Counterspeech Generation as an Alignment Problem: Leveraging Counterfactual Data Augmentation

Anonymous ACL submission

Abstract

Counterspeech is an effective strategy for mitigating online hate speech by generating constructive and non-toxic responses. While large language models (LLMs) have shown strong capabilities in this domain, their high computational cost limits practical deployment, motivating the use of smaller, more efficient models. However, smaller models often struggle to produce high-quality, contextually relevant counterspeech, especially in zero-shot settings. In this work, we frame counterspeech generation as an *alignment problem* and investigate data-centric approaches to improve the performance of smaller LLMs. We first evaluate standard alignment methods including supervised fine-tuning (SFT) and preference optimization via Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO). Our analysis shows that SFT often degrades performance, while preference-based methods provide limited and inconsistent improvements when trained on standard data. To address these limitations, we propose a **counterfactual data augmentation** framework that generates diverse training instances by systematically varying target attributes in hate speech inputs. These counterfactuals are used to construct richer preference datasets, enabling stronger supervision signals for alignment. Experimental results across four datasets and two open-source 7B models demonstrate that counterfactual augmentation significantly and consistently improves counterspeech quality as measured by our primary automated metric, substantially narrowing the gap to GPT-4 in several settings. We discuss the limitations of automated evaluation and outline directions for human validation.

1 Introduction

The rapid evolution of online platforms has simultaneously facilitated global communication and amplified the propagation of harmful content, most notably hate speech—expression that attacks or demeans individuals or groups based on attributes like race, religion, gender, or sexual orientation. Traditional moderation methods such as content removal and automated filtering often face criticism for limited effectiveness, potential censorship risks, and a failure to address the underlying motivations for hate speech itself.

Counterspeech emerges as a dialogue-driven, non-censorial alternative: any text that counteracts the potential hate expressed on social media against a community or a person belonging to a community, seeking to challenge or refute it through constructive dialogue rather than outright deletion (Benesch, 2014). Platforms like Meta actively advocate counterspeech to battle online hate.[1]

Manually crafting counterspeech at scale is infeasible given the volume of online hate. The advent of Large Language Models (LLMs) presents a promising avenue for automating counterspeech generation (Saha et al., 2024a). However, deploying such systems introduces significant practical constraints in terms of computational cost and latency. While larger LLMs achieve strong performance, their resource requirements make them impractical for large-scale deployment, motivating the use of smaller, more efficient models.

Smaller LLMs for counterspeech generation often produce generic, shallow outputs insufficiently tailored to the context. In this work, we make the following contributions:

- We frame counterspeech generation as an *alignment problem* and conduct a systematic evaluation of SFT and preference optimization (DPO and ORPO) on smaller open-source LLMs.

- We demonstrate that chain-of-thought (CoT)

[1] https://about.fb.com/news/2022/02/facebook-counterspeech-in-pakistan-uk/

prompting is the most effective zero-shot strategy, and that DPO finetuning significantly improves counterspeech quality while ORPO excels at producing argumentatively structured responses.

- We propose a novel **counterfactual data augmentation** pipeline that systematically varies target groups in hate speech to construct richer preference datasets, and show this consistently improves alignment across datasets and models.

- Using CF-augmented training, smaller 7B models substantially close the gap to GPT-4 on our primary automated quality metric; human evaluation remains an important open direction.

2 Related Work

Counterspeech has long been recognized as a powerful, non-censorial tool for combating online hate speech (Benesch, 2014). Early research focused on creating foundational datasets (Chung et al., 2019; Fanton et al., 2021); the scarcity of high-quality hate speech–counterspeech pairs has historically been the primary bottleneck for ML approaches (Qian et al., 2019).

Prior to the LLM era, automated counterspeech involved Seq2Seq models and controlled text generation frameworks (Saha et al., 2022; Zhu and Bhat, 2021). The advent of powerful LLMs shifted the paradigm, enabling effective zero-shot generation (Saha et al., 2024a). Studies comparing text-davinci-003, GPT-3.5, GPT-4, Llama-2, and Flan-T5 established that vanilla zero-shot performance can surpass traditional methods, but generated content often lacks the argumentative quality required for real-world impact.

On the alignment side, RLHF (Christiano et al., 2023) provides a mechanism for human feedback but is complex and computationally expensive. DPO (Rafailov et al., 2023) reformulates alignment as direct classification over preference pairs, while ORPO (Hong et al., 2024a) integrates SFT and preference objectives into a unified loss. Outcome-constrained LLMs have also been explored for counterspeech generation (Hong et al., 2024b).

