	Annotation	0.970	0.970	pass → pass
lgbtq:LGB vs. lgbtq:Trans/NB	Pooled	Coverage	0.563	0.667	fail → fail
race:latinx vs. race:middle_eastern	Pooled	Coverage	0.667	0.800	**fail → pass**
politics:liberal vs. politics:republican	Pooled	Coverage	0.667	0.857	**fail → pass**

Table 3: Tier 1+2 robustness comparison for every pair discussed in the main text and Appendix A, at every level where it is stable ($n \geq 30$ both groups). Pass/fail flips are bolded. This table covers 19 pair/level/metric rows; the complete 105-row comparison across all reporting-group granularities is available in the released code and data.

better-annotated *race:middle_eastern* (72.5% failure) rather than against each other. Coverage DI for *race:black vs. race:middle_eastern* shifts from 0.696 to 0.600, and remains failing under both tier sets.

The *lgbtq:Trans/NB* finding is robust in overall conclusion under both metrics, but the direction of the tier-restriction effect is no longer uniform across levels. Pooled coverage DI shifts from 0.563 (all tiers) to 0.667 (Tier 1+2) — a *less* disparate estimate, consistent with our expectation that transphobic entries are more densely represented in Tier 3 and their removal should narrow the presence-rate gap. At L3 specifically, however, coverage DI moves the other way, from 0.556 to 0.500, a marginally *more* disparate estimate under restriction. Coverage DI fails the four-fifths threshold under both tier sets at both levels regardless of direction. Annotation DI at L3 shifts from 0.855 to 0.966 and passes under both tier sets. We do not have a confirmed explanation for this pooled-vs.-level-specific directional inconsistency and do not offer a post-hoc mechanism here; we report it as an open question.

One finding does not survive tier restriction, and the way it fails is itself informative. At L2, the § 5.3 finding that *religion:jewish vs. religion:muslim* fails the four-fifths rule (DI = 0.747) disappears under Tier 1+2: it passes under the restricted glossary (DI = 0.829). This leaves *race:black vs. race:middle_eastern* and *race:latinx vs. race:middle_eastern* as the only L2 annotation-DI violations robust to tier restriction. At L3, there is nothing to lose: all five stable annotation-DI pairs pass under the full glossary already, before any tier restriction is applied, and all five continue to pass under Tier 1+2. Coverage DI at L3 shows a different, more stable pattern: all five pairs fail under both tier sets, and four of the five shift by no more than 0.038 — essentially flat. The fifth, *lgbtq:LGB vs. lgbtq:Trans/NB*, shifts further (0.556 to 0.500, a change of 0.056) but remains well below the four-fifths threshold under either tier set. This asymmetry — annotation DI's one L2 violation resolves under tier restriction while its L3 violations were never present under the full glossary to begin with, whereas coverage DI's violations persist essentially unchanged at every level — is consistent with this paper's central distinction between collection failure and annotation failure as separate mechanisms (§ 6.1): where annotation-quality violations do appear, they can be sensitive to which glossary tier is probed, while collection gaps are structural and persist regardless of which tier restricts the probe.

These findings do not exhaust the space of possible tier-restriction effects. Across the full pairwise comparison (105 pair × level × metric com-

binations across both fine-grained and collapsed reporting-group granularities), 10 rows (5 unique pair/level/metric combinations, each computed at two granularities) changed four-fifths-rule conclusions under Tier 1+2 restriction, all in the less-disparate direction (failing under the full glossary, passing under Tier 1+2): *religion:jewish vs. religion:muslim* at L2 annotation (discussed above), *race:latinx vs. race:middle_eastern* and *politics:liberal vs. politics:republican* at the pooled coverage level, and *politics:leftist vs. politics:liberal* at both L2 and pooled coverage. Most are not individually discussed in the main text; the complete comparison is available in the released code and data.

B Coverage Audit Detail

Figure 5 shows the full presence rate and type coverage heatmap across all reporting groups and coding levels. Red borders mark cells where type coverage falls below 100%, indicating categorical dogwhistle absence rather than surface-form undersampling.

Figure 7 shows token frequency for the top 10 groups at each coding level. At L2, *race:black* terms account for over 1,100 matches — more than four times the next-largest group (*religion:jewish*) — creating a substantial imbalance in the training signal available for different racial groups. At L3, the highest-frequency entry is minority, a broad aggregating category rather than a group-specific token; the next-highest, muslim, is group-specific. At L4, political terms (*liberal*, *republican*, *men*) dominate, consistent with the high ambient frequency of persona signals in non-hateful text.

