Learning What Is Said, Not How: Stance-Contrastive Learning for Generalizable Hate Speech Detection

Yucheng Lin, Dan Kienast, Jiaojiao Jiang
UNSW Sydney
High St, Kensington NSW 2052, Australia
{yucheng.lin,d.kienast,jiaojiao.jiang}@unsw.edu.au

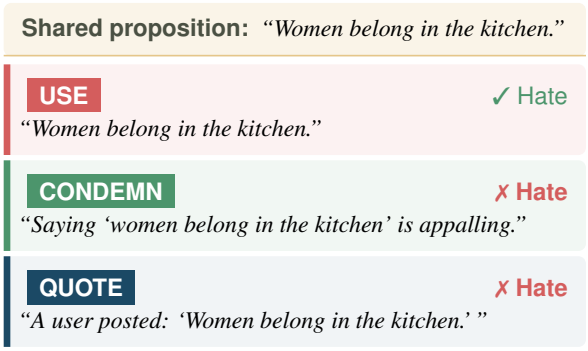
Abstract

Hate speech classifiers often misclassify counter speech, including condemnation and quotation of hateful content, as hateful because they fail to distinguish propositional content from speaker stance. We identify this issue as content stance conflation and address it as a representation learning problem. We learn stance invariant representations of the same proposition by constructing contrastive pairs that share content but differ in stance. We implement this idea through factorized annotation, stance controlled augmentation, and Stance Contrastive Learning (SCL). On the Implicit Hate Corpus, stance variant augmentation reduces counter speech false positives by 60%, and SCL further improves cross domain F1 by up to 6.5 points on top of the augmented signal. On SBIC, where use mention conflation is less prominent in the training labels, SCL maintains a low counter speech false positive rate. Semantic similarity pairing on the same augmented data yields no cross domain gain over the model trained without contrastive learning (0.623 vs. 0.625), which identifies the pair construction signal rather than contrastive learning itself as the determining factor.

1 Introduction

Hate speech on social media poses a persistent threat to online safety, and automated detection has become essential for content moderation at scale (Fortuna and Nunes, 2018; Poletto et al., 2021). Yet one critical failure mode remains unresolved: classifiers systematically flag *counter-speech*, posts that oppose, condemn, or quote hateful content, as hateful (Gligorić et al., 2024; Vidgen et al., 2021). This is not merely a nuisance. Counter-speech is among the most effective responses to online hate (Chaudhary et al., 2021). When moderation systems suppress it, the communities already targeted lose their voice, and trust in moderation erodes (Park et al., 2018; Bonaldi et al., 2024).

Figure 1: The **use–mention confound**. Three utterances share the same proposition but differ in stance. A typical classifier flags all three as hateful (✗ = false positive).



Prior work often explains this failure in terms of lexical bias: models over-associate identity terms (Dixon et al., 2018), profanity (Ethayarajh et al., 2022), or dialect markers (Sap et al., 2019) with toxicity. That explanation is real, but incomplete. Counter-speech often restates the very proposition it seeks to reject. "*Saying 'women belong in the kitchen' is appalling*" and "*Women belong in the kitchen*" express opposite speaker intent while sharing the same underlying proposition. Standard hate speech datasets assign a single binary label per utterance, collapsing this distinction between *what is said* and *how it is said*. We refer to this failure as **content–stance conflation**.

This distinction can be framed as the difference between propositional content and speaker stance. The same proposition can be *used* to endorse hate, *mentioned* to condemn it, or neutrally *quoted* (Austin, 1975; Searle, 1969; Sperber, 1981; Saka, 1998). For hate speech detection, the practical implication is straightforward: models that fail to separate proposition from stance will confuse hateful endorsement with opposition to hate itself.

This observation also reveals a limitation of contrastive learning. Contrastive methods are now widely used to improve generalization in hate speech detection (§2), but they typically assume that semantic similarity is a useful pairing signal.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1301–1318
November 6-10, 2026 ©2026 Association for Computational Linguistics

Under content–stance conflation, that assumption becomes problematic. Because hateful use and counter-speech mention share the same proposition, they are naturally nearby in embedding space. Semantically grounded contrastive methods (e.g., Ahn et al., 2024) can therefore pull them *closer*, reinforcing rather than removing the confound. Inference-time prompting (Gligorić et al., 2024) can correct some individual cases, but it does not reshape the representation space learned by the classifier.

We therefore argue that the central issue is not contrastive learning *per se*, but the *pair-construction signal* it uses. To address this, we propose a framework that factorizes content from stance at three levels. At the **annotation** level, three independent LLMs separately assess content hatefulness and speaker stance (§3.2). At the **augmentation** level, these factorized labels guide the generation of condemnation and quotation variants that preserve proposition while changing stance (§3.3). At the **representation** level, **Stance-Contrastive Learning** (SCL) uses these same-proposition, different-stance pairs as contrastive positives, encouraging content representations that are invariant to stance while preserving distinctions between different propositions (§3.4).

This paper makes three contributions. **First**, we propose SCL, to our knowledge the first contrastive learning method for hate speech detection that uses stance difference, rather than semantic similarity or class identity, as the positive-pair construction signal (§3.4), and show empirically that this choice, rather than contrastive learning itself, determines whether contrastive training improves cross-domain generalization (§5.1). **Second**, we formalize content–stance conflation as a representation-learning problem and show that semantic-pairing contrastive methods leave it unresolved (§3.1). **Third**, we introduce a *factorized* multi-agent annotation pipeline with dimension-aware aggregation and targeted single-dimension deliberation (§3.2), together with a stance-controlled augmentation procedure that synthesizes same-proposition, different-stance pairs not naturally present in standard hate speech corpora (§3.3).

2 Related Work

2.1 Hate Speech Detection and the Use–Mention Confound

Hate speech detection has advanced rapidly (Fortuna and Nunes, 2018; Poletto et al., 2021; Davidson et al., 2017), yet implicit hate remains difficult to identify reliably (ElSherief et al., 2021). Prior work has largely explained these errors in terms of lexical bias: models over-associate identity terms, profanity, and dialect markers with toxicity (Dixon et al., 2018; Ethayarajh et al., 2022; Sap et al., 2019; Zhou et al., 2021). While this bias is real, it does not fully explain why counter-speech is repeatedly misclassified. In many cases, the problem is that models fail to distinguish between *using* hateful language to endorse a proposition and *mentioning* it to reject or report it.

Recent work shows that this use–mention distinction remains difficult even for large language models. Gligorić et al. (2024) show that LLMs confuse the offensive use of slurs with mentioning them to condemn, even under inference-time prompting. Bonaldi et al. (2024) similarly show that LLM safety guardrails cannot reliably distinguish rebutting hateful language from using it to cause harm. Existing work therefore identifies the failure mode, but largely treats it as a problem of bias mitigation or inference-time disambiguation. In contrast, we treat it as a representation learning problem and encode the distinction directly into the learned space.

2.2 Contrastive Learning for Hate Speech

Contrastive learning (van den Oord et al., 2018; Chen et al., 2020) is increasingly used to improve generalization in hate speech detection. Existing methods mainly differ in how they define positive pairs. Some rely on latent semantic relations, such as implication-based pairing (Kim et al., 2022) or semantic clustering within each label class (Ahn et al., 2024). Others use task-specific heuristics, including shared labels (Khosla et al., 2021) or hard positives mined from ERM failures (Jiang, 2025). Related methods also disentangle nuisance factors such as platform variation (Sheth et al., 2024) or stylistic realization (Zhao et al., 2025).

However, these factors are orthogonal to the problem studied here. Platform variation and stylistic expression do not capture a speaker's stance toward a shared proposition. Related stance-aware models also fall short of this setting. HAMiSoN-

Table 1: Contrastive pair-construction strategies for hate speech detection.

Method	Pair Signal	Stance?
ImpCon (2022)	Shared implication	No
SharedCon (2024)	Semantic cluster	No
Label-SupCon (2021)	Same label	No
CCL (2025)	ERM failure	No
CATCH (2024)	Platform factor	No
CADET (2025)	Style flip	No
SCL (ours)	Stance difference	**Yes**

MTL (Rodriguez-Garcia and Centeno, 2024) treats stance as an auxiliary prediction target rather than a signal for structuring the representation space, and CDTT (Ran et al., 2025) uses stance across posts rather than toward the proposition expressed within a single utterance. None constructs contrastive pairs by holding proposition constant while varying stance.

2.3 LLM-Based Data Annotation

Recent studies show that LLMs can match or even exceed crowd-worker agreement in annotation tasks (Gilardi et al., 2023; Tan et al., 2024). Reliability can be improved through multi-agent debate (Du et al., 2024; Liang et al., 2024), cascaded escalation (Jung et al., 2025), or reconciliation across model families (Chen et al., 2024). However, most of this work treats annotation as a single-label decision problem. Hate speech annotation is inherently multi-dimensional, and prior work has called for more structured supervision (Davani et al., 2024; Calabrese et al., 2022). Our two-dimensional scheme separates *content hatefulness* from *speaker stance*, providing the factorized supervision needed to construct same-proposition, different-stance training pairs.