Counterfactual methods in NLP. Counterfactual data augmentation has been explored primarily in *classification* settings: Wachter et al. (2018) introduced counterfactual explanations for model decisions, and subsequent work applied demographic substitution to improve *classifier* fairness and robustness in tasks such as toxicity detection and coreference resolution (Zmigrod et al., 2019; Garg et al., 2019). Our work differs in two key respects. First, we operate in the *generative alignment* setting rather than classification, where the goal is to improve the quality of generated text. Second, we use counterfactual inputs not to train a classifier but to construct *preference pairs for DPO/ORPO*: the same hate speech post, rewritten to target a different demographic group, provides a natural contrastive signal when the model's response quality differs across the two versions. This preference-construction mechanism—counterfactual inputs as sources of diversity in a preference dataset—is, to our knowledge, novel in counterspeech and in generative alignment more broadly.

3 Preference Optimization Background

Preference Optimization (PO) refers to a family of techniques that align pretrained LLMs using comparative feedback via triplets $\langle x, y_w, y_l \rangle$, where x is the input prompt, y_w is the preferred (winning) response, and y_l is the rejected (losing) response.

Direct Preference Optimization (DPO). Rafailov et al. (2023) reframes the RLHF objective. Instead of training a separate reward model, DPO directly optimizes the language model using a binary cross-entropy loss that maximizes the log probability of the preferred response y_w relative to the rejected y_l according to the Bradley-Terry model:

$$\mathcal{L}_{\text{DPO}} = -E\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

where π_θ is the fine-tuned model, π_{ref} is the reference (SFT) model, σ is the sigmoid function, and β controls divergence from the reference model.

Odds Ratio Preference Optimization (ORPO). Hong et al. (2024a) combines a standard SFT objective with an odds-ratio-based preference loss into a single unified framework, eliminating the need for a separate reference policy:

$$\mathcal{L}_{\text{ORPO}} = \mathcal{L}_{\text{SFT}} - \lambda\, E\left[\log \sigma\left(\log \frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}\right)\right]$$

We apply both methods to finetune Qwen2-7B-Instruct and Mistral-7B-Instruct on preference datasets constructed from the CrowdCounter benchmark.

4 Datasets and Models

Evaluation datasets. We use four publicly available hate speech–counterspeech datasets: **Reddit** (Qian et al., 2019) (5,257 instances; 500 test samples), **Gab** (Qian et al., 2019) (14,614 instances; 500 test samples), **CONAN** (Chung et al., 2019) (408 English instances; 82 test samples), and **CONAN-MT** (Fanton et al., 2021) (3,718 multilingual instances; 744 English test samples).

Alignment dataset. For preference-based finetuning (DPO, ORPO) and counterfactual generation, we use **CrowdCounter** (Saha et al., 2024b), which provides multiple human-written counterspeech responses per hate speech instance. Preference pairs are constructed using an LLM-as-a-judge (Gemini 2.5 Flash) that ranks responses on effectiveness, politeness, and argumentative strength; higher-ranked responses are treated as chosen and lower-ranked ones as rejected.

Models. We use **Qwen2-7B-Instruct** and **Mistral-7B-Instruct** as our primary open-source models for finetuning and evaluation. We additionally evaluate **GPT-4** as a strong proprietary baseline representing near upper-bound performance.

5 Prompting Strategies

We evaluate four prompting strategies in the zero-shot setting:

Vanilla. A minimal prompt asking the model to generate counterspeech for the given hate speech without any additional instructions.

(+) Explanations. The model is additionally asked to provide an explanation for the generated counterspeech, encouraging deeper reasoning about the response.

(+) Targets. The model is prompted to identify and state the target community of the hate speech, pushing it to infer the victim group before generating the counterspeech.

Chain-of-Thought (CoT). A two-step strategy: the model first explains *why* the input is hateful ($\sim$15 words), then uses that reasoning as an additional cue to generate the final counterspeech. This is the most natural reasoning chain for our task (Wei et al., 2023).

For preference finetuning experiments (Sections 8 and 9), we use the Vanilla prompt for generation consistency, as it is the simplest and most widely applicable setting.

6 Evaluation Metrics

We evaluate generated counterspeech using both referential and quality-based metrics:

BLEURT-20 (Sellam et al., 2020) measures semantic similarity between generated responses and reference counterspeech. As a learned metric, it correlates with human judgments, though its sensitivity is limited for open-ended generation tasks like counterspeech where valid responses can differ substantially in phrasing from references. Differences in BLEURT of less than 0.02 should not be considered practically significant without further statistical validation.