C Annotation Audit Detail

Figure 7 disaggregates the token frequency imbalance underlying the L2 *race:black* finding in §5.2, and surfaces a complication at higher coding levels: the highest-frequency L3 entry, *minority*, is an aggregating category rather than a group-specific dogwhistle. (Self-referential white-supremacist terms, which would otherwise rank alongside it, are excluded from the figure entirely, per its caption.) The prominence of pan-minority language at L3 reflects patterns excluded from our per-group disparity analysis (§ 4.4.3) precisely because they cannot be attributed to a single target group, meaning the L3 token frequency distribution is not directly comparable to L2 and L4, where every top-10 token is group-specific.

Figures 8 and 9 corroborate the primary audit's L2 findings at the pooled and absolute-count levels, respectively. Figure 8 confirms that the elevated failure rates reported in §5.2 and §6.2 are not an artifact of any single coding level: *race:black*, *race:latinx*, *origin:undocumented*, and *lgbtq:Trans/NB* all post failure rates exceeding 60% once Case A and Case B counts are summed across L2, L3, and L4. Figure 9 shows the absolute scale behind this pattern, and the change in x-axis scale across its three panels is itself informative: the maximum match count drops by roughly an order of magnitude from L2 to L3, visually reinforcing the collection-failure pattern from § 5.1. at higher coding levels, not only do fewer surface forms appear in the union at all, but the ones that do appear are matched far less frequently.

D Label IAA and Resolution Rules

Metric	Labels Task	Glossary Task
N items	587	340
Exact match %	65.42%	71.47%
Mean Jaccard similarity	0.745	0.773
Macro-avg Cohen's κ[a]	0.608	0.470
Krippendorff's α (Jaccard dist.)	0.721	0.744

[a] Calculated using $N = 54$ distinct labels for the Labels Task and $N = 33$ distinct labels for the Glossary Task.

Table 4: Inter-annotator agreement across both annotation tasks.

Table 4 reports inter-annotator agreement across both annotation tasks, computed on raw pre-resolution labels. Labels are treated as sets; Krippendorff's α uses Jaccard distance ($d = 1 -$ Jaccard) as the disagreement metric.

We classified each disagreement case into one of five categories: exact agreement, both annotators assigned *_unknown*, additive disagreement (one annotator's set is a strict subset of the other's), partial overlap, and disjoint (no shared labels). Among true disagreements, 12.9% of labels-task cases and 36.1% of glossary-task cases involved disjoint assignments, a substantially higher rate in the glossary task, driven by three systematic patterns we identified through manual inspection. Disagreements clustered into four categories, each addressed by a named resolution rule applied before merging:

