Validating Automatic Evaluation of Controllable Counterspeech Generation: Rankings Matter More Than Scores

Yi Zheng, Björn Ross*, Walid Magdy
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
Y.Zheng-77@sms.ed.ac.uk, b.ross@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

Counterspeech generation has emerged as a promising approach to combat online hate speech, with recent work focusing on controlling attributes used in counterspeech, such as strategies or intents. While these attributes are often evaluated automatically using classifiers, a key goal of this evaluation is to compare the performance of different generation models. However, the validity of such evaluation results is questionable when the classifiers themselves have only modest performance. This paper examines the automatic evaluation of counterspeech attributes using a multi-attribute counterspeech dataset containing 2,728 samples. We investigate when automatic evaluation can be trusted for model comparison and address the limitations of current evaluation methodologies. We make concrete recommendations for how to perform classifier validation before model evaluation. Our classifier validation results demonstrate that even limited classifiers can produce trustworthy model rankings. Therefore, we argue that when comparing counterspeech generation models, a classifier's ability to rank generation models is a more direct measure of its practical utility than traditional classification metrics, e.g., accuracy and F1.

1 Introduction

Counterspeech is defined as a dialogue-driven method that seeks to counteract and curb the spread of online hate speech[1]. Significant research has focused on generating fluent, non-toxic and diversified counterspeech using natural language processing (NLP) techniques. There is a recent focus on generating counterspeech with various strategies, including warning of consequences, using empathy or affiliation, fact-checking, etc (Mathew et al., 2019; Chung et al., 2021). In this paper, we focus on the automatic evaluation of the attributes of counterspeech, i.e., the controllability of generation. Large Language Models (LLMs) have made it unprecedentedly easy to generate high-quality counterspeech (Saha et al., 2024), and researchers use classifiers to evaluate counterspeech attributes automatically. It has created a critical need for robust automatic evaluation. A primary goal of evaluation is to enable objective model comparison, allowing researchers and practitioners to compare the performance of different LLMs and select the one that best suits their objectives (Walker, 2025; Nucci, 2025). These comparisons inform high-stakes decisions, yet little research has examined how we can validate automatic evaluation of counterspeech generation – how do we know when to trust a classifier? This leads to our over-

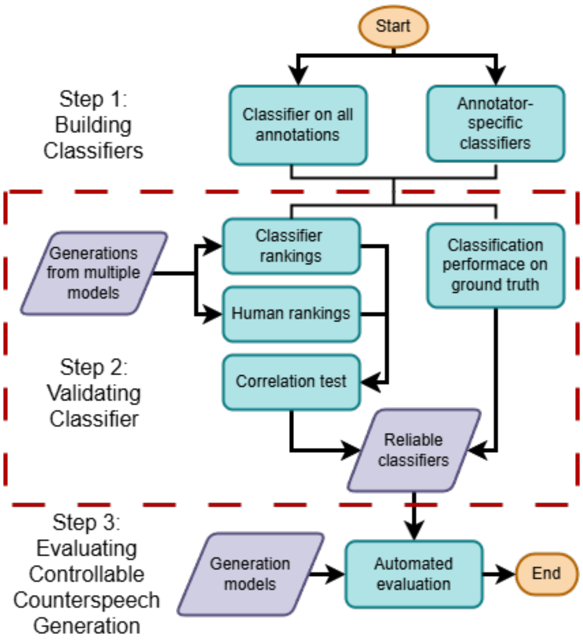
Start
Step 1: Building Classifiers
Classifier on all annotations
Annotator-specific classifiers
Generations from multiple models
Classifier rankings
Human rankings
Classification performace on ground truth
Step 2: Validating Classifier
Correlation test
Reliable classifiers
Step 3: Evaluating Controllable Counterspeech Generation
Generation models
Automated evaluation
End

Figure 1: Three-step process for automatic evaluation of controllable counterspeech generation models, highlighting classifier validation integrated as a key but often overlooked intermediate step between step 1 and 3.

*Corresponding Author

[1] Warning: This work contains offensive and hateful text that some might find upsetting, and that does not represent the views of the authors.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 4131–4146
March 24-29, 2026 ©2026 Association for Computational Linguistics

Paper	Dataset		Auto evaluation		Human evaluation	
	Source	Annotators	Classifier	Metric	Done?	Annotators
Saha et al. (2022)	Public	–	Third party	none	yes	5 experts
Gupta et al. (2023)	Own	Experts	Fine-tuned	Accuracy	yes	60 experts
Saha et al. (2024)	Public	–	Fine-tuned	F1	no	–
Hengle et al. (2024)	Own	Unknown	Fine-tuned	none	yes	35 experts
Wang et al. (2024)	Own	ChatGPT	ChatGPT	none	no	–
Kumar et al. (2025)	Own	Experts	Fine-tuned	Accuracy	no	–

Table 1: Summary of the evaluation approach of counterspeech attributes used in existing research.

arching research question: *When can we trust the automatic evaluation of controllable counterspeech generation?* We argue that model comparison results should be trustworthy when the classifier's rankings of multiple generation models are consistent with human judgements, regardless of absolute performance scores.