Counterspeech Score (CS-Score). We use an LLM-as-a-judge framework with `gemini-2.5-flash-lite` to evaluate responses on: (i) relevance to the hate post, (ii) tone (empathetic/toxic/neutral), (iii) argumentative strength, and (iv) whether the response explicitly acknowledges the targeted group. The judge is provided with both the original hate speech and the generated counterspeech, and returns a single holistic floating-point score (0–1) after reasoning across all four dimensions jointly (see Appendix A); this differs from averaging sub-scores, preventing any single dimension from dominating. This rubric operationalizes criteria established in prior counterspeech literature and validated against human judgments in Saha et al. (2024a). The CS-Score is our *primary evaluation metric*.

We note an important design property of our evaluation setup: while Gemini models are used in the *training pipeline* for target extraction, counterfactual generation, response generation, and preference labeling, the CS-Score judge evaluates *model outputs at inference time on held-out test sets*. The judge never sees training pipeline outputs; its only input is the hate speech and the evaluated model's generated response. This functional separation mitigates, though does not fully eliminate, risks of circularity. We treat the absence of human evaluation as a genuine limitation (see Section 11) and note that the pattern of gains across datasets—largest on Reddit and Gab, the most informal and structurally

diverse corpora—is more consistent with genuine generalization than with optimization toward a single model family's stylistic preferences.

Argumentative Score (Arg). A RoBERTa-base model trained on argument mining data (Stab et al., 2018) measures the structural quality of arguments. *Caution:* as we discuss in Section 8, this metric is susceptible to high scores from repetitive or malformed outputs that superficially resemble argumentative structure; it should not be interpreted in isolation.

Counter-Argumentative Score (C-Arg). A BERT-based model (Saha et al., 2024a) assesses how effectively the response counters the original hate speech. All models achieve near-ceiling performance (>0.97) on this metric, indicating saturation; we report it for completeness but focus our analysis on CS-Score. Full C-Arg results are provided in Appendix C.

7 Results: Prompting Strategies

Table 1 summarizes results for the five LLMs (GPT-3.5, text-davinci-003, GPT-4, Llama-2, Flan-T5) across four datasets and four prompt variants.

CoT is the best prompting strategy. Chain-of-thought prompting dramatically improves performance across all models and datasets, yielding improvements of up to ∼49% in BLEURT and ∼58% in CS-Score over vanilla prompting. The (+) Explanations strategy is second-best in both metrics. This confirms that structuring the model's reasoning about why a post is hateful leads to more contextually appropriate counterspeech.

Earlier OpenAI models vs. GPT-4. We observe empirically that text-davinci-003 and GPT-3.5 score higher than GPT-4 on both referential and quality metrics for this task. A plausible explanation is that older models generate text semantically closer to the human-written reference counterspeech in our datasets, which were collected before GPT-4 existed and may reflect stylistic patterns closer to those models' outputs. We present this as an empirical observation rather than a definitive claim about model capability; the finding may also partly reflect metric sensitivity to reference style rather than intrinsic counterspeech quality. Decoder-only models generally outperform the encoder-decoder Flan-T5 across all conditions.

Actionable evaluation. Using topic modeling on the top-100 highest-scoring responses per metric, we find that the best counterspeeches by BLEURT address hate targeting *migrants* and *c*nt*, while the best by CS-Score address hate targeting *muslim*, *islam*, and *jews*. This suggests the two metrics are sensitive to different target communities, and both should be considered for a complete evaluation.

8 Results: Preference Optimization on Standard Data

Tables 2 and 3 report results for the two open-source 7B models under baseline (zero-shot), SFT, DPO, and ORPO conditions. GPT-4 (vanilla) is included as a SOTA reference.

DPO consistently improves CS-Score. Across both models and all datasets, DPO achieves the best or second-best CS performance. This is because DPO directly optimizes pairwise preferences that implicitly encode qualities such as relevance, tone, and target-group appropriateness—the exact dimensions evaluated by the CS-Score.

SFT degrades performance. SFT performs particularly poorly, especially for Qwen, where CS-Scores collapse (e.g., 0.067 on GAB). Without a corrective preference signal, the model falls into repetitive loops and produces vague, generic outputs—a consequence of exposure bias and distributional mismatch on the CrowdCounter training set.

ORPO excels at argumentative structure. While ORPO underperforms DPO on CS-Score, it achieves the highest Argumentative Scores across nearly all conditions for the 7B models. Its monolithic objective that mixes SFT and preference learning encourages a more structured, argument-like style. However, this comes at the cost of holistic counterspeech quality.