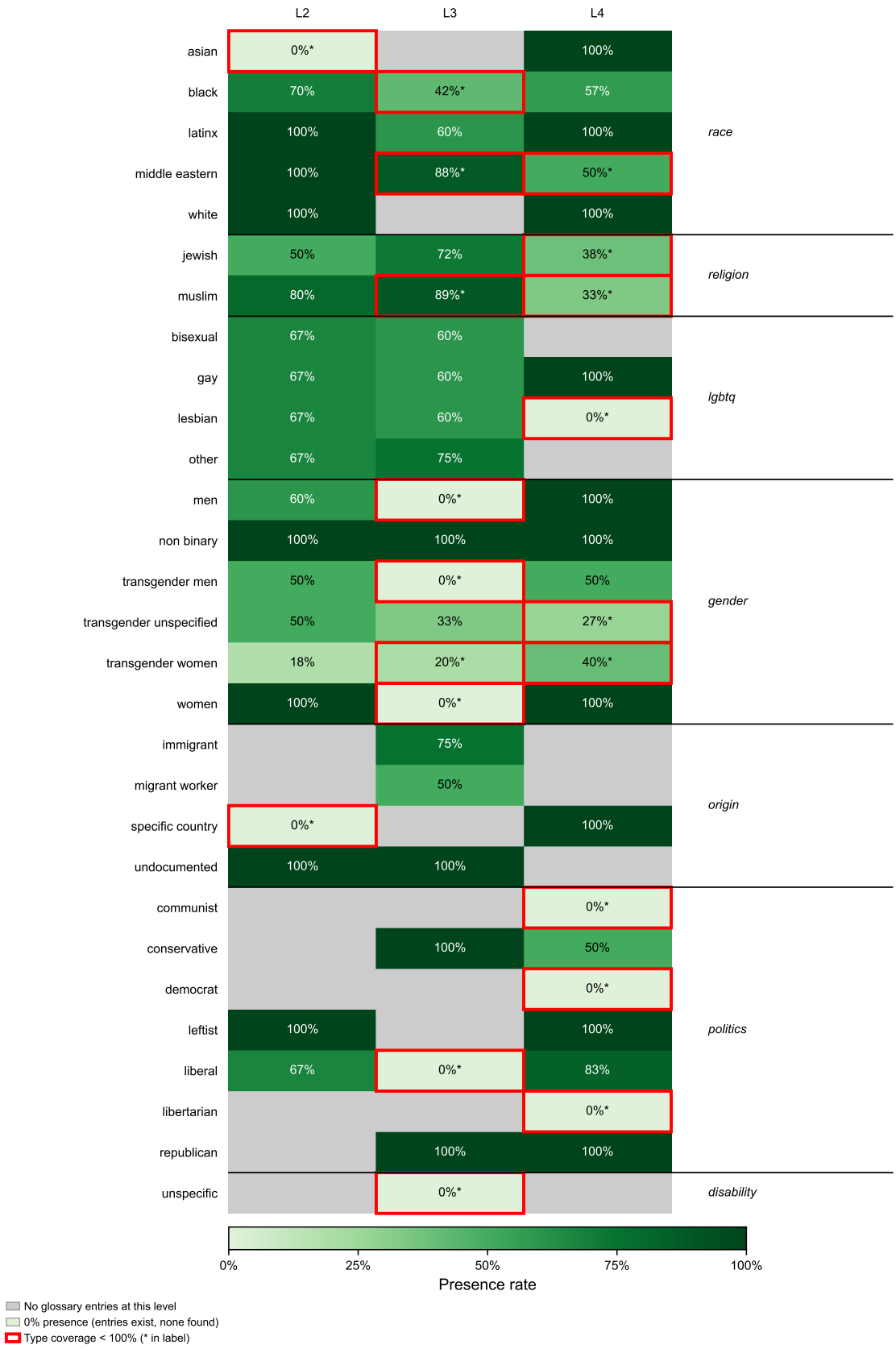


Figure 5: Presence rate by reporting group and coding level. Cell color encodes the proportion of Mendelsohn surface forms matched in the benchmark union (0%: light green; 100%: dark green). Gray cells indicate no glossary entries at that level for that group. Red borders and asterisks mark cells where type coverage falls below 100%, indicating entire dogwhistle *categories* absent from the union rather than merely undersampled surface forms.

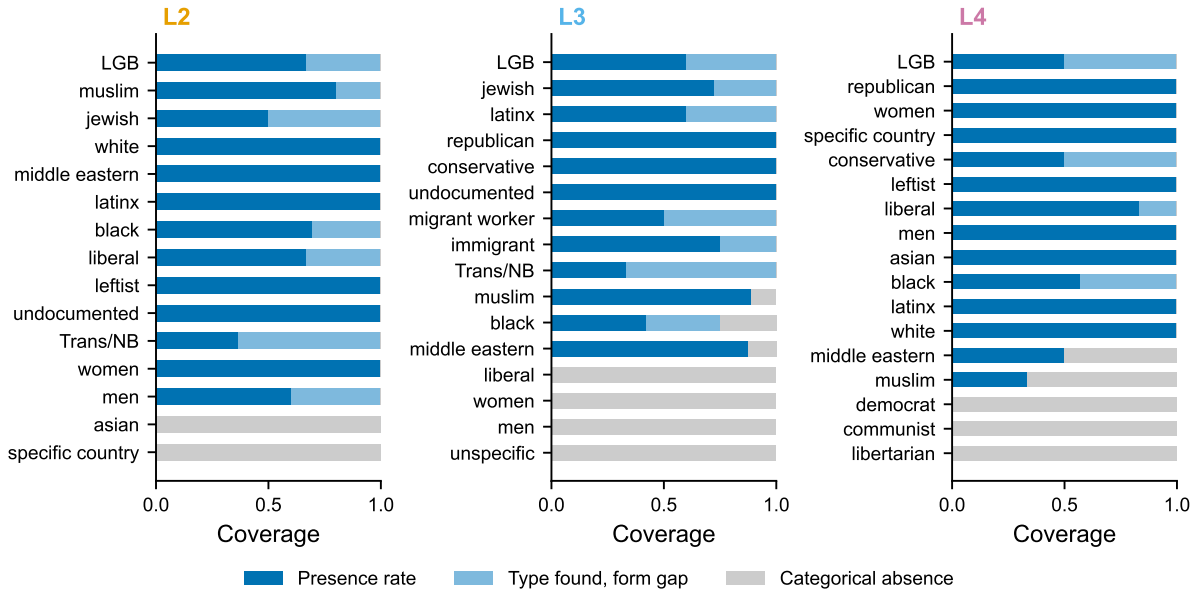


Figure 6: Presence rate and type coverage by target group and coding level. Each bar represents one group at a given level. The dark blue segment shows presence rate (fraction of glossary surface forms appearing in the union); the medium blue extension shows the gap between presence rate and type coverage (dogwhistle categories are represented but not all surface forms are found); the gray extension shows categorical absence (entire dogwhistle categories missing from the union). When the full bar falls short of 1.0, collection failure is categorical and cannot be remediated by re-annotation.

