

Counterfactual Robustness and the Calibration Trap in Antisemitism Detection

Giacomo Case, Yasmin Chaqor and Leonardo Cartesegna

EPFL, Lausanne, Switzerland

{giacomo.case, yasmin.chaqor, leonardo.cartesegna}@epfl.ch

Abstract

Classifiers for antisemitism detection inherit a shortcut from keyword-sampled training data: predictions flip when a Jewish identity term is rewritten to another religion, even though the gold label is preserved, conflating antisemitism with legitimate criticism of Israel. We audit this fragility jointly along three axes (positive-class performance, calibration, counterfactual stability) and report a previously unquantified tension. While four group-conditional post-hoc calibrators all improve global calibration (ECE), they inflate the soft counterfactual flip rate by 15–80%. A formal proposition shows intercept-free rescaling cannot change hard flips. *Orbit-constrained calibration* resolves it: it penalises calibrated probability differences on development-set identity-swap pairs, recovering the calibration gain (ECE -46%) with the hard rate unchanged; it is the only calibrator we test that lowers ECE without inflating soft flips. Separately, we attack invariance at training time: CCI (Counterfactual Consistency Invariance) adds a Jensen–Shannon consistency penalty between each tweet and its religiously rewritten counterfactuals. This cuts the hard flip rate by 49–73% across three encoder backbones and two model families, while preserving or improving F_1^+ (up to 0.72). Plain counterfactual augmentation reaches a lower flip rate at $3\times$ the training-set size; raising CCI’s multiplier closes that gap at a matched predicted-positive rate. Three prompted 7–8B LLMs cap at F_1^+ 0.45 and flip an order of magnitude more often. Finally, Integrated Gradients show identity-token attribution nearly unchanged: the gain comes from coupling pair scores, not from suppressing identity salience.

Content warning: this paper contains masked antisemitic slurs and offensive content.

1 Introduction

Antisemitic abuse online causes documented harm to Jewish communities, and platform moderation of it fails in both directions: hateful content survives while legitimate political speech is removed. The boundary is hard. Under the International Holocaust Remembrance Alliance (IHRA) Working Definition ([International Holocaust Remembrance Alliance, 2016](#)), criticism of Israeli policy “similar to that levelled against any other country” is not antisemitic, while antisemitism may surface through coded tropes rather than identity terms ([Becker et al., 2025](#)).

Supervised detectors make this boundary worse. Corpora collected through Jewish-related keywords inherit identity-token shortcuts: in GOLDSTANDARD2024 ([Jikeli et al., 2024](#)), positive base rates range from 11.4% on *Jews* to 88.3% on *Z*oNazi*. Empirical risk minimisation fits these keyword marginals rather than the IHRA boundary, so predictions flip under religious-identity swaps (*Jew* $\rightarrow$ *Muslim*) that preserve the label, the operational signature of a model that reads identities instead of content.

Despite a decade of identity-term bias work in hate speech detection, antisemitism detection has drawn relatively little attention: the one prior WOA treatment ([Ron et al., 2023](#)) contributes an annotation framework and an anti-Jewish hate corpus, but does not study detection robustness or the calibration–fairness interaction we audit here. We treat it as a stress test for counterfactual-fairness methods, on a target whose criticism-vs-hate boundary is contested by design. Concretely, we probe the religious-identity axis of that boundary (does renaming the targeted group flip the verdict?); its geopolitical axis (criticism of Israel motivates the problem but stays outside our swap operator (§3)).

Contributions.¹

- *A calibration–invariance tension, and a post-hoc calibrator that resolves it.* Across four group-conditional calibrators (temperature, Platt, isotonic, HKRR multicalibration (Hébert-Johnson et al., 2018)), every one that lowers Expected Calibration Error (ECE) inflates the soft counterfactual flip rate. A formal proposition shows that intercept-free rescaling cannot move hard flips at the symmetric threshold. We then introduce *orbit-constrained calibration*, which penalises calibrated probability differences on development-set counterfactual swap pairs. This is the one calibrator we test that escapes the trade-off. To our knowledge this is the first quantification of the trade-off, and the first post-hoc fix, for abuse detection.
- *A three-axis audit of antisemitism detection* (positive-class F_1 , F_1^+ ; ECE; counterfactual flip rates) with Holm-corrected significance over six seeds. This includes a null result: group error-rate parity (false-negative equality difference, FNED) neither improves nor degrades.
- *CCI (Counterfactual Consistency Invariance)*, a training-time Jensen–Shannon consistency loss over religiously symmetric identity swaps. Where Counterfactual Logit Pairing (CLP) (Garg et al., 2019) and its extension to hate speech (Davani et al., 2021) penalise an unbounded L_p logit gap, our JS penalty is a bounded surrogate that Pinsker’s inequality ties to the hard flip rate. Its gradient vanishes on already-consistent pairs, rather than pressuring confident pairs whose probabilities already agree. CCI cuts the hard counterfactual flip rate (CFR_{hard}) by 49%/58%/73% across DeBERTa-v3-base/large and RoBERTa-large at F_1^+ up to 0.72, transferring across tokeniser and model families. Counterfactual augmentation reaches a lower flip rate at 3× the training-set size; CCI attains that regime as a drop-in loss, closes the remaining gap when its multiplier is raised (App. H), and flips less than augmentation at a quarter of the labels (App. G).

¹Code, configurations and the pair manifest: [10.5281/zenodo.22699004](https://zenodo.org/record/22699004). No tweet text is redistributed; see the Ethics Statement.

2 Related Work

Identity-term shortcuts and counterfactual pairing. Dixon et al. (2018) introduced FPED/FNED and documented identity-term shortcuts as the dominant weakness of hate-speech corpora. Keyword balancing (Jikeli et al., 2023) acts at the sampling level and does not constrain the learned function on individual pairs. Garg et al. (2019) penalise the L_p logit difference between a sentence and its identity swap; Davani et al. (2021) extend this to hate speech. Both operate in unbounded logit space and have been validated within a single backbone family.