3 Method

To correct content–stance conflation, the model must receive supervision that keeps propositional content fixed while varying speaker stance. We therefore build the training signal in three stages. First, **factorized annotation** separates content hatefulness from speaker stance. Second, **stance-variant augmentation** constructs groups of utterances that share a proposition but differ in stance. Third, **Stance-Contrastive Learning** (SCL) encodes this structure into the representation space. Figure 2 summarizes the full framework.

3.1 Problem Formulation: Content–Stance Conflation

Let x denote an utterance and let $\mathbf{z} = f(x)$ denote its content representation. Content–stance conflation can be formulated as a representation learning problem. An ideal $\mathbf{z}$ should preserve *what is said* while suppressing *how it is said*. Concretely, if an endorsement, a condemnation, and a quotation express the same proposition, then their content representations should be close,

$$\mathbf{z}(x_{\text{use}}) \approx \mathbf{z}(x_{\text{condemn}}) \approx \mathbf{z}(x_{\text{quote}}), \quad (1)$$

whereas utterances expressing different propositions should remain distinguishable,

$$\mathbf{z}(x) \not\approx \mathbf{z}(x'), \text{when their propositions differ.} \quad (2)$$

This is a training-time inductive bias. The model is encouraged to rely on proposition-level content rather than stance-correlated surface cues. Final hate prediction is still determined by supervised classification, which must distinguish hateful endorsements from non-hateful condemnations and quotations.

3.2 Factorized Annotation with Multi-Agent Consensus

To construct same-proposition, different-stance supervision, annotation must separate content hatefulness from speaker stance, which a single binary label cannot provide. For each input x_i, three LLM analyst agents from diverse model families perform a two-step sequential analysis. Each analyst m first estimates a content hatefulness probability $p_{i,m} \in [0, 1]$ and then, conditioned on that proposition-level reading, predicts a three-way stance label $s_{i,m} \in \{\textsc{Support}, \textsc{Oppose}, \textsc{Neutral}\}$.

The two dimensions are aggregated separately. The content probability is the analyst mean p_i, paired with a confidence score c_i that is high only when analysts agree *and* the mean lies near 0 or 1; stance is aggregated by majority vote with a confidence reflecting vote unanimity (ties default to NEUTRAL). When the analysts disagree on a dimension, the first analyst (Qwen3-32B) is re-invoked with the others' rationales and asked to resolve only the contested dimension, so that a stance revision cannot implicitly revise the proposition. Full aggregation formulas, confidence definitions, and the ambiguity diagnosis are given in Appendix A.

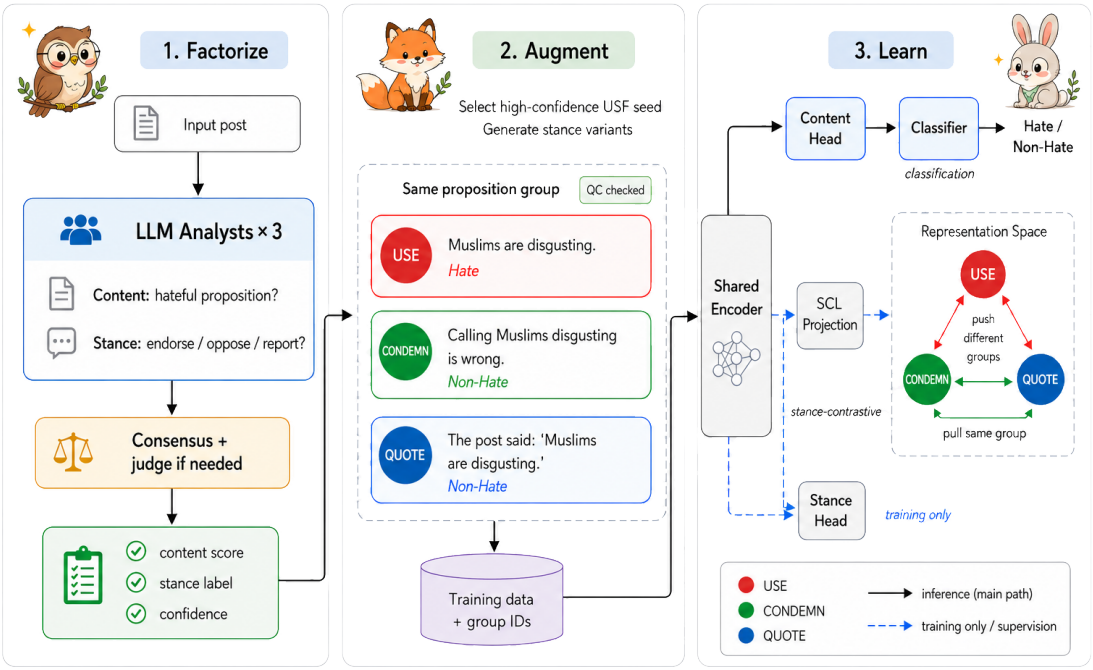


Figure 2: Framework overview. **1. Factorize:** LLM analysts extract content and stance signals from the input post. **2. Augment:** High-confidence hateful seeds are rewritten into CONDEMN and QUOTE variants within the same proposition group. **3. Learn:** The model trains a shared encoder with content classification and stance-contrastive supervision, while the stance head is used only during training.

The pipeline produces a factorized supervision tuple $a_i = (p_i, s_i, c_i)$ for every input. These outputs serve two roles: p_i and c_i provide soft supervision for representation learning, while (p_i, s_i) determine which examples are eligible for stance-variant augmentation.

3.3 Stance-Variant Augmentation

Naturally occurring hate speech datasets contain very few groups that hold proposition constant while varying stance. We therefore synthesize stance variants from high-confidence hateful originals.

Selecting source examples. We first identify hateful seed examples using the factorized labels:

$$\mathcal{H} = \{\, i \;:\; p_i \geq \gamma,\; s_i = \textsc{Support} \,\}, \quad (3)$$

where γ is a selection threshold. These examples most clearly instantiate hateful endorsement and therefore provide the cleanest starting point for controlled stance variation.

Generating stance variants. For each $i \in \mathcal{H}$, an LLM generates two non-hateful variants:

- **Condemn variant** x_i^{cond}: a counter-speech utterance that explicitly rejects the hateful proposition. It receives hate label 0 and stance label OPPOSE.
- **Quote variant** x_i^{quot}: a neutral report or citation of the same proposition. It receives hate label 0 and stance label NEUTRAL.

Each triplet

$$G_i = \left(x_i,\; x_i^{\text{cond}},\; x_i^{\text{quot}}\right) \quad (4)$$

is assigned a proposition-group identifier g_i, indicating that all members of G_i should share the same proposition despite differing in stance.

Quality control. Variant generation follows a three-stage pipeline of multi-candidate generation, cross-model judging, and feedback-based revision. An independent judge scores each candidate on proposition preservation, stance correctness, and naturalness (weighted $0.50/0.35/0.15$; proposition is weighted highest because proposition drift directly corrupts the same-proposition positives required by $\mathcal{L}_{\text{SCL}}$). Candidates below threshold trigger up to two revision rounds (Appendix F). The resulting augmented data supplies the same-proposition, different-stance groups required by SCL.

3.4 Stance-Contrastive Learning

The core design choice in SCL is the *pair-construction rule*. Positives are not defined by semantic similarity or class identity, but by shared proposition and different stance.

Encoder and heads. Each input x_i is encoded into a contextual representation $\mathbf{h}_i$. A content head maps $\mathbf{h}_i$ to a content representation

$$\mathbf{z}_i = f_{\text{content}}(\mathbf{h}_i), \quad (5)$$

and predicts the probability that the input is hateful:

$$\hat{y}_i = \sigma(\mathbf{w}^\top \mathbf{z}_i + b). \tag{6}$$

A projection head maps $\mathbf{z}_i$ into a contrastive space,

$$\mathbf{u}_i = \frac{f_{\text{proj}}(\mathbf{z}_i)}{\|f_{\text{proj}}(\mathbf{z}_i)\|_2}, \tag{7}$$

and an optional stance head produces stance logits

$$\hat{\mathbf{s}}_i = f_{\text{stance}}(\mathbf{h}_i). \tag{8}$$

Binary stance collapse. For contrastive pairing, the three-way stance label is collapsed into a binary distinction: $\tilde{s}_i = \text{AFFIRM}$ if $s_i = \text{SUPPORT}$, and $\tilde{s}_i = \text{NON-AFFIRM}$ otherwise (i.e., $s_i \in \{\text{OPPOSE}, \text{NEUTRAL}\}$). This collapse both increases the number of valid positive pairs per mini-batch and avoids over-interpreting the often subjective boundary between OPPOSE and NEUTRAL.