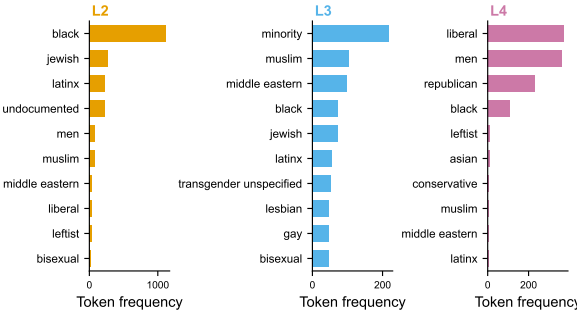


Figure 7: Top-10 groups by token frequency at each coding level. Token frequency counts the total number of times matched surface forms appear across the benchmark union. Self-referential terms are excluded from these counts.

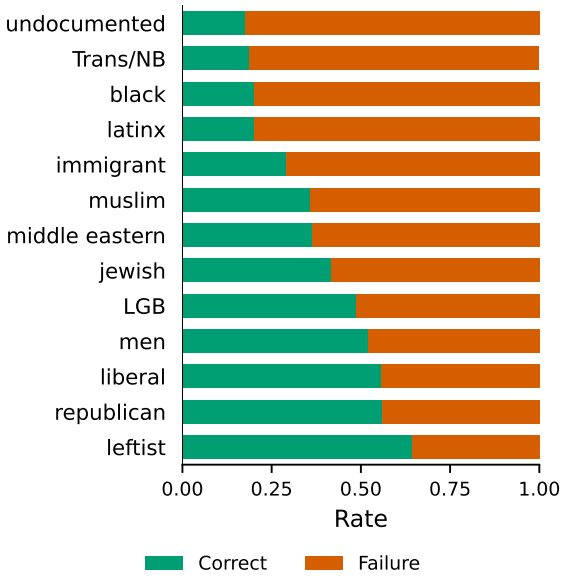


Figure 8: Correct labeling rate (teal) vs. annotator failure rate (vermillion) by reporting group, pooled across coding levels. For each group, rates are computed by summing Case A and Case B counts across L2, L3, and L4 and recomputing from pooled totals. Only groups with at least 30 total matches after pooling are shown.

1. **Unknown vs. specific** (labels task: 70 cases, 11.9%; glossary task: 18 cases, 5.3%): one annotator assigned *_unknown* where the other assigned specific labels. Resolution: adopt the specific label, treating *_unknown* as absence of information rather than a label value.
2. **Transgender granularity** (glossary task: 15 cases, 4.4%): one annotator used *gender_transgender_women* where the other used *gender:transgender_unspecified* for TERF-adjacent dogwhistles. Resolution: use *gender:transgender_unspecified*, since the terms in question target trans people broadly rather than trans women specifically.
3. **Self-referential vs. target framing** (glos-

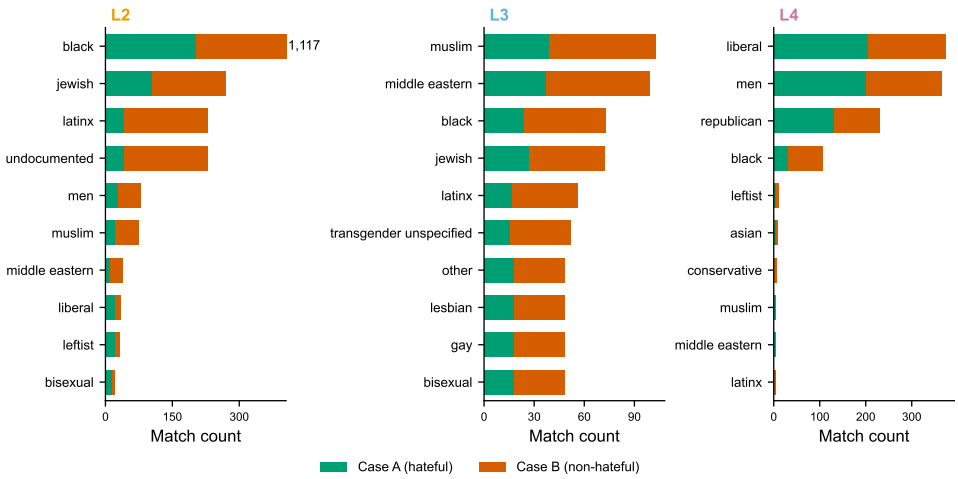


Figure 9: Absolute Case A (correctly labeled hateful; green) vs. Case B (mislabeled non-hateful; vermillion) match counts by target group and coding level, for the top-10 groups by total match count at each level. Note the change in x-axis scale across panels.

sary task: 8 cases, 2.4%; labels task: 2 cases, 0.3%): one annotator used *self_referential_white_supremacist* where the other assigned target-group labels. Resolution: use *self_referential* only when the term has no identifiable external target; otherwise use the target-group label.