Counterfactual evaluation and training. Counterfactual invariance in abuse detection has mostly been treated as an evaluation problem. HateCheck (Röttger et al., 2021) probes identity-swap behaviour with templated functional tests, and Fryer et al. (2022) generate flexible counterfactuals for fairness probing; both diagnose models but leave training untouched. Counterfactual fairness (Kusner et al., 2017) formalises invariance over a causal graph, rarely available for noisy social-media text; our operator instead operationalises the causal intervention via identity-term rewriting. Our flip rate is the thresholded, paired analogue of perturbation sensitivity analysis (Prabhakaran et al., 2019). Closest to our approach is counterfactual data augmentation (CDA) (Lu et al., 2020; Zmigrod et al., 2019; Kaushik et al., 2020), which trains on rewritten pairs without an explicit consistency term. CCI extends this paradigm by coupling the rewrites with a Jensen–Shannon penalty. Recent work in abuse detection probes adjacent axes without coupling them to calibration: counterfactual token fairness for Dutch detection (Bauer et al., 2025), identity-swapped functional tests across Singapore languages (Ng et al., 2024), and concept-sufficiency diagnosis without a corresponding mitigation (Nejadgholi et al., 2023).

Calibration and multicalibration. Temperature scaling (Guo et al., 2017) is monotone and cannot move decisions across the symmetric threshold (§3). Multicalibration (Hébert-Johnson et al., 2018) guarantees subgroup calibration but, as we quantify, breaks individual invariance. The inherent conflict between calibration and fairness is well established: Pleiss et al. (2017) prove that perfect calibration and equalised group error

rates are jointly unachievable. That result concerns a different axis from ours (calibration vs. counterfactual invariance under identity swaps), and the tension we document is not an impossibility: an orbit-constrained map avoids it. [Fillies et al. \(2025\)](#) taxonomise mitigation methods by pipeline stage. While our orbit calibrator and CCI loss fit neatly into their post-processing and in-training categories, existing per-method schemas fail to capture the interaction between these stages. Similarly, recent audits of dataset robustness and representation ([Antypas and Camacho-Collados, 2023](#); [Korre et al., 2023](#); [Brate et al., 2024](#); [Tonneau et al., 2024](#); [Sun et al., 2025](#)) study these axes in isolation, our work exposes and resolves the tension between them.

3 Method

Setup. Let $(X, Y, S) \sim \mathcal{D}$, where X is a tweet, $Y \in \{0, 1\}$ its IHRA label, and S its dataset collection-keyword stratum; FNED is reported over the four values of S ; ECE is global unless stated. Calibration instead uses $G = g(X) \in \{\text{JEWS, ISRAEL, K*KEs, Z*ONAZI, NONE}\}$, the group our identity inventory infers from the text itself. The two need not agree, and a rewritten input is regrouped from its own text. A classifier f_θ produces logits $(z_0(X), z_1(X))$; we use the decision logit $z_\theta(X) = z_1(X) - z_0(X)$, posterior $p_\theta(X) = \sigma(z_\theta(X))$, and hard prediction $\hat{Y}_\theta(X) = \mathbb{1}\{p_\theta(X) \geq 1/2\}$.

Counterfactual operator. An operator Ψ generates a set of up to five rewrites per tweet, replacing a Jewish religious-ethnic token (*jew, jewish, hebrew, judaism*), while preserving case, with one of $\{\text{muslim, christian, hindu, buddhist, atheist}\}$. A pair (X, X') is admissible only if it survives three filters. First, no slur may be involved (*k*ke, z*onazi, chr*skiller*), since slur-bearing text changes meaning under substitution. Second, the mention must be religious-ethnic rather than a purely geopolitical proxy (*Israel, Zionist* with no religious anchor). Third, implicit conspiracy-only references (*globalist, Rothschild, Soros*) are excluded because they encode the target without naming it. The operator targets the religious-ethnic component of the shortcut: geopolitical-only swaps are excluded because their labels need not survive the rewrite under the IHRA boundary. The filter removes the religious-ethnic anchor, not every residual reference: 18% of admissible

pairs retain a geopolitical proxy the rewrite does not move, so label preservation on the remainder is an assumption rather than a guarantee. Under that label-preservation assumption, Ψ induces a pair distribution $\mathcal{Q}$; its test realisation $\mathcal{P}$ has 2,117 pairs generated from 611 of the 1,060 test tweets.

Objective. Intuitively, we train the standard classifier while requiring that admissible identity rewrites do not change its decisions (Figure 1). Formally,

$$\min_{\theta} \mathbb{E}_{\mathcal{D}}[\ell_{\text{focal}}(f_\theta(X), Y)] \quad \text{s.t.} \quad \text{CFR}_{\text{hard}}(\theta) \leq \varepsilon, \quad (1)$$

where ℓ_{focal} is the focal loss ([Lin et al., 2017](#)) (the corpus is 4.79:1 imbalanced) and $\text{CFR}_{\text{hard}}(\theta) = \mathbb{E}_{\mathcal{Q}} \mathbb{1}\{\hat{Y}_\theta(X) \neq \hat{Y}_\theta(X')\}$ is the hard flip rate, non-differentiable since $\hat{Y}_\theta$ thresholds a continuous score. The companion soft metric is $\text{CFR}_{\text{soft}}^{L_1}(\theta) = \mathbb{E}_{\mathcal{Q}} |p_\theta(X) - p_\theta(X')|$.