Pair construction. Given a mini-batch of size B, with proposition-group identifiers $\{g_i\}_{i=1}^B$ and binary stances $\{\tilde{s}_i\}_{i=1}^B$, the positive set for anchor i is

$$P(i) = \{\, j \;:\; g_j = g_i,\ \tilde{s}_j \neq \tilde{s}_i,\ j \neq i \,\}. \tag{9}$$

Thus, an anchor is pulled toward examples that express the *same proposition* but a *different stance*.

Stance-contrastive loss. Pairwise similarity in the projection space is

$$a_{ij} = \frac{\cos(\mathbf{u}_i, \mathbf{u}_j)}{\tau}, \tag{10}$$

where $\tau > 0$ is a temperature hyperparameter. Let

$$\mathcal{A} = \{\, i : |P(i)| > 0 \,\} \tag{11}$$

be the set of anchors with at least one valid positive. The stance-contrastive loss is

$$\mathcal{L}_{\text{SCL}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(a_{ij})}{\sum_{\substack{k=1 \\ k \neq i}}^{B} \exp(a_{ik})}. \tag{12}$$

The denominator includes all non-anchor examples in the mini-batch, so samples from different proposition groups act as implicit negatives.

Group-aware batch sampler. Same-proposition, different-stance positives are sparse: only 707 of the 11,199 IHC proposition groups (6.3%) are stance-diverse. A group-aware sampler that places members of stance-diverse groups into the same mini-batch raises the per-anchor valid-positive ratio from 0.001 (random) to 0.231, and the fraction of batches carrying $\mathcal{L}_{\text{SCL}}$ signal from 3% to 100% (Appendix H).

Content distillation. The LLM-derived content probability p_i provides soft supervision for the content head. We define a confidence-weighted distillation loss

$$\mathcal{L}_C = \frac{1}{N} \sum_{i=1}^{N} c_i \,\text{BCE}(\hat{y}_i,\ p_i), \tag{13}$$

where BCE denotes binary cross-entropy.

Supervised objectives. The main hate classification loss is

$$\mathcal{L}_Y = \frac{1}{N} \sum_{i=1}^{N} \text{BCE}(\hat{y}_i,\ y_i), \tag{14}$$

where $y_i \in \{0, 1\}$ is the gold hate label. An optional auxiliary stance loss is applied to the stance head:

$$\mathcal{L}_S = \frac{1}{N} \sum_{i=1}^{N} \text{FL}(\hat{\mathbf{s}}_i,\ s_i), \tag{15}$$

where FL denotes focal loss.

Final objective and inference. The full training objective is

$$\mathcal{L} = \lambda_Y \mathcal{L}_Y + \lambda_C \mathcal{L}_C + \lambda_{\text{SCL}} \mathcal{L}_{\text{SCL}} + \lambda_S \mathcal{L}_S, \tag{16}$$

where λ_Y, λ_C, λ_{SCL}, and λ_S are loss weights. When the auxiliary stance head is not used, we set $\lambda_S = 0$.

At inference time, only the content head is active. The predicted binary hate label is

$$\hat{y}_i^{\text{bin}} = \mathbf{1}[\hat{y}_i > \tau^*], \tag{17}$$

where τ^* is selected on the validation set. The projection head and stance head are used only during training to impose the desired inductive bias.

4 Experimental Setup

A comprehensive evaluation of the proposed framework should address three questions. First, does stance-aware factorization improve hate speech detection under the original training distribution. Second, do the resulting representations transfer across datasets with different stance–label correlations. Third, is the observed improvement tied to the specific contrastive signal rather than to architectural changes alone.

Table 2: Dataset statistics. Augmentation adds non-hateful stance variants while preserving the original number of hateful examples. DynaHate is used only for cross-domain evaluation.

Dataset	Split	Hate	Non-Hate	Role
IHC	Train	3,280	7,919	Primary
	Valid	1,108	2,625	
	Test	1,072	2,662	
+ Aug.	Train	3,280	9,206	
SBIC	Train	18,093	17,411	Suppl.
	Valid	2,530	2,143	
	Test	2,643	2,055	
+ Aug.	Train	18,093	26,773	
DynaHate	Test	2,271	1,853	Cross

4.1 Datasets

Experiments are conducted on three hate speech datasets, summarized in Table 2. IHC (ElSherief et al., 2021) serves as the primary training corpus. It contains 11,199 examples of implicit hate speech and has a markedly imbalanced label distribution, with 29.3% hateful examples. After stance-variant augmentation, the IHC training split expands to 12,486 examples. The augmented set contains 703 condemn variants and 584 quote variants across 707 proposition groups; 82% of groups contribute both variant types, and the remainder contribute single-stance supervision retained at the same quality threshold.

SBIC (Sap et al., 2020) covers a broader range of social bias types and is roughly three times the scale of IHC. Its augmented version adds 4,778 condemn variants and 4,584 quote variants, yielding 4,796 proposition groups. In the main evaluation, IHC and SBIC serve as each other's training set and cross-domain test set.

DynaHate (Vidgen et al., 2021) is used as an additional cross-domain robustness benchmark and is reserved exclusively for evaluation.

4.2 Baselines

The comparison includes one non-contrastive baseline and several contrastive or factorized baselines. **BERT (No CL)** fine-tunes `bert-base-uncased` with binary cross-entropy only. **Label-SupCon** (Khosla et al., 2021) constructs positive pairs directly from hate and non-hate labels. **SharedCon** (Ahn et al., 2024) applies k-means clustering over sentence embeddings within each label class and treats within-cluster samples as positives. **SharedCon-F** is the original single-head implementation of SharedCon. An additional baseline, **FACTORIZED**, shares our dual-head architecture but replaces training-time contrastive alignment with stance-conditioned logit correction:

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{z} + b + \alpha \sum_k p_S^k \delta_k), \qquad (18)$$

where p_S^k is the predicted probability of stance class k. The offsets $\{\delta_k\}_{k=1}^3$ and the scaling factor α are learned jointly with the encoder by SGD; $\{\delta_k\}$ are initialized as $(+0.5, -0.5, 0)$ for SUPPORT, OPPOSE, and NEUTRAL respectively, and stance-head gradients from $\mathcal{L}_Y$ are stopped so that the auxiliary head is shaped by $\mathcal{L}_S$ rather than by the main task.

4.3 Metrics

Five metrics are reported. **Test F1** denotes in-domain macro F1. **SBIC F1** and **IHC F1** denote cross-domain macro F1 on the corresponding held-out dataset. **Dyna F1** denotes macro F1 on DynaHate.

Because the central failure mode in this paper is counter-speech misclassification, we also report **CS-FPR**, the counter-speech false positive rate. Let

$$\mathcal{M} = \{x_i : y_i = 0 \ \wedge \ s_i = \text{NON-AFFIRM}\} \qquad (19)$$

denote the set of non-hateful examples whose binary stance is non-affirming, including condemn, quote, and neutral cases. CS-FPR is the fraction of $\mathcal{M}$ that is incorrectly classified as hateful:

$$\text{CS-FPR} = \frac{|\{x_i \in \mathcal{M} : \hat{y}_i = 1\}|}{|\mathcal{M}|}. \qquad (20)$$

Unless stated otherwise, CS-FPR is computed on the validation split of the training corpus at a fixed decision threshold of 0.5, whereas the F1 columns use the threshold τ^* selected on validation; all values are means over five seeds. Appendix G examines how stable $\mathcal{M}$ remains under both LLM and human stance labels.

4.4 Implementation

We use `bert-base-uncased` as the primary encoder and replicate all main experiments on `roberta-base` (Appendix K). Both encoders use a maximum sequence length of 128. The `[CLS]` representation (768-dim) is fed to a content head (768→384→1 MLP with ReLU and dropout $p=0.1$), an SCL projection head (384→128 with ℓ_2 normalization), and an optional stance head (768→3 linear). Models are trained for 8 epochs with AdamW (lr $= 4.3 \times 10^{-5}$, batch size 64).

Default loss weights are $\lambda_Y = 1.87$, $\lambda_C = 1.17$, $\lambda_{\text{SCL}} = 1.0$, $\lambda_S = 2.01$ when the stance head is used, and $\tau = 0.07$; the sensitivity and filter-robustness studies in Appendices J and I use $\lambda_Y = 1.0$. Unless noted otherwise, all numbers are averaged over five random seeds. The three LLM analysts are Qwen3-32B, Gemma-3-27B-IT, and GPT-OSS-20B. Model specifications for variant generation are in Appendix F, and verbatim prompt templates are in Appendix N. Training code, augmentation regeneration scripts, and prompt templates will be released under a permissive license upon acceptance.