4. **Pan-group specificity explosion** (labels task: 6 cases, 1.0%): one annotator decomposed pan-group terms (e.g., *"colored people"*) into multiple specific race labels; the other used *unknown_minority*. Resolution: use *unknown_minority*, since decomposition assigns false specificity.

After applying these rules, 30 cases remained unresolved (12 glossary, 18 labels). These consisted of genuine taxonomy disputes, primarily concerning whether civic dogwhistles (e.g., *"voter fraud," "thin blue line"*) target *race_black* specifically or *unknown_minority* broadly. Such cases were adjudicated manually by the lead annotator. Resolution rules changed the final label set relative to the naive superset for 6.5% of glossary-task items (22 of 340) and 1.4% of labels-task items (8 of 587) — a subset of the 41 and 78 items respectively where a named rule applied, since one rule (*unknown vs. specific*) and one *self-referential vs. target framing* case produced the same label set the naive superset would have given anyway. Among items that did change, mean Jaccard similarity between naive and resolved labels was 0.500 (glossary) and 0.268 (labels), confirming that where rules mattered, they made substantive rather than cosmetic changes. Of the 587 labels-task items and 340 glossary-task items, resolution methods were: exact agreement (281 / 222), both unknown (105 / 21), additive superset (99 / 40), named rules (78 / 41), partial overlap superset (6 / 4), and manual adjudication (18 / 12).

E Disparity Audit Detail

Detailed Disparate Impact (DI) ratios across all taxonomy levels and demographic groups are visualized in Figures 10 and 11.

F Cross-Level Consistency

Figure 12 tracks how presence rate and correct-labeling rate change across coding levels for the nine target groups with non-missing values for both metrics at L2, L3, and L4. The presence-rate trajectories (left) are heterogeneous rather than uniformly declining, complicating any simple "coverage degrades with coding sophistication" reading of §5.1. *race:middle_eastern* and *gender:transgender_unspecified* decline monotonically across all three levels; *gender:transgender_women* increases monotonically from a low base; *gender:non binary* sits at ceiling throughout; and *race:black*, *race:latinx*, *religion:jewish*, *religion:muslim*, and *lgbtq:gay* are non-monotonic, either dipping at L3 before recovering by L4 (*race:black*, partially; *race:latinx*, fully; *lgbtq:gay*, sharply, to ceiling) or peaking at L3 before dropping sharply by L4 (*religion:jewish*, *religion:muslim*). This heterogeneity is consistent with §5.1's distinction between presence-rate

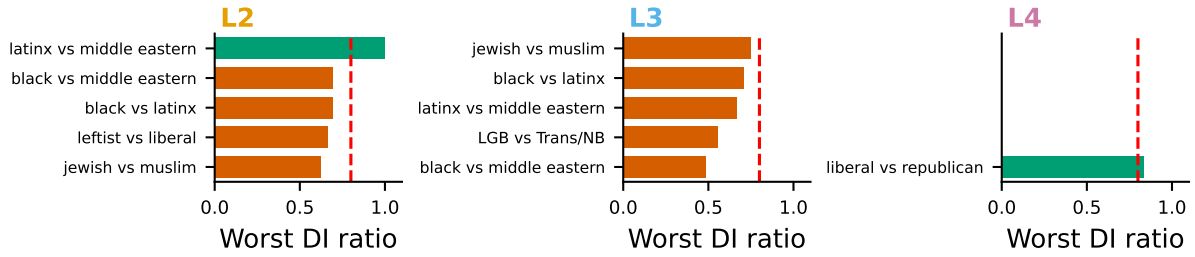


Figure 10: Worst (minimum) pairwise coverage DI ratio by fine-grained target pair and coding level, for stable pairs only. Blue bars pass the 4/5 threshold (≥ 0.80); red bars fail.