Smooth surrogate. Pinsker’s inequality applied to each KL summand of the binary Jensen–Shannon divergence gives $\frac{1}{2}|p - p'|^2 \leq \text{JS}_b(p, p')$, so JS_b upper-bounds pairwise disagreement outside a vanishing margin of the threshold. Relaxing (1) with a fixed multiplier $\lambda=1$, held at one value across every backbone and cell rather than tuned per condition (App. H), yields the training loss

$$\begin{aligned} \hat{\mathcal{L}}_B(\theta) = & \frac{1}{|B|} \sum_{i \in B} \ell_{\text{focal}}(f_\theta(X_i), Y_i) \\ & + \frac{\lambda}{N_B} \sum_{i \in B} \sum_{k=1}^{K_i} \text{JS}_b(p_\theta(X_i), p_\theta(X'_{ik})). \end{aligned}$$

Here B is a minibatch of training tweets, $K_i \leq 5$ (zero for tweets with no admissible rewrite) is the number of admissible counterfactuals for X_i , and $N_B = \sum_{i \in B} K_i$; the consistency term is zero when $N_B=0$. Valid pairs, not tweets, carry equal weight. Operationally, the JS_b term is a smooth, bounded stand-in for the flip rate: its gradient vanishes on pairs the model already treats consistently and pulls disagreeing pairs toward their Bernoulli mean. An L_1 divergence and a learned temperature head are ablated in App. D.

Proposition 1 (Sign-preservation of temperature scaling). *For any $T > 0$, $\mathbb{1}\{\sigma(z/T) \geq 1/2\} = \mathbb{1}\{z \geq 0\}$; hence hard predictions at the symmetric threshold are pointwise T -invariant and $\text{CFR}_{\text{hard}}^T(\theta) \equiv \text{CFR}_{\text{hard}}(\theta)$.*

Proposition 1 rules out every intercept-free monotone rescaling of the decision logit at the symmetric threshold; a post-hoc remedy must move the

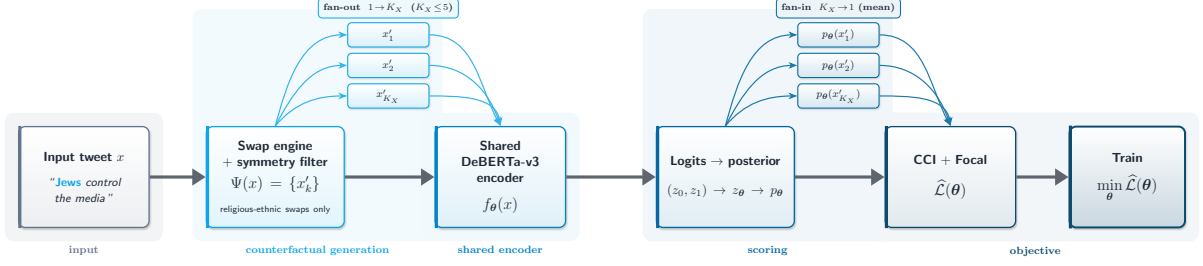


Figure 1: CCI pipeline. A shared encoder scores a tweet x and its religiously rewritten counterfactuals x'_k ; the training loss couples their predictions with a Jensen–Shannon penalty.

decision boundary itself (an affine or per-cell correction). We show below that the natural per-cell candidate, multicalibration, trades this invariance for soft-flip inflation, so a training-time relaxation is required. Training details are in Appendix B.

4 Validation

Data and protocol. GOLDSTANDARD2024 (Jikeli et al., 2024) (CC BY 4.0): 11,311 English tweets keyword-sampled around $\{\text{Jews}, \text{Israel}, K^* \text{kes}, Z^* \text{oNazi}\}$, annotated under the IHRA definition by the dataset’s curators. We use a stratified 80/10/10 split (8,479/1,060/1,060; 187 test positives) of the 10,599 tweets remaining after removing duplicate texts, which the corpus contains and which would otherwise leak across splits, and report F_1^+ , 15-bin ECE, CFR_{hard} and $\text{CFR}_{\text{soft}}^{L_1}$ over $\mathcal{P}$, and the false-negative equality difference (FNED) (Dixon et al., 2018). Adapting paired bootstrap resampling (Dror et al., 2018) to seed-level test-set statistics, each contrast resamples the six training seeds with replacement (10^4 resamples; the unit is the seed-level test-set scalar, not individual examples or pairs), with Holm–Bonferroni correction applied separately within two pre-specified families at family-wise $\alpha=0.05$: 54 pivot-vs-comparator tests and 6 within-architecture transfer tests. Because the 10^4 resamples floor the bootstrap p -value at 10^{-4} when all six seeds agree in direction, we also report an exact two-sided sign test over seeds (minimum $p = 2 \cdot 2^{-6} \approx 0.031$, uncorrected) as a direction-consistency check. All cells aggregate six seeds (42–47).

Main result. Table 1 reports a CFR_{hard} reduction of 49% on DeBERTa-v3-base, 58% on DeBERTa-v3-large and 73% on RoBERTa-large versus the same-architecture focal-mask baseline (adjacent C rows). Holm-corrected paired boot-

strap rejects equality on every counterfactual contrast, both CFR metrics against all four base-architecture baselines and in both large-backbone cells ($p_{\text{boot}}=10^{-4}$ for these comparisons). An exact seed-level sign test agrees in every cell ($p=0.031$).

Error-rate parity. At fixed backbone the FNED difference is within seed noise in every same-architecture contrast after Holm correction (Model C, its temperature-scaled variant, Davani $\lambda=0.5$, and post-hoc multicalibration); the largest gap, a nominal improvement over multicalibration whose per-cell correction widens group error-rate disparity, does not survive correction (uncorrected $p=.03$). F_1^+ is preserved or improved at every cell.