5 Results and Analysis

We first examine how the choice of pair-construction signal changes the effect of contrastive learning, and then turn to the contribution of each component in the pipeline. Additional validation, including prompting comparisons, human evaluation, robustness analyses, and qualitative examples, is summarized at the end of the section and detailed in the appendix.

5.1 Contrastive Signal Comparison

Table 4 reports results under four training conditions, corresponding to two training datasets and two data regimes. The *Standard* columns are the most informative starting point, because every method there is trained on the same original examples. Differences in those columns therefore reflect differences in the learning signal rather than differences in training data.

Three findings are especially notable. The first concerns semantic pairing. Under the standard setting, the two semantic-pairing methods have the two highest CS-FPR values among all contrastive methods on both corpora. SharedCon and SharedCon-F reach 0.167 and 0.142 on IHC and 0.150 and 0.160 on SBIC, whereas every stance-based variant stays at or below the non-contrastive BERT baseline of 0.146 on IHC and 0.139 on SBIC. Semantic similarity groups hateful use with non-hateful mention of the same proposition, so it offers no mechanism for separating them.

A second pattern appears in the factorized baselines. On standard data, every example forms its own singleton proposition group, which means that $\mathcal{L}_{\text{SCL}} = 0$. The rows marked with † therefore isolate the effect of factorized supervision through $\mathcal{L}_{\text{C}}$ and $\mathcal{L}_{\text{S}}$ without stance-based contrastive alignment. Under this condition, FACTORIZED achieves the highest SBIC F1 of 0.662 and Dyna F1 of 0.619, while SCL+$\mathcal{L}_{\text{S}}$ reaches 0.653 on SBIC and the same 0.619 on Dyna with a lower CS-FPR (0.124 vs. 0.160). Relative to BERT, both factorized variants improve cross-domain performance. Even before true same-proposition, different-stance groups are introduced, separating content from stance already yields a more transferable representation.

The full advantage of SCL becomes visible once augmentation introduces such groups. On augmented IHC, SCL raises SBIC F1 from 0.597 to 0.642 and SCL+$\mathcal{L}_{\text{S}}$ raises it to 0.662, the highest transfer score in the table ($p = .007$ against BERT+aug), while CS-FPR stays at 0.064 and 0.069, less than half the standard-data level of 0.146. In contrast, across the standard-data methods, CS-FPR ranges from 0.167 for SharedCon to 0.095 for SCL+SharedCon. Taken together, these results indicate that the usefulness of contrastive learning depends primarily on the signal used to define positive pairs.

Per-sample predictions from a single seed illustrate the two mechanisms. Among models trained on the same augmented data and the same dual-head architecture, SharedCon introduces false positives on counter-speech examples that BERT classifies correctly with high confidence. These cases typically share topical vocabulary with hateful posts, which is exactly the feature used by semantic clustering to form positive pairs (Appendix Table 16, part a). By contrast, SCL+$\mathcal{L}_{\text{S}}$ corrects BERT false positives on posts containing identity-group vocabulary whose stance is labeled as NON-AFFIRM (Appendix Table 16, part b).

Table 3: Component matrix for Table 4. $\mathcal{L}_{\text{Y}}$: BCE on hate label. $\mathcal{L}_{\text{C}}$: distillation from LLM content probability. $\mathcal{L}_{\text{SCL}}$: stance-contrastive loss on same-proposition, different-stance pairs. $\mathcal{L}_{\text{S}}$: focal CE on the stance head. "Content" inference uses $\sigma(\mathbf{w}^\top \mathbf{z} + b)$; "stance-corrected" adds $\alpha \sum_k p_S^k \delta_k$.

Method	$\mathcal{L}_{\text{Y}}$	$\mathcal{L}_{\text{C}}$	$\mathcal{L}_{\text{SCL}}$	$\mathcal{L}_{\text{S}}$	Inference
BERT, BERT+aug	✓				content
Content-only	✓	✓			content
SCL	✓	✓	✓		content
SCL+$\mathcal{L}_{\text{S}}$	✓	✓	✓	✓	content
FACTORIZED	✓	✓		✓	stance-corrected

5.2 Pipeline Decomposition

The augmented setting reveals a clear division of labor across the pipeline. Augmentation alone sharply reduces counter-speech errors. On IHC, BERT+aug lowers CS-FPR from 0.146 to 0.058, with SBIC F1 moving from 0.615 to 0.597. Content

Table 4: Contrastive signal comparison on BERT-base (five seeds). In-domain and cross-domain macro F1 together with CS-FPR (↓). Green denotes our methods. † indicates that $\mathcal{L}_{\text{SCL}} = 0$ on standard data because all groups are singletons. ‡ indicates that $\mathcal{L}_{\text{S}}$ is trained on the LLM stance labels of the training corpus. CS-FPR is measured on the validation split of the training corpus at threshold 0.5; F1 uses the validation-selected threshold. The active loss term and inference path for each row are summarized in Table 3.

	Train: IHC								Train: SBIC							
	Standard				Augmented				Standard				Augmented			
	IHC	**SBIC**	**Dyna**	**FPR↓**	**IHC**	**SBIC**	**Dyna**	**FPR↓**	**SBIC**	**IHC**	**Dyna**	**FPR↓**	**SBIC**	**IHC**	**Dyna**	**FPR↓**
BERT	.764	.615	.566	.146	.770	.597	.567	**.058**	.832	.617	.618	.139	.831	.616	.623	.155
Content-only	.765	.633	.593	.133	**.775**	.625	.583	.075	**.837**	.618	.620	.127	**.838**	**.619**	**.630**	**.124**
Label-SupCon	.768	.632	.580	.121	.770	.614	.578	.108	.833	.614	**.624**	.131	.837	**.619**	**.630**	.131
SharedCon	.769	.639	.568	.167	.770	.623	.582	.068	.834	**.623**	.609	.150	.832	.616	.628	.147
SharedCon-F	.761	.637	.561	.142	.766	.613	.576	.099	.831	.614	.602	.160	.832	.613	.619	.168
FACTORIZED	.760	**.662**	**.619**	.160	.763	.640	**.614**	.073	.836	.608	.622	**.118**	.835	.611	.623	.136
SCL†	.765	.633	.593	.133	.772	.642	.593	.064	**.837**	.618	.620	.132	.835	.615	.622	.128
SCL+$\mathcal{L}_{\text{S}}$†‡	.769	.653	**.619**	.124	.772	**.662**	.602	.069	.835	.615	.617	.121	.835	.602	.617	.146
SCL+SharedCon†	**.777**	.647	.594	**.095**	.772	.658	.605	.066	**.837**	.618	.612	.130	.837	.611	.616	.139

distillation then recovers transfer performance by transferring proposition-level supervision from the annotation pipeline into the encoder. The Content-only model raises SBIC F1 to 0.625 while maintaining a low CS-FPR of 0.075.

What SCL adds is a stronger bias toward stance-invariant content representations. On augmented IHC, adding $\mathcal{L}_{\text{S}}$ raises SBIC F1 to 0.662 and Dyna F1 to 0.602, while CS-FPR stays low at 0.069. Relative to augmented-data BERT, the full pipeline gains 6.5 percentage points on SBIC F1 ($p = .007$). The same progression holds on RoBERTa-base, where SCL improves SBIC F1 from 0.586 to 0.621 while reducing CS-FPR from 0.066 to 0.055 (Appendix K).

A likely explanation is distributional asymmetry. Because IHC was collected from accounts promoting hate (ElSherief et al., 2021), stance and label are strongly correlated, allowing models to rely on stance-correlated cues that become fragile across domains. By encouraging stance-invariant content representations, SCL removes this shortcut.

5.3 Additional Validation

We also compared SCL with prompting-based disambiguation. Use–mention prompting reduces CS-FPR, but our trained encoder achieves higher task F1 at substantially lower inference cost (Appendix D).

Human evaluation on 500 samples supports the reliability of the factorized labels in high-confidence regions and confirms that the binary stance collapse is more stable than the original three-way scheme. LLM–human agreement on content reaches 85.4% for high-confidence hateful examples, while the binary stance distinction attains 72.9% agreement with Cohen's $\kappa = 0.45$. Among non-hateful samples, 97.1% of the human-defined NON-AFFIRM cases are also included in the LLM-defined set used for CS-FPR. Full details are reported in Appendix E and Appendix G.

The transfer gains of SCL hold across five random seeds, two encoder architectures, the hyperparameter range tested (Appendix J), and the augmentation-filter strictness (Appendix I). The remaining counter-speech false positives concentrate on utterances whose stance depends on pragmatic context beyond the text itself, identifying a clear boundary for future extensions. Prompting comparisons and qualitative examples are in Appendix L and Appendix M.