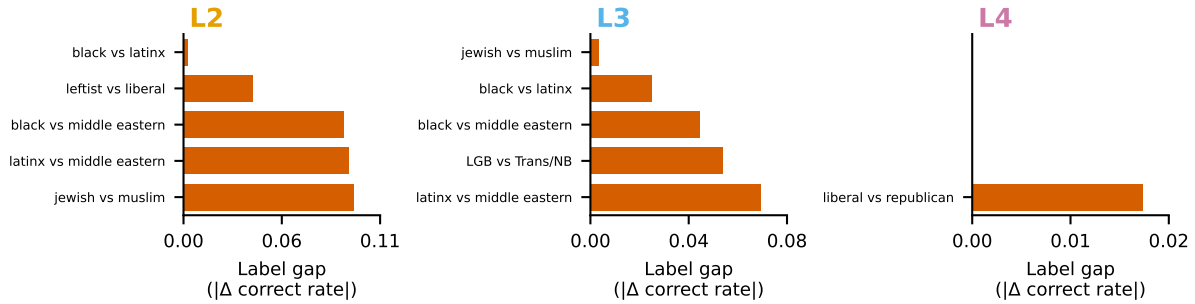


Figure 11: Pairwise annotation labeling rate gap (|Δ correct rate|) by fine-grained target pair and coding level. Bars show the absolute difference in correct labeling rate between each stable pair. All bars are shown in red because labeling rate gaps represent a disparity finding regardless of direction. Only stable pairs ($n \geq 30$ for both groups at that level) are shown.

undersampling and categorical type-coverage absence: a group's raw presence rate can move in either direction level-to-level even as its categorical coverage status (present vs. entirely absent) remains fixed.

Correct-labeling rate (right) is comparatively stable across levels for the groups with solid match counts throughout: *race:black* (0.18–0.33), *religion:jewish* (0.38–0.56), and *religion:muslim* and *race:middle_eastern* at L2–L3 (0.27–0.38) all move within a narrow band. Several small-n cells swing to a ceiling or floor at L4 instead: *race:middle_eastern* and *religion:muslim* both reach 1.0 and *race:latinx* drops to 0.0, each on fewer than five matched posts, consistent with small-sample instability rather than a genuine level effect. The most dramatic swing overall remains *gender:transgender_women*, whose correct-labeling rate jumps from 0.0 at L2 to 1.0 at L3 before partially reverting to 0.25 by L4; given this group's presence rate of only 18–20% at those same levels, the underlying match count is small enough that this is almost certainly a small-n artifact rather than a genuine swing in annotation quality, consistent with the stability caveats applied elsewhere in the audit (§ 4.4.3).

G Implicit Hate Comparison Detail

In Figure 13, the *religion:jewish vs. religion:muslim* pair (evaluated at L3, where both corpora are stable) shows the largest shift: the union improves both the worst pairwise DI (Δ = +0.50) and the absolute label gap (Δ = -0.12) relative to Implicit Hate alone. The narrowed gap partly reflects leveling down rather than improvement: the union degrades muslim correct-labeling at L3 (Δ = -0.15) far more than jewish (Δ = -0.03), consistent with the volume-dilution mechanism in § 5.4. The only pair exhibiting a genuine trade-off is *lgbtq:LGB vs. lgbtq:Trans/NB*, which improves coverage DI (Δ = +0.24) while marginally widening the label gap (Δ = +0.006).

H FPR Estimation Details

To estimate the false positive rate of surface-form matching, we drew a stratified random sample of 75 posts per coding level (225 total) from matches against the full glossary, excluding self-referential and pan-minority targets. Two annotators independently judged each post as *Hateful*, *Not Hateful*, or *Unclear*, with access only to the post text and matched surface form. Coding level, target group,