Against augmentation. Counterfactual augmentation (CDA, the same admissible swaps as hard-labelled rows) reaches a lower hard flip rate ($\text{CFR}_{\text{hard}} .012$ at $F_1^+ .662$): what cuts flips is training-time exposure to the swap set, as data or as the consistency penalty, and CCI attains it as a drop-in loss without the $3\times$ training-set expansion CDA requires. That margin closes when the multiplier is raised: at $\lambda=5$ CCI reaches $\text{CFR}_{\text{hard}} .009$ at $F_1^+ .666$ and a predicted-positive rate of .147 against CDA’s .154 (paired $\text{CFR}_{\text{hard}} -.003$, $p=.11$; $F_1^+ +.004$, $p=.59$; App. H). A keyword-masking factorial (App. E) shows that the $\lambda=1$ margin is not a masking artefact and that CCI’s reduction survives matched masking: the C-to-CCI CFR_{hard} reduction is 50% with masking on and 55% with it off, CDA keeps its $\approx 33\%$ edge in both columns, and masking itself moves CFR_{hard} by $\leq .004$ while lowering F_1^+ (most for CCI, whose swaps operate on the identity tokens masking removes).

Portability. The largest reduction crosses tokeniser families (SentencePiece vs. BPE) and model families. CCI generates its rewrites in text space and applies the penalty to predicted proba-

Method	$F_1^+ \uparrow$	ECE $\downarrow$	CFR $_{\text{hard}} \downarrow$	CFR $_{\text{soft}}^{L_1} \downarrow$	FNED $\downarrow$
C (DeBERTa-base)	.660 $\pm$.019	.077 $\pm$.034	.036 $\pm$.008	.032 $\pm$.006	.797 $\pm$.159
+ TempScaling	.660 $\pm$.019	.045 $\pm$.014	.036 $\pm$.008	.033 $\pm$.009	.797 $\pm$.159
+ Multicalibration	.646 $\pm$.022	.035 $\pm$.005	.033 $\pm$.006	.057 $\pm$.013	.957 $\pm$.064
Davani CLP ($\lambda=0.5$)	.672 $\pm$.015	.094 $\pm$.033	.040 $\pm$.005	.036 $\pm$.005	.786 $\pm$.036
CDA (augment-only)	.662 $\pm$.012	.070 $\pm$.028	.012 $\pm$.004	.013 $\pm$.001	.775 $\pm$.105
CCI (DeBERTa-base)	.677 $\pm$.008	.098 $\pm$.053	.018 $\pm$.003	.014 $\pm$.002	.793 $\pm$.180
C (DeBERTa-large)	.704 $\pm$.020	.097 $\pm$.053	.042 $\pm$.003	.037 $\pm$.009	.612 $\pm$.051
CCI (DeBERTa-large)	.721 $\pm$.012	.096 $\pm$.057	.017 $\pm$.009	.016 $\pm$.003	.655 $\pm$.056
C (RoBERTa-large)	.689 $\pm$.016	.068 $\pm$.044	.047 $\pm$.002	.048 $\pm$.001	.708 $\pm$.064
CCI (RoBERTa-large)	.694 $\pm$.005	.077 $\pm$.052	.013 $\pm$.006	.013 $\pm$.002	.759 $\pm$.115
LLM Llama-3.1-8B, Guided-CoT	.450	–	–	–	–
LLM Mistral-7B, best prompt	.417	–	.168 †	–	–
LLM Qwen-2.5-7B, best prompt	.373	–	.238	–	–

Table 1: Test metrics on GOLDSTANDARD2024 ($n=1,060$); CFR $_{\text{hard}}$ and CFR $_{\text{soft}}^{L_1}$ over $\mathcal{P}$ ($|\mathcal{P}|=2,117$). Encoder cells are mean $\pm$ std over six seeds; Model C is the focal-loss baseline with keyword masking: during training only, and with probability 0.3, the eight collection keywords (*jews, jewish, israel, israeli, zionist, zionazi, kike, kikes*) are replaced by the tokeniser’s mask token. It is named Model C throughout. LLM F_1^+ is over parseable responses from each model’s best prompt, whereas the reported LLM CFR $_{\text{hard}}$ values are zero-shot († Mistral CFR $_{\text{hard}}$ over its 69% parsed pairs; [.116, .427] full-denominator); the LLM dashes need probabilities the models do not expose. Coverage in App. C. Bold = per-column winner.

Calibrator	ECE $\downarrow$	CFR $_{\text{soft}}^{L_1}$ infl.	CFR $_{\text{hard}}$
(uncalibrated)	.077	–	.0358
Per-group temperature	.044	+47%	.0358
Per-group Platt	.030	+15%	.0347
Per-group isotonic	.030	+48%	.0334
Multicalibration (patching)	.035	+80%	.0325
Orbit (ours)	.041	–0%	.0358

Table 2: Post-hoc calibration of Model C (DeBERTa-v3-base, six seeds), fitted on dev and evaluated over $\mathcal{P}$. Every unconstrained group-conditional calibrator improves ECE but inflates the soft counterfactual flip rate, and the intercept-bearing ones also perturb CFR $_{\text{hard}}$; only the orbit-constrained temperature avoids both.

bilities, so the same procedure runs unchanged on either backbone; the observed reductions demonstrate that portability rather than invariance to tokenisation or to logit scaling, which the Jensen–Shannon penalty does not have.

Mechanism. Integrated Gradients (Sundarajan et al., 2017) show the median fraction of attribution mass on identity tokens is nearly unchanged ($0.046 \rightarrow 0.044$; paired Wilcoxon over 500 test examples, Holm-corrected $p < 10^{-2}$, $d_z=0.11$), so the gain does not come from suppressing identity-token salience. The JS $_b$ term instead couples each pair’s scores toward their Bernoulli mean, visible directly as the soft flip rate falling from .032 to .014 (Table 1).