6 Conclusion

Counter-speech misclassification arises in large part because hate speech detectors conflate propositional content with speaker stance, leaving hateful endorsement, condemnation, and quotation close in representation space even though they express sharply different communicative intent.

Our framework factorizes content from stance through annotation, augmentation, and contrastive alignment, supplying supervision that preserves proposition while varying speaker stance. The critical factor is the signal used to construct positive pairs. On IHC, the primary corpus, semantic-similarity pairing leaves counter-speech errors unresolved and transfers worse than stance-based pairing, whereas stance-based pairing improves cross-domain generalization while preserving the counter-speech gains of stance-variant augmentation. Moderation systems should therefore be trained to recognize not only what is said, but also

whether that content is being endorsed, opposed, or merely reported.

Limitations

Scalability. The framework uses LLM-based stance-variant generation, which adds a one-time preprocessing cost. The 707 proposition groups on IHC cover 6.3% of the training set. This targeted augmentation is sufficient for the gains reported in §5, and broader coverage is left to future work. The contrastive objective requires same-proposition, different-stance examples to co-occur within a batch, which we satisfy through a group-aware sampler rather than random shuffling.

Annotation and evaluation. The annotation pipeline uses three open-source LLM families. The binary stance labels used by $\mathcal{L}_{\text{SCL}}$ are robust across analyst choices: majority binary stance is defined for all 11,193 originals with three analyst votes (per-analyst seed-set diagnostics in Appendix A). Replication across generator-model families is a useful direction for follow-up work. The human validation deliberately oversamples difficult and ambiguous cases, so the reported agreement rates are conservative relative to population-level performance. CCL (Jiang, 2025) is omitted from the quantitative comparison because no public implementation is available.

Augmentation distribution. The generated condemn and quote variants may differ stylistically from naturally occurring counter-speech. This introduces a possible distribution shift between synthetic training instances and organic non-hateful mentions of hateful propositions. The drop in cross-domain F1 observed for augmentation alone, from 0.615 to 0.597 in Table 4, is consistent with such an effect, although subsequent stance-aware learning recovers and surpasses the standard-data baseline.

Language and domain. All experiments are conducted in English. The use–mention distinction is realized differently in languages with distinct quotation conventions, reporting structures, or stance-marking systems; reported speech in Chinese and evidential morphology in Turkish are two obvious examples. The framework is most useful where same-proposition, different-stance evidence is recoverable from the training distribution, and the contrastive objective's marginal contribution scales accordingly with the prevalence of use–mention pairs in the training labels.

Ethical Considerations

The datasets used in this study, namely IHC, SBIC, and DynaHate, contain offensive and harmful language and were previously released with appropriate ethical review. Our annotation and augmentation pipeline necessarily processes and generates text that refers to hateful propositions, including condemn and quotation variants created for training purposes. These variants are intended to distinguish hateful use from non-hateful mention, not to expand or circulate harmful content. No newly generated hateful content is released.

Human validation was performed by three NLP researchers recruited from different institutions and countries, ensuring diversity in cultural and linguistic backgrounds. All annotators were informed of the nature of the task and the presence of offensive content before participation. Annotators were compensated above the local minimum wage in their respective countries of residence. Content warnings were provided at the beginning of each annotation session.

Even with these precautions, deployment of a hate speech detector remains socially consequential. False positives can suppress counter-speech, especially when users are responding to hate directed at their own communities. That risk is central to the present work. Although the proposed framework reduces one important class of such errors, it does not eliminate them entirely. Any deployment would therefore still require careful monitoring, clear avenues for appeal, and particular attention to how errors affect already marginalized groups.

Acknowledgments

This work was supported by the Australian Research Council (ARC) under the DECRA Fellowship DE240101049. This research includes computations using the computational cluster Katana, supported by Research Technology Services at UNSW Sydney (PVC (Research Infrastructure), UNSW Sydney, 2010).

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10444–10455.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. Is safer better? The impact of guardrails on the argumentative strength of LLMs in hate speech countering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3446–3463.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

M Chaudhary, C Saxena, and H Meng. 2021. Countering online hate speech: An nlp perspective. psyarxiv.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7066–7085.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021.

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

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

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.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with V-usable information. *arXiv preprint arXiv:2110.08420*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. Nlp systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959.

Tianming Jiang. 2025. Learn from failure: Causality-guided contrastive learning for generalizable implicit hate speech detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8858–8867.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: LLM judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

PVC (Research Infrastructure), UNSW Sydney. 2010. Katana.

Hongyan Ran, Di Zhang, Xiaohong Li, Huifang Ma, Caiyan Jia, and Yaogong Feng. 2025. Unsupervised contrastive domain adaptive rumor detection with test-time classifier adjustment. *Information Processing & Management*, 63(2):104341.

Raquel Rodriguez-Garcia and Roberto Centeno. 2024. HAMiSoN-MTL at ClimateActivism 2024: Detection of hate speech, targets, and stance using multitask learning. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 89–95.

Paul Saka. 1998. Quotation and the use-mention distinction. *Mind*, 107(425):113–135.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Paras Sheth, Raha Moraffah, Tharindu S Kumarage, Aman Chadha, and Huan Liu. 2024. Causality guided disentanglement for cross-platform hate speech detection. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 626–635.

Dan Sperber. 1981. Irony and the use-mention distinction. *Radical pragmatics*, pages 295–318.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682.

Chengshuai Zhao, Shu Wan, Paras Sheth, Karan Patwa, K Selçuk Candan, and Huan Liu. 2025. Causality guided representation learning for cross-style hate speech detection. *arXiv preprint arXiv:2510.07707*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3143–3155.

A Annotation Aggregation Details

Each analyst m produces a content hatefulness probability $p_{i,m} \in [0,1]$ and, conditioned on it, a three-way stance label $s_{i,m} \in \{\text{SUPPORT}, \text{OPPOSE}, \text{NEUTRAL}\}$. The content probability is aggregated as the analyst mean,

$$p_i = \frac{1}{M} \sum_{m=1}^{M} p_{i,m}, \tag{21}$$

paired with an agreement-based confidence score

$$c_i = \max\big(0,\ 1 - 4\sigma_p^2\big) \cdot \big|\, 2p_i - 1 \,\big|, \tag{22}$$

where σ_p^2 is the cross-analyst variance of $\{p_{i,m}\}$. The two factors penalize analyst disagreement and proximity to the decision boundary respectively. Stance labels are aggregated by majority vote,

$$s_i = \text{maj}\big(\{s_{i,m}\}_{m=1}^{M}\big)\,, \tag{23}$$

with stance confidence 1.0, 0.8, and 0.3 for unanimous, 2-of-3, and tied votes respectively; tied votes default to NEUTRAL.

A sample is diagnosed as CLEAR when $\sigma_p < 0.15$ *and* the stance vote has at least 2-of-3 agreement; the disagreement dimension(s) otherwise determine one of CONTENT_AMBIGUOUS, STANCE_AMBIGUOUS, or FULLY_AMBIGUOUS. When the diagnosis indicates ambiguity, the first analyst (Qwen3-32B) is re-invoked in a second pass with the other analysts' rationales as additional context, and asked to resolve only the contested dimension. If content is already stable, stance deliberation conditions on the agreed content interpretation, preventing a stance revision from implicitly revising the proposition itself.

Per-analyst high-confidence seed-set diagnostics yield pairwise Jaccard overlaps of 0.158, 0.179, and 0.450 across the three analysts; the downstream binary stance labels used by $\mathcal{L}_{\text{SCL}}$ remain stable because they aggregate the majority signal, with majority binary stance defined for 100% of originals.

B Annotation Agreement

Table 5 reports the full agreement statistics for the three LLM analysts on IHC and SBIC. The two dimensions behave differently: content hatefulness elicits consistent agreement, while stance shows more spread, especially on IHC, motivating the binary collapse used in the main experiments (Appendix G). Most disagreements concentrate in the content dimension, so targeted deliberation is triggered more often by content ambiguity than by stance ambiguity.

Table 5: Inter-annotator agreement among three LLM analysts on the IHC and SBIC training splits (11,193 and 35,495 posts with three analyst votes).

Metric	IHC	SBIC
Content hatefulness		
ICC	0.53	0.65
Pearson (best pair)	0.67	0.75
Mean absolute deviation	0.15	0.12
Stance label		
Fleiss' κ	0.14	0.30
Unanimous agreement	29.2%	48.8%
2-of-3 agreement	66.9%	49.7%
Diagnosis distribution		
CLEAR	39.9%	54.1%
CONTENT_AMBIGUOUS	56.1%	44.5%
STANCE_AMBIGUOUS	1.1%	0.3%
FULLY_AMBIGUOUS	2.8%	1.1%
Deliberation rate	60.1%	45.9%

C Statistical Significance

Table 6 reports paired two-sided t-tests over five random seeds for SCL+$\mathcal{L}_S$ against the main baselines. On augmented data, the full pipeline achieves statistically significant gains over BERT on both SBIC and Dyna. Under the standard setting, the strongest significance appears on Dyna and against the semantic-pairing baseline, which is consistent with the main claim that the pairing signal determines whether contrastive learning helps or harms.