DPO narrows the gap to GPT-4 on automated metrics. For Mistral-7B under the CoT prompting strategy, DPO-finetuned models achieve a CS-Score of 0.733 vs. GPT-4's 0.735 on our automated judge. While this near-parity is encouraging, we caution that these results rely on automated evaluation; human assessment is needed to confirm whether this gap closure holds in practice.

C-Arg score is saturated. All models achieve C-Arg scores above 0.97, confirming that modern instruction-tuned LLMs are already proficient at

Table 1: Results across datasets and prompt variants (Br: BLEURT, Arg: Argumentative Score, CS: Counterspeech Score, C-arg: Counter-Argumentative Score). Best results per dataset per metric are **bolded**.

		Reddit				Gab				CONAN				CONAN-MT			
Model	**Strat.**	Br	Arg	CS	C-arg	Br	Arg	CS	C-arg	Br	Arg	CS	C-arg	Br	Arg	CS	C-arg
GPT-3.5	Van.	-.46	.35	.76	.46	-.80	.29	.63	.38	-.74	.36	.73	.45	-1.04	.19	.33	.24
	Exp.	-.58	.40	.79	.47	-.55	.40	.82	.44	-.74	.35	.71	.44	-.62	.43	.72	.41
	Tgt.	-.62	.35	.77	.44	-.56	.39	.81	.43	-.63	.43	.88	.44	-.63	.39	.72	.41
	CoT	**-.41**	.41	.83	.44	**-.41**	.43	**.89**	.43	**-.41**	.43	**.89**	.43	-.56	.48	**.86**	.33
Davinci	Van.	-.45	.35	.78	.44	-.37	.38	.85	.42	-.74	.35	.73	.46	-.47	.43	.77	.38
	Exp.	-.50	.39	.76	.47	-.44	.39	.81	.43	-.75	.36	.72	.43	-.63	.44	.75	.39
	Tgt.	-.46	.35	.71	.44	-.40	.42	.84	.45	-.62	.41	.82	.46	-.79	.51	.83	.43
	CoT	-.42	.41	.83	.44	-.41	.43	**.89**	.43	-.81	.43	**.89**	.43	**-.56**	**.48**	.85	.33
GPT-4	Van.	-.51	.21	.61	.55	-.54	.23	.60	.48	-.81	.31	.65	.54	-.57	.38	.69	.46
	Exp.	-.63	.26	.70	.55	-.63	.28	.77	.51	-.85	.30	.59	.56	-.57	.41	.81	.51
	Tgt.	-.80	.18	.53	.54	-.75	.19	.62	.50	-.77	.27	.73	.53	-.79	.33	.76	.55
	CoT	-.52	.28	.73	.53	-.49	.30	.78	.47	-.49	.39	.88	.50	-.57	.44	.88	.42

Table 2: Counterspeech Score (CS) for Mistral-7B-Instruct and Qwen2-7B-Instruct. Green: best; Yellow: second-best (excluding GPT-4).

Dataset	**Mistral**				**GPT-4**
	Base	DPO	ORPO	SFT	Base
CONAN	.740	.777	.723	.651	.796
CONAN-MT	.776	.802	.748	.637	.828
GAB	.603	.676	.601	.556	.675
Reddit	.591	.655	.578	.546	.711
	Qwen				
CONAN	.774	.797	.753	.243	.796
CONAN-MT	.812	.813	.782	.286	.828
GAB	.746	.719	.680	.067	.675
Reddit	.718	.720	.659	.106	.711

Table 3: BLEURT-20 scores for Mistral-7B and Qwen2-7B.

Dataset	**Mistral**				**GPT-4**
	Base	DPO	ORPO	SFT	
CONAN	.40	.40	.39	.38	.42
CONAN-MT	.45	.46	.41	.37	.45
GAB	.38	.40	.39	.36	.29
Reddit	.38	.40	.38	.36	.33
	Qwen				
CONAN	.41	.41	.38	.21	.42
CONAN-MT	.47	.47	.41	.21	.45
GAB	.43	.42	.39	.21	.29
Reddit	.42	.41	.38	.21	.33

generating text that directly opposes hate speech. This metric lacks discriminative power and should be treated as a necessary but not sufficient condition for quality.

9 Counterfactual Data Augmentation for Alignment

9.1 Motivation

A key limitation in aligning LLMs for counterspeech is the lack of diversity and coverage in training data. Existing datasets capture a limited set of target groups and linguistic patterns, leading to models that fail to generalize across varied real-world scenarios. Additionally, standard preference training may not sufficiently capture fine-grained differences in response quality.

We hypothesize that *counterfactual data augmentation* can address both issues by: (i) increasing diversity through controlled variations of hate speech inputs, and (ii) constructing stronger preference signals by exposing models to semantically similar but contextually varied instances.