The calibration–invariance tension. Group-conditional post-hoc calibration improves global ECE but pulls counterfactual pairs apart: a tweet

(group g) and its swap (group g') pass through different per-group corrections, so their probabilities diverge even when their hard labels agree. Table 2 quantifies this on Model C across four calibrators fitted on dev. All four cut ECE (from .077 to .030–.044) but inflate CFR $_{\text{soft}}^{L_1}$ by 15–80%; the intercept-bearing maps (Platt, isotonic, multicalibration) additionally perturb CFR $_{\text{hard}}$, while intercept-free per-group temperature leaves it bit-exact (Proposition 1) yet still inflates the soft rate by 47%. Calibration and counterfactual invariance are in tension for all four standard calibrators. Multicalibration, the one that most directly targets subgroups, implemented as a full-sweep variant of HKRR patching that corrects every violating group-bin cell per pass rather than the single worst, leaves worst-group ECE over the four keyword groups unchanged (.221, equal to the uncalibrated base).

Orbit-constrained calibration restores the frontier. The tension is not intrinsic to calibration; it is a property of correcting each group independently. We add an orbit penalty to per-group temperature: fit $\{T_g\}$ on dev to minimise NLL plus β times the mean absolute calibrated probability difference over dev swap pairs. This penalises pair-score disagreement without constraining the group temperatures to be equal. Being intercept-free it preserves CFR $_{\text{hard}}$ bit-exactly for any β . β is selected on dev; at $\beta=0$ the penalty is absent, though that member is not identical to the sepa-

rate per-group temperature baseline, whose optimisation bounds and small-group handling differ. At inference it applies $T_{g(x)}$ by the input’s own group, needing no test-time swaps. It cuts ECE by 46% (.077 $\rightarrow$.041, below unconstrained per-group temperature’s .044) with zero mean soft-flip inflation (Table 2): the only calibrator we test that lowers ECE without trading away counterfactual invariance. CCI itself is a consistency loss, not a calibrator (its ECE matches or trails its baseline); the two are complementary, training-time invariance plus orbit-aware post-hoc calibration.

LLM and out-of-distribution audit. The three $\leq 8B$ instruction-tuned models cap at F_1^+ of 0.450, 23–27 points below the CCI cells, and flip an order of magnitude more often (zero-shot $\text{CFR}_{\text{hard}}=.238$ Qwen at full coverage, .168 Mistral over parseable pairs, versus .013–.018 for CCI): scale and instruction-tuning alone do not confer counterfactual stability. On the Jewish-target subsets of HateXplain (Mathew et al., 2021) and ToxiGen (Hartvigsen et al., 2022), baseline and CCI rank comparably (PR-AUC differences within $[-0.007, +0.021]$). The apparent fixed-cutoff gap is a prevalence-shift artefact (mean oracle thresholds for both fall in $[0.06, 0.18]$), not a ranking deficit.

5 Conclusion

On a contested, under-studied target, naive group-conditional calibration improves calibration only by breaking counterfactual invariance. An orbit-constrained map recovers the calibration gain without that cost. The CCI consistency loss cuts hard flips by 49% to 73% across three backbones at equal or higher F_1^+ , as a drop-in term without the $3\times$ training-set expansion augmentation requires; it closes augmentation’s remaining flip-rate margin when its multiplier is raised, and flips less than augmentation at a quarter of the labels. Counterfactual robustness and subgroup calibration need not conflict, provided the post-hoc map respects the swap orbits. One null result bounds the scope: group error-rate parity (FNED) is neither improved nor degraded.

Limitations

The swap vocabulary covers explicit identity terms only; coded antisemitism (*Rothschild, globalist*) is tracked but not swapped, and the symmetry filter is heuristic. The operator also rewrites one occur-

rence of an identity term per pair, so 713 pairs keep a residual anchor to the source group; restricting to the rest moves the reported CFR_{hard} reduction by at most 4.1 points, below what a size-matched placebo null resolves at 80% power (App. F). Label preservation itself is not measured here, but if rewrites broke labels at any real rate, a loss that pulls each tweet toward its rewrites would pull F_1^+ down; instead F_1^+ rises in all three cells (.660 $\rightarrow$.677, .704 $\rightarrow$.721, .689 $\rightarrow$.694) and its seed spread falls (.019 $\rightarrow$.008, .020 $\rightarrow$.012, .016 $\rightarrow$.005), which is evidence against wholesale breakage rather than a measurement of its rate. All training data and every headline number come from a single English Twitter corpus under one annotation scheme, and the out-of-domain check covers two Jewish-target subsets only; the IHRA definition is contested, and alternatives such as the Jerusalem Declaration (Jerusalem Declaration on Antisemitism, 2021) draw the Israel-criticism boundary differently, so our labels, and what counts as a “preserved” label under a swap, inherit a normative choice. CCI does not improve group error-rate parity (FNED within seed noise) or calibration (ECE degrades or remains on par with the baseline quality). The LLM cells are single-run point estimates evaluated solely on parseable responses; their CFR_{hard} is measured zero-shot for two of the three models (Qwen at full coverage, Mistral over the 69% of pairs both sides parse), with the Llama pair sweep left to future work. Counterfactual augmentation (CDA) reaches a lower CFR_{hard} than CCI at $\lambda=1$ and full supervision (Table 1), so a consistency penalty is not the only route to robustness there; raising the multiplier to $\lambda=5$ matches CDA’s rate at a matched predicted-positive rate (App. H), without the $3\times$ training-set expansion CDA requires.

Ethics Statement

Dual use. A detector of antisemitism can be repurposed to suppress legitimate political speech; predictions that flip when a religious group is renamed represent one mechanism by which moderation silences protected criticism, and reducing these flips is our primary goal. We recommend deployment as a triage tool with human review for borderline scores, not autonomous action. The swap operator generates no abusive content and ships with the code; we publish no tweet text and no trained checkpoint, so that care falls to whoever

trains and deploys one.

Annotation framework. Our labels follow the IHRA Working Definition because it is the framework of GOLDSTANDARD2024; the methodological contribution is framework-independent (the loss enforces invariance with respect to whichever labels annotators provide), and practitioners adopting a different framework inherit that framework’s boundary judgments. We take no position in the definitional debate and disclose the dependency.