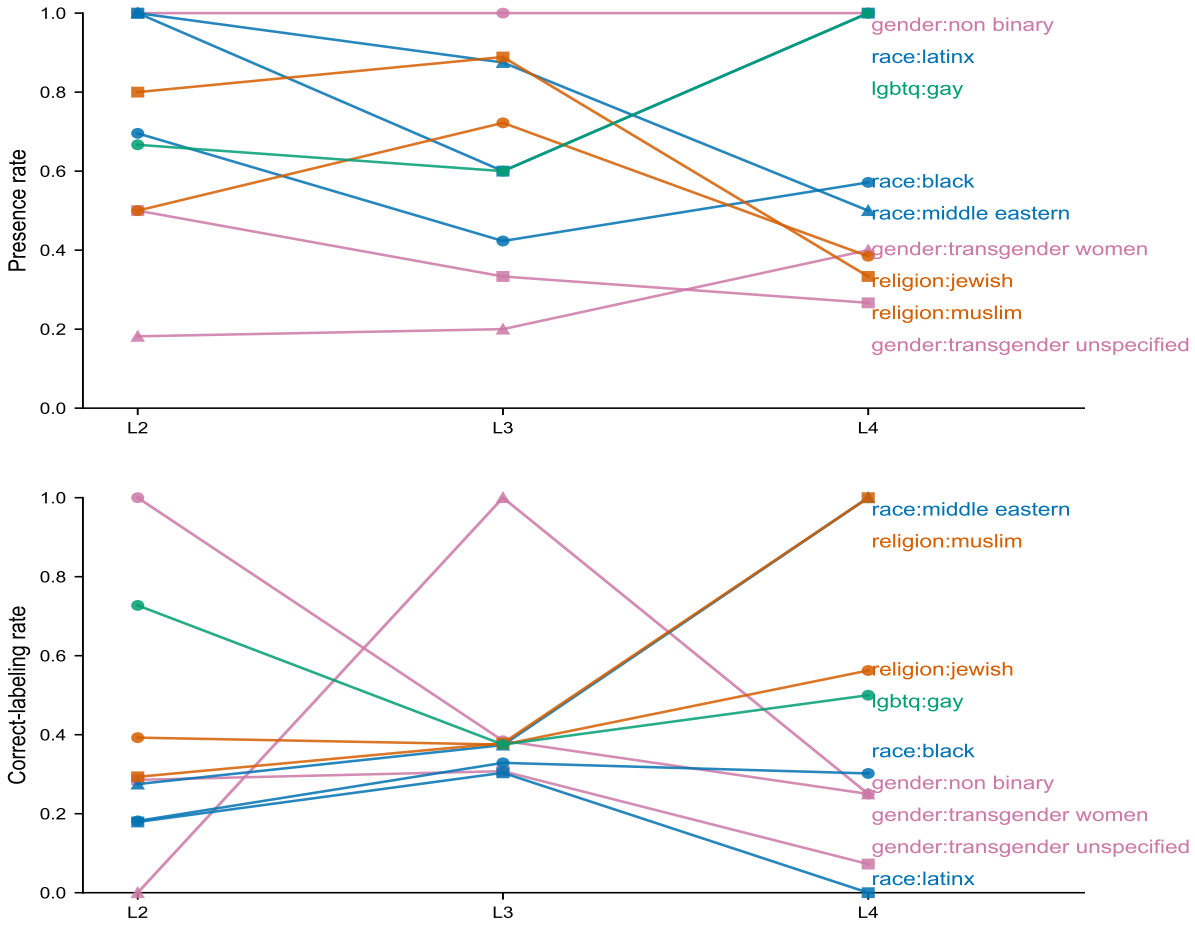


Figure 12: Presence rate (left) and correct-labeling rate (right) across coding levels L2, L3, and L4, for the nine target groups with non-missing values for both metrics at every level. Each line traces one group's trajectory across levels; color denotes target category (race, religion, lgbtq, gender, politics) and marker shape distinguishes individual groups within a category. Labels are placed at each line's L4 endpoint.

benchmark label, and dogwhistle term were hidden to prevent anchoring. Annotators were kept blind to our hypothesis regarding FPR and coding level until after annotation was complete.

Inter-annotator agreement was moderate overall (Cohen's $\kappa = 0.437$, three-way; $\kappa = 0.521$ excluding Unclear items, $N = 194$; percent agreement 69.8%). Per-level κ decreases monotonically from L2 to L4 (0.510, 0.454, 0.349), consistent with our taxonomy's prediction that higher-level terms are harder to disambiguate from non-hateful uses.

FPR was computed on resolved judgments. Items where both annotators agreed on a definitive label retain that label (152 of 225). Items where exactly one annotator marked Unclear adopt the other annotator's definitive label (26): one annotator had sufficient context to judge, and treating uncertainty as abstention would discard real signal. Items where both annotators marked Unclear (5) or gave opposing definitive judgments (42) are excluded, the latter because no principled resolution exists without adjudication. This yields N = 65, 58, and 55 per level; FPR is the proportion of resolved Not Hateful judgments, with Wilson 95% confidence intervals reported in Table 2 in § 8. Excluding opposing judgments removes the most contested items from the denominator; since disagreement rises with coding level (per-level κ decreases), higher-level FPRs are estimated on progressively more consensual subsets.

FPR does not increase monotonically across levels; L3 (0.362) exceeds L4 (0.345). Additionally, confidence intervals overlap substantially, precluding level-to-level comparisons. Unclear rates decrease with coding level (L2: 12.0%; L3: 10.7%; L4: 9.3%), suggesting that higher-level terms tend to appear in contexts that are easier to classify as clearly non-hateful (high ambient frequency in political discourse) rather than genuinely ambiguous.