9.2 Counterfactual Augmentation Pipeline

Our pipeline consists of three stages, detailed below and illustrated in Algorithms 1–3.

Validity of counterfactual substitutions. We acknowledge that hate speech is often culturally and rhetorically specific, meaning demographic substitution can produce semantically implausible or structurally distorted examples. Several pipeline design choices mitigate this. First, replacements are sampled only from *different* demographic categories (e.g., a religion-targeted post is substituted with a gender or nationality target, not another religion), ensuring structural diversity rather than near-identical variants. Second, the substitution prompt instructs the model to verify grammatical and contextual fit, and to return an empty string when substitution is not possible—posts where the target is too deeply embedded in culturally specific

Algorithm 1 Counterfactual Creation

Require: CrowdCounter dataset D
Ensure: Counterfactual hate speech dataset D_{cf}
Initialize empty D_{cf}
for each hate speech instance $h \in D$ **do**
  Identify target group words T_h using Gemini 2.5 Flash Lite
end for
Group all targets $\{T_h\}$ into categories (race, religion, gender, etc.) using Gemini 2.5 Pro
for each $h \in D$ **do**
  Select one target word $t \in T_h$
  Sample 3 replacement words $\{t'\}$ from *different* categories
  Generate 3 counterfactuals by replacing all occurrences of t with t' via Gemini 2.5 Flash Lite
  Add generated counterfactuals to D_{cf}
end for
return D_{cf}

Algorithm 2 Preference Dataset Construction

Require: Original dataset D, Counterfactual dataset D_{cf}
Ensure: Preference dataset D_{pref}
Initialize empty D_{pref}
for each hate speech $h \in D$ **do**
  Retrieve gold counterspeech responses $\{g_i\}$ and counterfactuals $\{h'\}$ from D_{cf}
  for each counterfactual h' **do**
    Generate counterspeech $\{c_j\}$ using Gemini 2.5 Flash Lite
  end for
  Score all $\{g_i\}$ and $\{c_j\}$ using Gemini 2.5 Flash (LLM-as-judge)
  Construct candidate pairs: Gold vs. Counterfactual (g_i, c_j) and Counterfactual vs. Counterfactual (c_j, c_k)
  Remove pairs with score difference < 0.1
  Select top-3 pairs by highest score difference
  Add selected pairs (with h as prompt) to D_{pref}
end for
return D_{pref}

rhetoric tend to fail gracefully. Third, the score-difference threshold (≥ 0.1) filters preference pairs where the counterfactual produced indistinguishable response quality, which naturally excludes degenerate or incoherent substitutions. Qualitative examples of successful and failed counterfactual generations are provided in Appendix D.

Dataset size and diversity. An important open question is whether gains stem from counterfactual *diversity* or simply more preference data. We note that the CrowdCounter dataset is fixed in size; the CF pipeline generates preference pairs from a structurally different input distribution (modified posts targeting varied demographic groups) rather than simply sampling more pairs from the same distribution. Constructing a matched-size non-CF preference dataset to isolate the diversity effect is a key direction for future work; we treat this as an important limitation of the current study.

Algorithm 3 Counterfactual Alignment Pipeline

Require: CrowdCounter dataset D
Ensure: Finetuned model M^*
$D_{\text{cf}} \leftarrow$ COUNTERFACTUALCREATION(D)
$D_{\text{pref}} \leftarrow$ PREFERENCEDATASETCONSTRUCTION(D, D_{cf})
Finetune base models using DPO/ORPO on D_{pref}
return finetuned model M^*

Implementation details. Target words are extracted as-is (not normalized) to preserve variation. One target word per hate speech is selected for replacement. For both original and counterfactual inputs, counterspeech responses are generated using the same strong LLM (Gemini 2.5 Flash Lite) and the same prompt. Gold responses from CrowdCounter are often lower-scoring than LLM-generated responses due to their short length; hence pairs are not assumed to be gold-is-better—the judge score determines chosen vs. rejected.

10 Results: Counterfactual-Augmented Alignment

Tables 4–8 compare models trained on standard preference data (DPO, ORPO) against those trained on counterfactual-augmented data (CF-DPO, CF-ORPO).

Table 4: Counterspeech Score (CS) – Mistral-7B-Instruct. Green : best; Yellow : second-best (excluding GPT-4).

Dataset	Base	DPO	ORPO	CF-DPO	CF-ORPO	SFT	GPT-4
CONAN	.740	.777	.723	.786	.790	.651	.796
CONAN-MT	.776	.802	.748	.822	.818	.637	.828
GAB	.603	.676	.601	.755	.762	.556	.675
Reddit	.591	.655	.578	.736	.752	.546	.711

Table 5: Counterspeech Score (CS) – Qwen2-7B-Instruct. Green : best; Yellow : second-best (excluding GPT-4).