Data and content. Our use of GOLDSTANDARD2024 (Jikeli et al., 2024) is a secondary analysis of a public CC BY 4.0 release through its own distribution channel: the tweets were gathered without explicit author consent, usernames are masked to @USER and never modelled, the released splits carry tweet identifiers without the author column, and no additional ethics-board review was required. The corpus contains slurs, Holocaust-related content, conspiracy theories, and dehumanising language; we reproduce offensive terms only where scientifically necessary, mask slurs with asterisks, and flag content with a warning. Full provenance, annotation protocol, compute and the artefact release are documented in App. A.

References

- Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*.
- Julie Bauer, Rishabh Kaushal, Thales Bertaglia, and Adriana Iamnitchi. 2025. *Towards fairness assessment of Dutch hate speech detection*. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*.
- Matthias J. Becker, Jordan Blatter, and Oksana Stanevich. 2025. *Decoding antisemitism online: linguistic and multimodal challenges in the age of AI*. *Frontiers in Communication*, 10.
- Ryan Brate, Marieke van Erp, and Antal van den Bosch. 2024. A Bayesian quantification of aporophobia and the aggravating effect of low-wealth contexts on stigmatization. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*.
- A. M. Davani, A. Omrani, B. Kennedy, M. Atari, X. Ren, and M. Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*.
- L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of AIES*.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of ACL*.
- Jan Fillies, Michael Wawerek, and Adrian Paschke. 2025. *A comprehensive taxonomy of bias mitigation methods for hate speech detection*. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*.
- Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229.
- S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of AIES*.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*.
- T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. 2022. ToxiGen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of ACL*.
- Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. 2018. Multicalibration: calibration for the (computationally-identifiable) masses. In *Proceedings of ICML*.
- International Holocaust Remembrance Alliance. 2016. Working definition of antisemitism.
- Jerusalem Declaration on Antisemitism. 2021. The Jerusalem declaration on antisemitism. <https://jerusalemdeclaration.org/>.
- G. Jikeli, S. Karali, D. Miehl, and K. Soemer. 2023. Antisemitic messages? a guide to high-quality annotation and a labeled dataset of tweets. *arXiv:2304.14599*.
- G. Jikeli, S. Karali, D. Miehl, and K. Soemer. 2024. Antisemitism on Twitter: a dataset for machine learning. *Zenodo v3*, doi:10.5281/zenodo.14448399.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.
- Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-Cedeño. 2023. Harmful language datasets: An assessment of robustness. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*.

- Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*.
- Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer.
- B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. 2021. HateXplain: a benchmark dataset for explainable hate speech detection. In *Proceedings of AAAI*.
- Isar Nejadgholi, Svetlana Kiritchenko, Kathleen C. Fraser, and Esma Balkır. 2023. [Concept-based explanations to test for false causal relationships learned by abusive language classifiers](#). In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*.
- Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*.
- Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of EMNLP-IJCNLP*, pages 5740–5745.
- Gal Ron, Effi Levi, Odelia Oshri, and Shaul R. Shenhav. 2023. [Factoring hate speech: A new annotation framework to study hate speech in social media](#). In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*. 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 ACL*, pages 41–58.
- Ling Sun, Soyoung Kim, Xiao Dong, and Sandra Kübler. 2025. [Debiasing static embeddings for hate speech detection](#). In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*.
- M. Sundararajan, A. Taly, and Q. Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of ICML*.
- Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*.
- Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of ACL*, pages 1651–1661.

A Data statement

Corpus. GOLDSTANDARD2024 (Jikeli et al., 2024) contains 11,311 English tweets posted between January 2019 and April 2023, curated at Indiana University by random keyword sampling (*Jews, Israel*, and the slurs *Z*oNazi, K*kes*) from a 10% public-tweet archive (Jikeli et al., 2023). The release contains tweet IDs, usernames, and full text, and includes slurs and Holocaust-related content.

Annotation. Each tweet was labelled by two trained annotators who viewed it in live context and applied the IHRA definition. Annotators had prior academic exposure to anti-semitism and spanned different religions and genders. Pre-adjudication agreement was Cohen’s $\kappa=0.66$, measured on the predecessor corpus; disagreements were discussed and unresolved tweets removed (Jikeli et al., 2023). Counterspeech that quotes antisemitism is deliberately labelled non-antisemitic, which reduces but does not remove the risk that trained models penalise benign mentions of Jewish identity.

Models and compute. The backbones are DeBERTa-v3-base (184M parameters), DeBERTa-v3-large (435M), and RoBERTa-large (355M). Each training run used a single NVIDIA A100 GPU; sweeps dispatched independent runs concurrently across a shared cluster pool.

Implementation. Models, tokenisers and generation use HuggingFace transformers (≥ 4.36) on PyTorch (≥ 2.1); the encoder checkpoints are microsoft/deberta-v3-base, microsoft/deberta-v3-large and roberta-large. The prompted models are meta-llama/Llama-3.1-8B-Instruct, mistralai/Mistral-7B-Instruct-v0.3 and Qwen/Qwen2.5-7B-Instruct, loaded in 4-bit via bitsandbytes and decoded greedily. Classification metrics use scikit-learn (≥ 1.3), the paired

bootstrap and sign tests use `scipy` (≥ 1.11), and Integrated Gradients uses `captum` (≥ 0.7).

Release. We release code, not data: the code, the configurations, the per-seed rollups behind every table, a per-split index of tweet ids, and a manifest of the 2,117 pairs with each substitution. No tweet text is redistributed; the corpus is fetched from its own archive and the code rebuilds the splits and the pairs exactly.