Table 6: Statistical significance of SCL+$\mathcal{L}_S$ vs. baselines (paired t-test, 5 seeds, two-sided). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Comparison	SBIC	Dyna	CS-FPR
Aug. data (Table 4)			
vs. BERT	$.007^{**}$	$.005^{**}$	.640
vs. Content	$.027^{*}$	.058	.759
vs. FACTORIZED	.104	.304	.783
vs. SCL	.082	$.032^{*}$	.591
Std. data (Table 4)			
vs. BERT	$.018^{*}$	$.001^{**}$	.453
vs. SharedCon	.153	$<.001^{***}$	.230
vs. SCL	$<.001^{***}$	$.003^{**}$	.625
vs. FACTORIZED	.080	.877	.207

D Prompting-Based Comparison

A natural alternative to training-time factorization is to resolve use–mention confusion at inference time through prompting. Table 7 compares our approach with prompting-based baselines on IHC.

Zero-shot prompting yields a CS-FPR of 0.128. Once few-shot examples are introduced, the error rate rises further to 0.167, suggesting that lexical overlap in demonstrations can reinforce the confound rather than correct it. Explicit use–mention instructions substantially reduce CS-FPR to 0.048, but the improvement comes at the cost of prompt engineering and roughly $290\times$ more parameters at inference, which we use as a coarse proxy for relative inference cost.

Under the same evaluation setting, SCL+$\mathcal{L}_S$ trained on augmented data reaches a higher in-domain F1 of 0.772, compared with 0.747 for use–mention prompting, while using approximately $290\times$ fewer parameters at inference time. Prompting can therefore mitigate the failure mode on a case-by-case basis, but training-time factorization provides a more efficient and more accurate solution.

Table 7: Comparison with LLM prompting on IHC. Cost is the parameter ratio ($32\text{B}/110\text{M} \approx 290$), used as a coarse proxy for relative inference cost rather than a measured FLOP count. Prompting rows are scored on the IHC test set, and the trained-encoder rows are the five-seed validation-split means of Table 4.

Method	Params	Cost	Test F1	CS-FPR↓
LLM prompting (Qwen3-32B)				
Zero-shot	32B	$\sim 290\times$	.725	.128
Few-shot	32B	$\sim 290\times$	.720	.167
Use-mention	32B	$\sim 290\times$	.747	**.048**
Trained encoders				
BERT (std)	110M	$1\times$	.764	.146
BERT+aug	110M	$1\times$	.770	.058
SCL+$\mathcal{L}_S$+aug	110M	$1\times$	**.772**	.069

E Human Validation Details

Validation design. The human validation set contains 500 samples divided across two Label Studio projects. *Project 3* ($n = 350$) evaluates LLM annotations for content and stance on original posts. *Project 4* ($n = 150$) evaluates proposition preservation, stance correctness, and naturalness for generated variants. The latter includes 75 condemn and 75 quote variants, each distributed across five diagnostic subcategories of 15 examples. Because the sampling intentionally oversamples borderline cases and minority classes, the resulting agreement

rates should be read as conservative. Three annotators, NLP researchers from different institutions and countries, ensuring diversity in cultural and linguistic backgrounds, labeled each sample independently, producing 1,050 annotations for Project 3 and 450 for Project 4. Annotators were provided with detailed task definitions, per-dimension evaluation criteria, Likert scale definitions, and content warnings about the presence of hateful language. A two-round annotation protocol was used. Annotators first labeled independently and then reviewed cases of disagreement in a second round.

LLM annotation quality. Table 8 summarizes LLM–human agreement on original examples. Agreement is highest at the extremes of the content scale. High-confidence hateful samples ($p_c^{\text{agg}} \geq 0.8$, $n = 82$) reach 85.4% agreement, while confident non-hateful samples ($p_c^{\text{agg}} \leq 0.3$, $n = 113$) reach 95.6%. The most difficult region is the borderline interval ($p_c^{\text{agg}} \in [0.4, 0.6]$, $n = 129$), where agreement falls to 55.0%. This is exactly the part of the space where deliberation is expected to matter most. On IHC, 60.1% of training samples trigger deliberation, whereas high-confidence regions are almost always diagnosed as CLEAR.

Table 8: LLM–human agreement on original text annotations (Project 3, $n = 350$), stratified by content probability.

Confidence region	n	Agreement
LLM vs. human (content)		
$p_c \geq 0.8$ (hateful)	82	85.4%
$0.5 \leq p_c < 0.8$	153	56.9%
$0.4 \leq p_c \leq 0.6$ (borderline)	129	55.0%
$p_c \leq 0.3$ (non-hateful)	113	95.6%
Overall	350	76.3%
LLM vs. human (stance)		
SUPPORT	150	72.0%
OPPOSE	26	15.4%
NEUTRAL	174	67.8%
Overall	350	65.7%
Human inter-annotator		
Content (pairwise)	1,050	91.6%
Content (unanimous)	350	87.4%
Stance (pairwise)	1,050	83.1%
Stance (unanimous)	350	74.9%

Human agreement provides an upper bound on what the LLM pipeline can realistically achieve. Pairwise agreement among human annotators reaches 91.6% on content and 83.1% on stance (Table 8). The gap between LLM–human agreement and this human ceiling is concentrated in the borderline region, which supports the view that most residual disagreement reflects genuine ambiguity rather than arbitrary LLM noise.

Annotator guidelines. Annotators received a written rubric with five components. The task scope is to judge an utterance on two independent dimensions, content hatefulness and speaker stance toward that content. The content scale is a 1–5 Likert scale anchored at 1 (clearly non-hateful), 3 (ambiguous or context-dependent), and 5 (clearly hateful), with the rule that mentioning hateful content to condemn or quote it does not by itself raise the content score. The stance categories are SUPPORT when the speaker endorses the proposition, OPPOSE when the speaker explicitly rejects or condemns it, and NEUTRAL when the speaker reports, quotes, or discusses the proposition without endorsing or opposing it. Edge case rules cover ironic or sarcastic posts (labeled by intended pragmatic effect), posts with multiple propositions (labeled by the most salient one), and reported speech (treated as mention by default unless the framing makes endorsement clear). The variant evaluation in Project 4 rates proposition preservation, stance correctness, and naturalness independently on the same 1–5 scale. The full rubric document is included in the supplementary materials.

Dimensional asymmetry. Stance remains harder than content. Overall three-way stance agreement reaches 65.7%, compared with 76.3% for content, and the difference is driven most strongly by OPPOSE, where agreement falls to 15.4% ($n = 26$). This asymmetry is unsurprising. Fine-grained stance often depends on contextual assumptions, and the same text may be read either as opposition or as neutral reporting. That ambiguity motivates the binary collapse used in the main experiments, namely AFFIRM versus NON-AFFIRM (validated further in Appendix G). In this framework, stance functions as a noise-tolerant pairing signal rather than as a strictly reliable multiclass target.

Variant quality. The generated variants are evaluated in Table 9. Quote variants preserve the underlying proposition much more reliably than condemn variants, with 97.3% proposition preservation compared with 70.7%. This gap is expected because quotation often embeds the original proposition almost verbatim, whereas condemnation usually requires paraphrasing. Stance correctness reaches 69.3% for condemn variants and 54.7% for quote variants. The lower score for quote variants reflects a common annotator tendency to interpret

mention of hateful content as implicit opposition.

Table 9: Variant validation (Project 4, $n = 150$, 3 annotators). Prop. pres. = human majority judges that the variant preserves the original proposition. Stance corr. = majority matches intended stance.

Variant type	n	**Prop. pres.**	**Stance corr.**
Condemn (all)	75	70.7%	69.3%
High quality	15	80.0%	80.0%
Medium quality	15	73.3%	86.7%
Low proposition	15	33.3%	33.3%
Strategy mix	15	80.0%	60.0%
Target group	15	86.7%	86.7%
Quote (all)	75	97.3%	54.7%
High quality	15	100%	33.3%
Low naturalness	15	100%	86.7%
Framing mix	15	100%	33.3%
Longest	15	100%	46.7%
Random	15	86.7%	73.3%
Overall	150	84.0%	62.0%

The diagnostic subcategories show that judge scores track human ratings across the quality range. The deliberately stress-sampled "low proposition" condemn variants reach 33.3% proposition preservation, while the remaining condemn subcategories are at or above 70%. The score is therefore informative at the diagnostic level required for SCL training.