Dataset	Base	DPO	ORPO	CF-DPO	CF-ORPO	SFT	GPT-4
CONAN	.774	.797	.753	.805	.795	.243	.796
CONAN-MT	.812	.813	.782	.845	.829	.286	.828
GAB	.746	.719	.680	.798	.772	.067	.675
Reddit	.718	.720	.659	.781	.754	.106	.711

10.1 Analysis

CS-Score: CF augmentation delivers consistent, substantial gains. As shown in Tables 4 and 5, CF-DPO and CF-ORPO outperform all other variants across almost all datasets for both Mistral and

Table 6: BLEURT-20 scores with counterfactual augmentation (Mistral / Qwen).

	Mistral					GPT-4
Dataset	Base	DPO	ORPO	CF-DPO	CF-ORPO	
CONAN	.40	.40	.39	.40	.39	.42
CONAN-MT	.45	.46	.41	.44	.42	.45
GAB	.38	.40	.39	.40	.40	.29
Reddit	.38	.40	.38	.39	.40	.33
	Qwen					
CONAN	.41	.41	.38	.40	.40	.42
CONAN-MT	.47	.47	.41	.45	.43	.45
GAB	.43	.42	.39	.41	.40	.29
Reddit	.42	.41	.38	.40	.40	.33

Table 7: Argumentative Scores (Mistral / Qwen) with CF augmentation.

	Mistral					GPT-4
Dataset	Base	DPO	ORPO	CF-DPO	CF-ORPO	
CONAN	.702	.685	.744	.696	.730	.691
CONAN-MT	.707	.701	.727	.728	.746	.682
GAB	.771	.730	.807	.738	.738	.831
Reddit	.778	.765	.825	.760	.759	.830
	Qwen					
CONAN	.691	.685	.760	.685	.738	.691
CONAN-MT	.691	.685	.718	.724	.747	.682
GAB	.787	.798	.846	.716	.723	.831
Reddit	.795	.785	.860	.726	.754	.830

Qwen. Gains are most pronounced on Reddit and Gab (e.g., Mistral CF-ORPO: 0.591→0.752 on Reddit; Qwen CF-DPO: 0.746→0.798 on Gab), which are the most informal and structurally diverse corpora. This pattern—larger gains on the most varied datasets—is consistent with our hypothesis that CF augmentation improves generalization across demographic groups, and is harder to explain as simple optimization toward a particular judge's stylistic preferences. On CONAN-MT, Qwen CF-DPO reaches 0.845, compared to GPT-4's 0.828 on our automated metric; we report this result while noting it requires human validation before stronger claims can be made.

We hypothesize three contributing mechanisms, which we present as explanatory candidates rather than validated ablations: (i) **improved target generalization**—the model is exposed to a broader distribution of demographic targets, potentially learning *how* to respond rather than memorizing responses to specific groups; (ii) **harder negatives in preference training**—CF pairs may create more informative preference contrasts, sharpening distinctions in tone and empathy; (iii) **reduction of template responses**—variation across CF inputs may discourage dataset-specific phrase patterns. Isolating these mechanisms through targeted ablations is an important direction for future work.

BLEURT: Small differences should not be overclaimed. The BLEURT differences across methods are mostly in the 0.01–0.02 range (Table 6), which are too small to interpret without significance testing. The key takeaway is simply that CF training *does not degrade* semantic similarity to references: CF variants perform at par with the best non-CF methods across both models, confirming that CS-Score gains do not come at the cost of reference-level regression.

Argumentative Score: Standard ORPO still leads; metric should be interpreted cautiously. CF methods do not outperform standard ORPO on the Arg metric (Table 7). As discussed in Section 8, this metric is susceptible to high scores from superficially argumentative but incoherent outputs; the fact that SFT—which produces degenerate outputs for Qwen—scores highest on Arg confirms that it should not be read as a holistic quality indicator. CF training appears to produce more concise, target-grounded responses that score lower on surface argumentative markers while being better quality counterspeech by the CS-Score rubric. This divergence itself is informative: it suggests that Arg and CS-Score capture genuinely different dimensions of counterspeech quality.

C-Arg: Results in Appendix. C-Arg scores are reported in Appendix C. All models achieve near-ceiling performance (>0.97), confirming that instruction-tuned LLMs already oppose hate speech effectively at the surface level. The metric lacks discriminative power; we include it for completeness only.