B Training details

AdamW, learning rate 2×10^{-5} (base) / 1×10^{-5} (large), weight decay 0.01 (zero on biases and LayerNorm), 10% linear warm-up, effective batch 16, gradient clipping at 1, sequence length 128, focal $\gamma=2$, $\alpha=0.75$, fp32, patience-3 early stopping on dev macro- F_1 , six seeds per cell. The swap operator draws three of the five alternative groups per identity occurrence under a fixed seed, caps rewrites at five per tweet, and substitutes the number- and form-matched surface form (*judaism* to *islam*, *jews* to *muslims*).

C LLM evaluation details

Model	Prompt	F_1^+	Valid (%)
Llama-3.1-8B-Instruct	zero-shot	0.441	99.5
	few-shot	0.443	99.4
	Guided CoT	0.450	83.9
Mistral-7B-Instruct-v0.3	zero-shot	0.296	100.0
	few-shot	0.417	61.4
	Guided CoT	0.240	10.3
Qwen2.5-7B-Instruct	zero-shot	0.335	100.0
	few-shot	0.373	100.0
	Guided CoT	0.298	49.5

Table 3: Prompted-LLM results on the 1,060-tweet test set (greedy decoding); bold marks the best prompt per model. F_1^+ is over parseable responses; *Valid* is the fraction the rule-based parser mapped to a label. Low-validity cells (e.g. Mistral CoT, 10.3%) score a non-random subset and are not directly comparable.

D Extended results

Method	F_1^+	ECE	CFR_{hard}	$\text{CFR}_{\text{soft}}^{L_1}$
Davani $\lambda=1.0$	.660 $\pm$.017	.127 $\pm$.047	.034 $\pm$.011	.029 $\pm$.010
Davani $\lambda=2.0$	.667 $\pm$.019	.177 $\pm$.035	.042 $\pm$.009	.026 $\pm$.006
Davani $\lambda=1.0$ (no filt.)	.661 $\pm$.025	.162 $\pm$.055	.041 $\pm$.010	.028 $\pm$.004

Table 4: Davani CLP comparators (DeBERTa-v3-base), six seeds. “No filt.” disables the swap-symmetry filter, which raises CFR_{hard} from .034 to .041, so the filter removes spurious swaps at no cost in stability; none of these rows reaches CCI’s .018 (Table 1).

T_g	Div.	F_1^+	ECE	CFR_{hard}	$\text{CFR}_{\text{soft}}^{L_1}$
fixed	L_1	.651 $\pm$.012	.200 $\pm$.019	.007 $\pm$.002	.003 $\pm$.001
fixed	JS	.677 $\pm$.008	.098 $\pm$.053	.018 $\pm$.003	.014 $\pm$.002
learned	L_1	.647 $\pm$.010	.186 $\pm$.012	.008 $\pm$.002	.004 $\pm$.000
learned	JS	.662 $\pm$.015	.083 $\pm$.048	.015 $\pm$.008	.013 $\pm$.003

Table 5: CCI ablation (DeBERTa-v3-base): temperature head (fixed vs. learned) $\times$ divergence (L_1 vs. JS), six seeds. L_1 cuts CFR_{hard} more but costs F_1^+ and ECE; JS is the better trade-off. Fixed and learned T tie on CFR_{hard} ($p=0.25$), so we pivot on fixed- T (row 2).

Seed	42	43	44 [†]	45 [†]	46	47
Model C ECE	.063	.054	.129	.109	.067	.040
Model C recall	.647	.551	.722	.636	.572	.626
Model C F_1 epoch	7	5	3	3	5	4
CCI ECE	.049	.174	.078	.149	.046	.092
CCI recall	.631	.733	.668	.631	.610	.583

Table 6: Per-seed 15-bin ECE and recall for Model C and CCI (DeBERTa-v3-base); each method’s two largest ECE values are bold. Both methods show substantial calibration variation across seeds, and each one’s worst-calibrated seed is also its highest-recall seed. Checkpoints were selected by development macro- F_1 . Six seeds cannot establish bimodality or attribute the spread to the stopping rule.

Seed	Fitted T	CFR_{hard} (raw)	CFR_{hard} (TS)
42	1.400	.0321	.0321
43	1.244	.0293	.0293
44	0.602	.0444	.0444
45	0.566	.0463	.0463
46	1.388	.0255	.0255
47	0.985	.0373	.0373

Table 7: Empirical face of Proposition 1: post-hoc temperature scaling (with T fitted on dev, range 0.57–1.40) preserves CFR_{hard} bit-exactly on all six Model C seeds, since monotone rescaling cannot move any score across the symmetric threshold.

E Keyword-masking factorial

Metric	Method	mask off	mask on
CFR_{hard}	focal	.040	.036
	CDA	.012	.012
	CCI	.018	.018
F_1^+	focal	.667	.660
	CDA	.673	.662
	CCI	.677	.647

Table 8: Keyword-masking factorial (DeBERTa-v3-base, six seeds, test split). CCI’s CFR_{hard} reduction over Model C survives matched masking (50% with masking on, 55% off); CDA holds the lowest CFR_{hard} in both columns, so its margin at $\lambda=1$ is not a masking artefact. Masking itself moves CFR_{hard} by at most .004 and lowers F_1^+ , most for CCI, since it removes the identity tokens the swaps act on. Table 1 publishes C and CDA with masking on and CCI with it off.

F Validity of the counterfactual manifest

Placebo slices. The operator rewrites one occurrence of an identity term per pair, so 713 of the $|\mathcal{P}|=2,117$ pairs keep a residual anchor to the source group. For a slice of k pairs we recompute the CFR_{hard} reduction and compare it against 2,000 random slices of the same size, rejecting two-sided when $|s - \bar{s}_0|$ exceeds the 95th percentile of $|s_0 - \bar{s}_0|$. NO-JEWISH ($k=1,404$) retains no Jewish identity term and NO-ANCHOR ($k=1,137$) also drops geopolitical and coded terms; *shorter* ($k=1,416$, at or below the 66.5th percentile of length) is a size-matched sham, defined by a criterion unrelated to the defect. MDE_{80} is the smallest location shift of the reference distribution reaching 80% rejection: it measures sensitivity under that assumption, and is not an upper bound on an undetected effect.