The annotator-vs-benchmark agreement finding

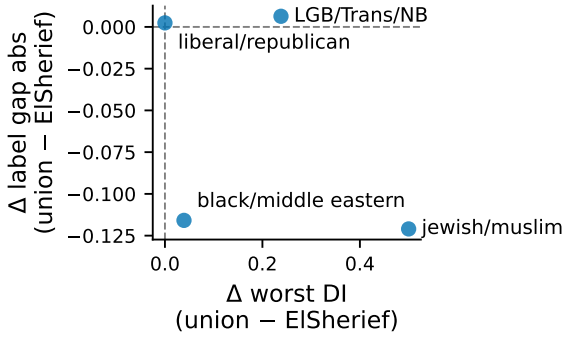


Figure 13: Change in worst pairwise coverage DI ratio and annotation label gap when expanding from the Implicit Hate corpus alone to the full benchmark union (Δ = union − Implicit Hate), for the four group pairs with stable estimates under both conditions. Points in the lower-right quadrant indicate pairs for which the union improves both metrics (higher worst DI, smaller absolute label gap); points in the upper-right indicate pairs for which the union improves coverage DI but widens the label gap.

provides independent validation of our RQ2 annotation failure result. Human annotators and benchmark labels agreed on only 46.2% of L2 matched posts, rising to 63.8% at L3 and 69.1% at L4 (Figure 16). The low L2 agreement is driven by posts that annotators judged *Hateful* but the benchmark labeled non-hateful (the benchmark FP category in the figure) consistent with Case B annotation failures in the primary audit and providing human-judgment evidence that benchmark mislabeling is not an artifact of surface-form matching.

Both annotators independently noted that their political orientation influenced their judgments, particularly for L3 concept and policy dogwhistles. This pattern is consistent with the design of L3 dogwhistles, which are specifically constructed to be interpretable as neutral policy language to an outgroup audience while conveying targeted meaning to an ingroup. The moderate IAA at L3 ($\kappa = 0.454$) may therefore reflect genuine semantic ambiguity rather than annotation noise, and the FPR estimates at L3 and L4 should be interpreted as observer-dependent lower bounds rather than fixed properties of the terms themselves.

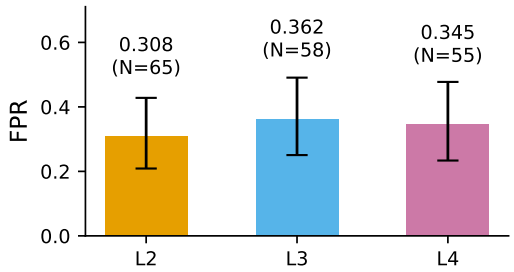


Figure 14: False positive rate (FPR) by coding level, with Wilson 95% confidence intervals. FPR is the proportion of resolved Not Hateful judgments per level; resolution rules and exclusion accounting in Appendix H).

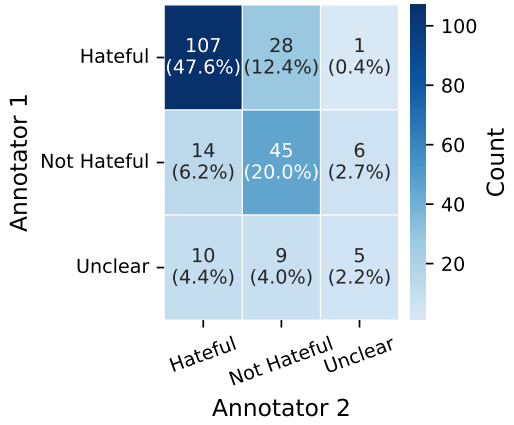


Figure 15: Inter-annotator agreement matrix for the FPR annotation task, pooled across all 225 sampled posts (75 per coding level). Cells show the count and percentage of posts receiving each combination of judgments from Annotator 1 and Annotator 2.

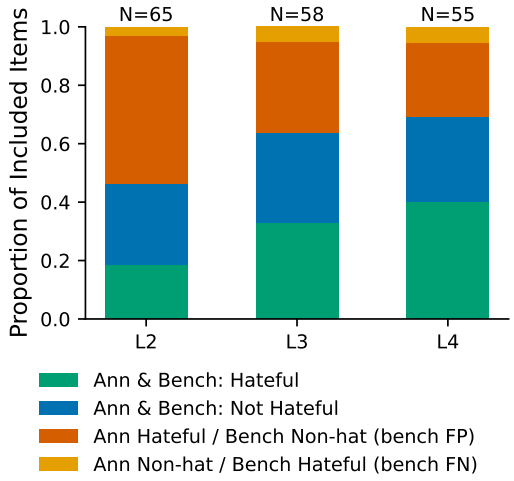


Figure 16: Agreement between annotator judgment and benchmark label by coding level, among items with a resolved annotator judgment (resolution rules in Appendix H).