11 Conclusion

We have presented a comprehensive study of counterspeech generation as an alignment problem. Our key findings are:

1. **Prompting**: Chain-of-thought is the most effective zero-shot strategy, improving CS-Score by up to ∼58% over vanilla prompting. Earlier OpenAI models outperform GPT-4 on referential metrics; we present this as an empirical observation rather than a capability ranking.

2. **Standard alignment**: DPO consistently improves holistic counterspeech quality over baselines; ORPO excels at argumentative structure as measured by the Arg metric; SFT

degrades performance due to distributional mismatch. The Arg and C-Arg metrics both exhibit pathological behavior (high scores for degenerate SFT outputs; near-ceiling saturation respectively), and should be interpreted carefully.

3. **Counterfactual augmentation**: Our proposed CF pipeline provides consistent and substantial improvements in CS-Score across all datasets and both models. The gains are largest on the most diverse corpora (Reddit, Gab), consistent with improved demographic generalization. On our automated metric, CF-augmented 7B models substantially narrow the gap to GPT-4; human evaluation is needed to confirm this.

These findings highlight the importance of data diversity and structured augmentation in aligning smaller language models for counterspeech generation. Future work should prioritize: (1) human evaluation to validate automated metric findings; (2) ablation studies isolating the role of counterfactual diversity from dataset size; (3) qualitative analysis of counterfactual generation quality across demographic groups; and (4) multilingual evaluation.

Limitations

Our study has several important limitations that future work should address.

Automated evaluation. Our primary metric (CS-Score) relies entirely on an LLM judge from the same model family used in the training pipeline. While the judge operates on held-out outputs rather than training data, and while the rubric is grounded in prior human-validated criteria, evaluation circularity remains a genuine concern. Human evaluation is the essential next step before strong claims about real-world effectiveness can be made.

Dataset size vs. diversity. We cannot currently isolate whether CF gains stem from increased input diversity or simply from having more preference pairs. A controlled ablation—training on a matched number of pairs drawn from the original distribution without counterfactual augmentation—would be necessary to answer this definitively, and we identify this as our most important open empirical question.

Counterfactual validity. Demographic substitution may produce implausible or distorted hate speech variants for culturally specific rhetoric. Our pipeline includes safeguards (cross-category sampling, contextual substitution, score-difference filtering), but systematic error analysis is lacking. Appendix D provides qualitative examples.

Statistical significance. We do not report significance tests or confidence intervals. BLEURT differences of <0.02 should not be interpreted as meaningful without further testing; conclusions in this paper are based on CS-Score, where the gains are large enough (>0.05 in most CF comparisons) to be unlikely to reverse under reasonable test conditions.

Scope. Our study is limited to English, two model architectures, and automated metrics; generalization to multilingual settings, other model families, and human judgment requires further study.

Ethics Statement

All experiments were conducted purely for research purposes; real-world deployment requires human oversight.

Handling of hate speech data. The datasets used contain real hate speech from Reddit, Gab, and other platforms, handled solely for the purpose of generating counterspeech. All data was used under existing research agreements and was not redistributed.

Counterfactual generation and stereotype risk. Our pipeline generates synthetic hate speech variants by substituting demographic targets. This carries a risk of producing harmful or misleading cross-group associations. We mitigate this through cross-category-only sampling (same-category substitution is explicitly prohibited) and contextual substitution that gracefully fails when replacement would be implausible. All counterfactual data remains internal to the training pipeline and is not released. Nevertheless, we acknowledge that no pipeline-level safeguard fully eliminates the risk of stereotype amplification, and we recommend against using the counterfactual generation components as standalone tools outside the alignment context described here.

Demographic coverage and bias. The target categories used in CF substitution are derived from the CrowdCounter dataset, which may over-represent certain demographic groups and under-represent others. Models trained on this pipeline may inherit biases in how they respond to hate speech targeting underrepresented groups. We rec-

ommend demographic coverage analysis before any deployment.

Limitations of automated counterspeech. Generated counterspeech should not be deployed without human moderation. The models studied here can and do produce imperfect, generic, or tonally inappropriate responses. The challenges documented in this paper—generic SFT outputs, metric saturation, CF distortions—underscore that no current automated system is ready for unsupervised deployment in content moderation contexts.

References

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. volume 31, pages 11–17.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN: Counter narratives through nichesourcing: A multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: A multi-target counter narrative dataset to fight online hate speech. 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 3226–3240.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 219–226.