Judge behavior across rubric dimensions. The pipeline divides quality assurance across two complementary components: the variant judge enforces proposition preservation, and the generation strategy (CONDEMN vs. QUOTE) anchors stance. All but one of the $1{,}287$ accepted variants pass the judge's weighted acceptance threshold (the exception is a flagged best-effort candidate), and human annotators confirm preservation of condemn variants in 70.7% of cases, with the remaining 29.3% being partial or thematic preservations rather than topic drift. The strategy-anchored design also yields high stance agreement: the judge assigns the maximum stance score to $1{,}286$ of $1{,}287$ accepted variants. The two-round revision cap rarely binds in practice: $1{,}285$ of the $1{,}287$ variants were accepted at the initial generation, one after the first revision round, and one exhausted both rounds and was kept as a flagged best-effort candidate. SCL is trained against this realistic proposition-level signal, and the downstream impact is characterized in the human validation results above.

F Augmentation Quality Control

Generated variants are produced through a three-stage pipeline consisting of multi-candidate generation, cross-model judging, and feedback-based regeneration. For each hateful original with $p_c \geq 0.7$ and majority stance SUPPORT, Qwen3-32B produces one candidate per prompting strategy. Condemn generation uses five strategies, namely refutation, empathy, humor, fact-checking, and calling-out. Quote generation uses four framings, namely news report, academic citation, moderation log, and social media quotation. In every case, the prompt instructs the generator to explicitly cite or paraphrase the hateful proposition, using temperature 0.7.

A separate judge model, Gemma-3-27B-IT, evaluates each candidate on proposition preservation, stance correctness, and naturalness, with weights 0.50, 0.35, and 0.15 respectively. These weights prioritize proposition preservation over stance correctness over naturalness, because proposition-changing variants directly corrupt the same-proposition positives required by $\mathcal{L}_{\text{SCL}}$; naturalness affects only stylistic distribution. Candidates must achieve a weighted score of at least $w \geq 3.0$, and the highest-scoring passing candidate is retained. When no candidate passes, the judge's critique is fed back to the generator for up to two additional rounds at temperature 0.9. If none of these attempts succeed, the best-effort candidate is kept and flagged. On IHC, this procedure yields 1,287 variants across 707 proposition groups.

G Binary Stance Validity and CS-FPR Reliability

The main text evaluates CS-FPR using a binary stance distinction, namely AFFIRM for SUPPORT and NON-AFFIRM for OPPOSE or NEUTRAL. This collapse is motivated by reliability. As Table 10 shows, agreement improves noticeably when the original three-way stance scheme is reduced to the binary form.

Table 10: Stance agreement: three-way vs. binary collapse (Project 3, $n = 350$). Binary collapse raises Fleiss' κ from 0.70 to 0.74 and LLM–human agreement from 65.7% to 72.9%.

Metric	3-way	Binary
Human inter-annotator		
Fleiss' κ	0.70	**0.74**
Pairwise agreement	83.1%	**86.9%**
Unanimous (3/3)	74.9%	**80.3%**
LLM vs. human majority		
Agreement	65.7%	**72.9%**
Cohen's κ	0.39	**0.45**

CS-FPR denominator stability. The denominator of CS-FPR is the set $\mathcal{M} = \{x_i : y_i = 0 \wedge s_i = \text{NON-AFFIRM}\}$. Among 150 non-hateful validation examples, the LLM pipeline identifies $|\mathcal{M}_{\text{LLM}}| = 146$, whereas human annotators identify $|\mathcal{M}_{\text{HUM}}| = 105$. Crucially, 102 of the 105 human-defined NON-AFFIRM examples, or 97.1%, are also included in the LLM-defined set. In other words, $\mathcal{M}_{\text{HUM}}$ is almost entirely contained within $\mathcal{M}_{\text{LLM}}$.

Direction of bias. The mismatch is also informative. The 44 examples in $\mathcal{M}_{\text{LLM}} \setminus \mathcal{M}_{\text{HUM}}$ are all labeled NEUTRAL by the LLM but SUPPORT by the human majority, with 70% unanimous agreement. These cases appear to be implicit endorsements that the LLM over-attributes to neutral reporting. Their mean $p_c = 0.44$ is considerably higher than that of the core overlap set ($p_c = 0.10$), indicating that they are borderline rather than clear-cut examples. Because these extra cases enlarge the denominator while also potentially adding false positives, CS-FPR computed from LLM labels is conservative. The improvements reported in Table 4 therefore reflect genuine reductions in error rather than artifacts of stance-label noise.

Binary vs. three-class pairing. The binary collapse is a pairing rule for $\mathcal{L}_{\text{SCL}}$, not a claim that OPPOSE and NEUTRAL are pragmatically identical: three-way stance labels are retained for variant generation and for the auxiliary head $\mathcal{L}_{\text{S}}$. We retrain SCL+$\mathcal{L}_{\text{S}}$ on augmented IHC with positives constructed under each rule (three seeds, $\lambda_Y = 1.0$, the setup of Appendix I). Binary pairing yields SBIC F1 = .666 and validation CS-FPR = .057; three-class pairing yields .636 and .066. The binary rule therefore improves both transfer and counter-speech handling, consistent with the reliability gains in Table 10.

H Group-Aware Sampler Diagnostic

Table 11 compares random and group-aware sampling on IHC augmented training. The group-aware sampler places at least one member of each binary stance from the same proposition group into each mini-batch, with the remaining slots filled by ordinary examples that serve as implicit negatives. Labels, losses, and the definition of $P(i)$ are unchanged.

Table 11: Sampler diagnostic on IHC augmented training. Stance-diverse groups are $707/11{,}199 = 6.3\%$ of all proposition groups.

Sampler	Anchor rate	Batches w/ signal	Group cov.
Random	0.001	3.0%	4.2%
Group-aware	**0.231**	100%	100%

I Augmentation Filter Sensitivity

We test whether the gains of SCL depend on treating noisy variants as gold paraphrases by varying the strictness of the augmentation filter and retraining SCL+$\mathcal{L}_{\text{S}}$ on IHC (three seeds, $\lambda_Y = 1.0$; the default row therefore differs slightly from the five-seed, $\lambda_Y = 1.87$ entry in Table 4). *Unfiltered* is the default pipeline. *Strict_prop5* keeps only top-scoring candidates with no feedback round, removing variants most likely to drift in proposition. *Quote_only* removes all condemn variants and trains on quote variants alone; this is the strongest stress test, since condemn is the variant type where judge confidence is broadest.

Table 12: Augmentation filter robustness on IHC. SBIC F1 / validation CS-FPR for SCL+$\mathcal{L}_{\text{S}}$ (three seeds, $\lambda_Y = 1.0$). All settings remain well below the standard BERT CS-FPR of 0.146.

Filter setting	SBIC F1	CS-FPR↓
Unfiltered (default)	.666	.057
Strict_prop5	.676	.078
Quote_only	.661	.090

Transfer F1 is stable across all three settings, and CS-FPR stays well below the BERT baseline of 0.146 in every setting (default .057, strict_prop5 .078, quote-only .090). The same-proposition, different-stance signal therefore provides robust supervision across filter strictness.

J Hyperparameter Sensitivity

SCL is robust to hyperparameter choice. Table 13 shows that, on augmented IHC, SBIC F1 stays

between 0.635 and 0.650 while the single-seed CS-FPR stays between 0.071 and 0.077 across a broad range of λ_{SCL}, τ, and d_p. The framework therefore does not require fine-tuning to a narrow hyperparameter setting to reproduce the reported gains.

Table 13: Hyperparameter sensitivity of SCL on IHC (augmented data). SBIC and Dyna F1 are five-seed means with $\lambda_Y = 1.0$; the CS-FPR column is from a single seed (789, final epoch) because per-seed validation CS-FPR was not retained for this sweep. * denotes the default setting.

Setting	SBIC	Dyna	CS-FPR↓
λ_{SCL} ($\tau = 0.07$, $d_p = 128$)			
0.1	.643	.602	.072
0.5	.645	.603	.074
1.0*	.639	.604	.072
2.0	.647	.604	.071
5.0	.637	.605	.071
τ ($\lambda_{\text{SCL}} = 1.0$, $d_p = 128$)			
0.05	.650	.606	.071
0.07*	.639	.604	.072
0.10	.643	.603	.073
0.20	.650	.602	.077
d_p ($\lambda_{\text{SCL}} = 1.0$, $\tau = 0.07$)			
64	.635	.602	.075
128*	.639	.604	.072
256	.643	.605	.075

K RoBERTa-base Replication

Table 14 reproduces the contrastive signal comparison with RoBERTa-base under the same four training conditions. On augmented IHC, SCL raises SBIC F1 from 0.586 (RoBERTa+aug) to 0.621 and lowers CS-FPR from 0.066 to 0.055, and SCL+SharedCon reaches the highest SBIC F1 (0.631). Unlike BERT, SharedCon does not raise CS-FPR on RoBERTa (0.096 vs. 0.099 on standard data; 0.048 vs. 0.066 on augmented data), but it yields the lowest cross-domain F1 in both regimes (0.561 and 0.583). The transfer benefit of stance-based pairing therefore replicates across encoders, whereas the counter-speech penalty of semantic pairing is encoder-dependent.