Backbone	Slice	shift	MDE_{80}	det.
DeBERTa-base	NO-JEWISH	−2.6	7.7	no
	NO-ANCHOR	+0.6	10.2	no
	<i>shorter</i>	+10.1	7.5	yes
DeBERTa-large	NO-JEWISH	+4.1	6.7	no
	NO-ANCHOR	+2.9	8.8	no
	<i>shorter</i>	−1.0	6.6	no
RoBERTa-large	NO-JEWISH	−0.2	5.0	no
	NO-ANCHOR	−0.1	6.6	no
	<i>shorter</i>	−3.5	5.1	yes

Table 9: Swap-operator validity. “shift” is the slice reduction minus the size-matched placebo mean, in points of relative CFR_{hard} reduction; MDE_{80} is the smallest shift the placebo null resolves at 80% power; “det.” marks slices the null rejects. Substantive slices are small-caps, the sham slice italic. Six seeds per cell, 2,000 placebo replicates.

Reading. No substantive slice is distinguishable from random slices of its own size, with $|\text{shift}|$ at most 4.1 points against a sensitivity of 5.0 to 10.2; unrestricted, the protocol recovers 48.8% against the 49% of Table 1. On DeBERTa-v3-base and RoBERTa-large the sham slice is rejected at the same k , so the test does resolve shifts of that size there; DeBERTa-v3-large has no such control. The residual anchor therefore leaves no detectable trace at this sensitivity; this is a sensitivity statement, not an equivalence bound.

G Label efficiency: CCI against augmentation

Design. At labelled fraction f all three methods receive the same labelled subset; only CCI additionally enforces its consistency term on the unlabelled remainder, which augmentation cannot use because it needs a label to copy onto each rewritten row. Six seeds per cell, eighteen at $f=0.25$. Fraction 1.0 is the main run of Table 1, and the sweep reproduces it (CDA CFR_{hard} .0117 at F_1^+ .662), which serves as the self-check.

f	n	$\Delta\text{CFR}_{\text{hard}}$	p_{boot}	sign	Holm
0.10	6	−.0106	.0001	6/6	rejects
0.25	18	−.0046	.0066	15/18	rejects
0.50	6	−.0017	.3008	4/6	no
1.00	6	+.0067	.0002	1/6	—

Table 10: CCI minus CDA on CFR_{hard} , paired over seeds; negative means CCI flips less. Holm at family-wise $\alpha=0.05$ over the sub-1.0 fractions, which are the pre-registered family; $f=1.0$ is the published cell and sits outside it. p_{boot} cannot fall below the resampling floor of 10^{-4} , which the $f=0.10$ cell reaches.

Reading. At $f=0.25$ CCI cuts CFR_{hard} 33% below CDA, with the paired bootstrap and the sign test agreeing, which is the standard the cells of Table 1 meet. At $f=0.5$ the difference is small (a standardised effect of 0.42) and six seeds do not resolve it; its size implies roughly an order of magnitude more would be needed. At $f=0.1$ CCI’s F_1^+ is 6.7 points lower, so that cell is not F_1^+ -matched and its larger gap is not interpretable as invariance.

Prevalence check. A model predicting positive less often has fewer predictions available to flip, and at $f=0.25$ CCI’s F_1^+ is .0095 below CDA’s. Regressing the per-seed $\Delta\text{CFR}_{\text{hard}}$ on the per-seed ΔF_1^+ leaves an intercept of −.0043 at zero ΔF_1^+ (uncorrected bootstrap $p=.010$, 95% interval [−.0079, −.0011]), so the gap does not reduce to the F_1^+ difference. The direct test, a comparison at matched predicted-positive rates, is the one App. H applies at full supervision; this cell has the regression only.

H Consistency multiplier

Design. The multiplier is held at $\lambda=1$ throughout the paper: one value across three backbones, six seeds and every cell of Table 1, rather than a value tuned per condition. This appendix reports what the multiplier controls. Three further arms at $\lambda \in \{0.5, 2, 5\}$ on DeBERTa-v3-base are identical to the published configuration in every other respect (masking off, six seeds each, the same stopping rule); $\lambda=1$ is the published run itself.

Arm	F_1^+	CFR_{hard}	pos. rate
$\lambda=0.5$	.657±.008	.0207±.0061	.177
$\lambda=1$	.677±.008	.0183±.0028	.159
$\lambda=2$	.662±.011	.0144±.0052	.155
$\lambda=5$	.666±.014	.0089±.0034	.147
CDA	.662	.0117	.154

Table 11: CCI at four multipliers against CDA: DeBERTa-v3-base, test split, mean±std over six seeds per arm; masking off for CCI and on for CDA, as in Table 1. pos. rate is the predicted-positive rate on the 1,060 test tweets, listed because a flip rate lowered by predicting fewer positives would be a move along the operating curve rather than added invariance.

Reading. In Table 11, CFR_{hard} falls monotonically in λ , from .0207 to .0089, while F_1^+ moves by .020 across the whole range. At $\lambda=5$ CCI reaches CDA’s flip rate: paired over seeds the difference is −.0028 in CFR_{hard} (bootstrap $p=.11$) and +.0040 in F_1^+ ($p=.59$), neither resolvable at six seeds, and the predicted-positive rates are .147 against .154, so the arms are close to matched rather than separated by a shift in prevalence. The CFR_{hard} margin CDA holds at $\lambda=1$ (Table 10, $f=1.0$ row, +.0067) is a setting we can move, and moving it closes the margin with F_1^+ and predicted-positive rate matched to CDA’s (.666 against .662; .147 against .154).

We report this as an ablation and do not re-select λ on it: Table 1 is the $\lambda=1$ model throughout, and λ was fixed before any of these arms were trained.