Jiwoo Hong, Jinhyuk Lee, Jaeho Kim, and Jaewoo Kang. 2024a. ORPO: Odds ratio preference optimization for large language models. In *International Conference on Machine Learning (ICML)*.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024b. Outcome-constrained large language models for countering hate speech.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024a. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024b. CrowdCounter: A benchmark type-specific multi-target counterspeech dataset.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 5157–5163.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Prompt Templates

Vanilla Prompt.

```
Generate a counterspeech for
the post delimited by triple
backticks in about 50 words
```{HS_input}```
```

LLM-as-a-Judge Prompt.

```
Critique the counterspeech for
the following: 1. Relevance
to the hate post 2. Tone
(empathetic, toxic, neutral)
3. Argumentative strength
4. Whether it addresses
the victim group explicitly.
Response Format: - Relevance:
High/Medium/Low - Tone: ...
- Argument Strength: ... -
Target Addressed: Yes/No -
Counterspeech Score: A floating
point between 0 and 1 (higher is
better).
```

Counterspeech Generation for Counterfactuals.

```
You are generating counterspeech
to respond to a harmful
statement. Write a response
that: directly addresses the
harmful idea; challenges or
corrects it with reasoning,
perspective, or empathy; does
NOT repeat slurs explicitly;
avoids generic phrases like "be
kind" or "everyone deserves
respect"; is clear, specific,
and constructive. Write
around 1--3 COMPLETE sentences.
Statement: {hatespeech}
```

B Hyperparameters

SFT (common). batch_size=2, grad_accum_steps=8, epochs=3, lr=2e-4, fp16=True, QLoRA 4-bit nf4, LoRA r=16, alpha=32, dropout=0.05, max_length=512.

DPO/ORPO (common). batch_size=2, grad_accum_steps=8, fp16=True, max_length=1024, max_prompt_length=512, QLoRA 4-bit nf4 double-quant, LoRA r=16, alpha=32, dropout=0.05, beta=0.1, gradient_checkpointing=True.

Mistral-specific (DPO/ORPO). epochs=5, lr=5e-6, warmup_ratio=0.05, padding_side=right.

Qwen-specific (DPO/ORPO). epochs=3, lr=1e-5, warmup_ratio=0.1, padding_side=left, ChatML format.

C CF Results: C-Arg Scores

Table 8: Counter-Argumentative Scores with CF augmentation (Mistral / Qwen). All values are near-saturated; included for completeness.

				Mistral			GPT-4
Dataset	Base	DPO	ORPO	CF-DPO	CF-ORPO	SFT	
CONAN	.996	.989	.991	.972	.981	.983	.982
CONAN-MT	.983	.980	.985	.988	.985	.981	.984
GAB	.979	.978	.976	.985	.975	.983	.983
Reddit	.975	.981	.982	.984	.983	.981	.981
				Qwen			
CONAN	.986	.985	.987	.986	.987	.978	.982
CONAN-MT	.986	.985	.980	.988	.988	.965	.984
GAB	.979	.984	.984	.984	.977	.965	.983
Reddit	.982	.989	.984	.989	.983	.968	.981

D Qualitative Analysis of Counterfactual Generation

Table 9 shows representative examples of counterfactual hate speech generated by our pipeline, including both successful substitutions and cases where the pipeline produces implausible or empty outputs.

The "partial success" case above illustrates the core risk in demographic substitution: replacing a slur targeting one group can produce a structurally similar hateful statement about another group. Our pipeline does not eliminate this entirely; the score-difference threshold provides a statistical filter (pairs where the model responds similarly to both versions are removed), but is not a guarantee. This is a genuine limitation we acknowledge in Section 11.
```

Table 9: Examples of counterfactual generation. "Fail" cases are filtered out by the empty-string fallback or score-difference threshold before entering preference training.

Type	Original hate speech	Counterfactual	Notes
Success	"[Muslims] are taking over our cities and replacing our culture." (target: religion)	"[Women] are taking over our cities and replacing our culture." (target: gender)	Grammatically and semantically coherent; structure preserved; counterfactual clearly targets a different group.
Success	"[Immigrants] are criminals who should be deported." (target: nationality)	"[Christians] are criminals who should be deported." (target: religion)	Plausible substitution; rhetorical structure intact; used in training.
Partial	"Those [n*****s] always act like animals in every neighborhood they move into."	"Those [Jews] always act like animals in every neighborhood they move into."	Substitution is grammatically valid but maps a racist slur's rhetorical framing onto a different group; may reproduce harmful stereotypes. Filtered by low score-difference if models respond similarly.
Fail	"[Go back to your country], you don't belong here."	*(empty string returned)*	Target phrase is idiomatic and culturally specific; model correctly returns empty to avoid implausible substitution. Excluded from D_{cf}.