L LLM Prompting Comparison

The comparison with prompting reveals that the two approaches are not simply interchangeable. Table 15 reports per-sample disagreement between SCL+$\mathcal{L}_S$+aug and use–mention prompting. The two methods disagree on roughly one fifth of the evaluation set, and the asymmetry of those disagreements is informative. SCL wins more often on hateful samples, which points to a recall advantage, whereas prompting wins more often on non-hateful samples, which points to a specificity advantage. This difference is consistent with the claim that training-time factorization and inference-time prompting correct related but not identical aspects of the use–mention confound.

Table 15: Per-sample disagreement: SCL+$\mathcal{L}_S$+aug vs. use-mention prompting ($n = 3{,}683$).

	Samples
SCL correct, Prompt wrong	400
Prompt correct, SCL wrong	391
Both correct	2,539
Both wrong	353
SCL wins by gold label	
Hateful (recall)	240 (60%)
Non-hateful (specif.)	160 (40%)
Prompt wins by gold label	
Hateful (recall)	111 (28%)
Non-hateful (specif.)	280 (72%)

M Qualitative Examples

The qualitative examples illustrate three distinct mechanisms. First, semantic pairing introduces false positives by grouping counter-speech with hateful posts that share topical vocabulary. Second, stance-aware representations correct false positives made by a standard augmented baseline. Third, the remaining errors arise on utterances whose stance depends on pragmatic context beyond the text, marking a clear boundary between text-internal use–mention cues and conversational pragmatic inference.

Table 16(a) contains examples where SharedCon flips correct BERT predictions into false positives. Table 16(b) shows the reverse pattern, namely counter-speech examples that BERT+aug misclassifies but SCL+$\mathcal{L}_S$+aug corrects. Table 16(c) lists pragmatic-context cases that fall outside the text-internal scope of SCL, where the utterance offers no explicit evidence that the hateful proposition is being opposed rather than endorsed.

N Prompt Templates

This appendix lists the operative content of the prompt templates used at each pipeline stage. Full templates with output-format specifications are released with the code.

Table 14: RoBERTa-base replication (cf. Table 4; five seeds, CS-FPR on the validation split at threshold 0.5). Green denotes our methods. † indicates that $\mathcal{L}_{\text{SCL}} = 0$ on standard data, so SCL and Content-only are the same model there.

	Train: IHC								Train: SBIC							
	Standard				Augmented				Standard				Augmented			
	IHC	**SBIC**	**Dyna**	**FPR↓**	**IHC**	**SBIC**	**Dyna**	**FPR↓**	**SBIC**	**IHC**	**Dyna**	**FPR↓**	**SBIC**	**IHC**	**Dyna**	**FPR↓**
RoBERTa	.784	.565	.481	.099	.784	.586	.542	.066	.842	.601	.627	.172	.840	**.605**	**.640**	.129
Content-only	.784	.595	.495	.112	.779	.595	.530	.054	**.845**	.607	**.640**	**.129**	.843	.601	**.640**	.122
Label-SupCon	.782	.590	.493	.131	.789	.574	.523	.089	.842	.593	.631	.137	.840	.596	.639	**.114**
SharedCon	.790	.561	.478	**.096**	**.790**	.583	.519	**.048**	.842	**.610**	.623	.168	.840	.594	**.640**	.142
SharedCon-F	.783	.580	.480	.125	.771	.618	.538	.057	.844	.601	.621	.151	.839	.596	.634	.159
FACTORIZED	.762	**.643**	**.521**	.163	.782	.585	**.551**	.066	.844	.608	.633	.150	.842	.598	.632	.134
SCL†	.784	.595	.495	.112	**.790**	.621	.541	.055	**.845**	.607	**.640**	**.129**	.842	.582	.627	.127
SCL+$\mathcal{L}_S$†	.766	.640	.519	.177	.787	.617	.548	.060	.843	.608	.635	.145	.844	.601	.627	.134
SCL+SharedCon†	**.791**	.591	.494	**.096**	.780	**.631**	.539	.060	.844	.606	.626	.133	**.847**	.598	.634	.130

N.1 Analyst Content Prompt

```
You are analyzing a social media post for
hateful propositional content.
Step 1: Identify the propositional content.
Step 2: Estimate the probability that this
content, considered independently of who
is uttering it, expresses hate toward a
protected group.
Step 3: Return JSON: {"content_hate_prob":
float in [0,1], "content_conf": float
in [0,1], "hateful_elements": [str],
"target_groups": [str], "reasoning": str}.
```

N.2 Analyst Stance Prompt

```
Given the post and its content analysis,
classify the speaker's stance toward the
propositional content as: SUPPORT (endorses
or spreads), OPPOSE (rejects or condemns), or
NEUTRAL (reports or quotes without taking a
side).
Return JSON: {"stance": one of [SUPPORT,
OPPOSE, NEUTRAL], "stance_conf": float in
[0,1], "reasoning": str}.
```

N.3 Generator: CONDEMN

Five strategies (refutation, empathy, humor, fact-check, calling-out) share the following frame; only the strategy keyword differs.

```
Rewrite the following hateful post as a
counter-speech utterance using the {strategy}
strategy. Your output MUST mention the
hateful proposition (cite, quote, or closely
paraphrase it) before opposing it. Generic
opposition that could apply to any hate
speech is unacceptable.
Original: "{text}". Target groups:
{target_groups}.
```

N.4 Generator: QUOTE_REPORT

Four framings (news_report, academic, documentation, social_media_quote) share the following frame.

```
Rewrite the following hateful post as a
neutral {framing}-style report. The output
must factually document the original claim
(quote or closely paraphrase its exact words)
without endorsing or condemning it. Generic
abstract summaries that lose the specific
claim are unacceptable.
Original: "{text}". Target groups:
{target_groups}.
```

N.5 Variant Judge

```
Rate the variant on three dimensions on a 1–5
scale:
(P) proposition preservation: does the
variant preserve the original hateful claim
verbatim or as a close paraphrase?
(S) stance correctness: does the
variant clearly take the intended stance
({intended_stance})?
(N) naturalness: does it read as a fluent
utterance?
Return JSON: {"prop": int, "stance": int,
"natural": int, "critique": str}.
```

N.6 Prompting Baselines

The use-mention baseline uses the following classification prompt.

```
Classify the post as HATE (1) or NOT-HATE
(0). Important: a post that mentions
hateful content for the purpose of
condemning, quoting, or neutrally reporting
it is NOT itself hate speech; only posts that
use hateful content to endorse hate are HATE.
Post: "{text}". Answer with a single digit.
```

The zero-shot and few-shot baselines use the same instruction without the use–mention clarification; few-shot adds three demonstrations drawn from the IHC training set, balanced across hate and non-hate.

Table 16: Qualitative analysis of counter-speech predictions (all augmented-data models, gold = 0). (a) SharedCon introduces false positives that BERT handles correctly. (b) SCL+$\mathcal{L}_S$ corrects BERT false positives. (c) Pragmatic-context cases where surface cues provide no stance signal.

(a) SharedCon worsens: BERT correct → SharedCon FP (topical vocabulary overlap)

Text	**Gold**	p_{BERT}	p_{SC}
dems don't worry about color hate is hate to them	0	.004	1.00
what happened to those idiots that are making laws?? they must be aliens from jupitor	0	.011	.994
fuhrer trump and racist white supremacist in full bloom. make america great again! vote hillary [. . .]	0	.006	.989

(b) SCL fixes: BERT FP → SCL+$\mathcal{L}_S$ correct (stance-invariant content repr.)

Text	**Gold**	p_{BERT}	p_{SCL}
this is ok tho & i'm mexican american & i fought 4 my country	0	.697	.232
(((hey fellow white people)))	0	.726	.281
you can kill the unborn if they're not human. bring it back to: what is the unborn? [. . .]	0	.657	.250
its not just about one race, but the fact of the matter is black people are being killed at a higher rate [. . .]	0	.454	.145

(c) Pragmatic-context cases: stance not derivable from text-internal markers alone

Text	**Gold**	p_{SCL}
. we . don't need any, until they prove to be safe for us, to be arround them.	0	.954
it would be very beneficial to let some jews join guys you need to be more diverse [. . .]	0	.936
the price of white existence is being called racist.	0	.879