Manually evaluating generated counterspeech is time-consuming and labour-intensive. Therefore, current research relies heavily on classifiers (fine-tuned/zero-shot/few-shot) for automatic evaluation of controllability, with examples in Gupta et al. (2023); Wang et al. (2024). However, these classifiers face significant limitations with only modest, sometimes even low classification performance. As observed by Chung et al. (2021), humour and hypocrisy in counterspeech are challenging to detect, with only 0.44 and 0.49 F1. As the authors suggested, this limited performance is likely caused by the challenges of the task, as counterspeech attributes often intersect, making it difficult for classifiers to distinguish between them. Also, data annotations for attributes are subjective, as evidenced by a fair-to-moderate inter-annotator agreement score (Cohen's κ: 0.4-0.6) reported in existing datasets (Poudhar et al., 2024). For the purpose of our study, we fine-tuned multiple classifiers on a dataset we annotated and observed moderate to strong performance (macro-F1: 0.64-0.85). This motivates *RQ 1: How critical is high classifier performance for automatic evaluation? Can imperfect classifiers still reliably rank LLMs?*

Since annotating counterspeech attributes is inherently subjective, we trained separate classifiers using each annotator's labels to investigate how this variation in opinion affects evaluation outcomes. This motivates *RQ 2: When comparing models, how does the variation in annotator opinions affect the results of automatic evaluation?*

In this paper, we examine the common practice of automatic evaluation of controllable counterspeech generation. Through validating classifiers, it was observed that even classifiers with imperfect F1 or accuracy can produce trustworthy model rankings. We therefore argue that when the goal is to compare generation models, researchers should prioritise validating a classifier's ranking alignment with human judgments over relying solely on its classification scores. This step can be integrated into existing evaluation pipelines, resulting in a three-step process (illustrated in Figure 1): (1) a classifier is required to detect counterspeech attributes; (2) validate the classifier to assess its trustworthiness for automatic evaluation; (3) use the classifier to compare generation models. We release a multi-attribute counterspeech dataset to support further research in this area.

2 Background

2.1 Controllable Counterspeech Generation

Counterspeech is a broad term that varies widely in its characteristics; the context in which it is delivered plays a role in determining the appropriate response. Mathew et al. (2019) observed that different target groups favour different counterspeech strategies. They observed that among the African-American community, counterspeech with strategies of "warning of consequences" and "denouncing hateful" received more likes than other types. With controllable counterspeech generation, it is possible to tailor the content to align with the preferences of target communities. Gupta et al. (2023); Hengle et al. (2024); Saha et al. (2024); Poudhar et al. (2024); Wang et al. (2024); Kumar et al. (2025) focus on generating attributes that have been previously analysed in various studies (Benesch et al., 2016; Chung et al., 2019; Mathew et al., 2019), including "warning", "questioning", "empathy", "facts", etc. Meanwhile, studies in Saha et al. (2022); Mun et al. (2023); Sougata and Rohini

(2023) focus on attributes outside this taxonomy, such as emotions or argument schemes.

2.2 Evaluation of Controllability

Previous studies on counterspeech generation without controllable styles have typically relied on ground truth comparisons for evaluation (Qian et al., 2019; Zhu and Bhat, 2021). In contrast, most recent work on controllable counterspeech generation adopts reference-free automatic evaluation methods to assess controllability. A primary goal of this evaluation is to compare different generation models, for example, to select the best-performing set of hyperparameters. To do this, some studies fine-tune classifiers to evaluate attribute accuracy (Gupta et al., 2023; Hengle et al., 2024; Saha et al., 2024; Kumar et al., 2025), while some use LLMs (Wang et al., 2024) or public models (Saha et al., 2022). As shown in Table 1, only some work report classifier classification performance (e.g. F1/accuracy), however, they proceed with automatic evaluation without further validation. Among all studies, automatic evaluation is applied only at the final stage to compare generation models, and it remains unclear whether it was used during model development. Saha et al. (2024); Wang et al. (2024); Kumar et al. (2025) rely solely on automatic evaluation, whereas Saha et al. (2022), Gupta et al. (2023) and Hengle et al. (2024) supplement it with human evaluation (with annotators different from those in dataset creation) on a subset of generations to confirm automatic evaluation results.

Some studies have also built classifiers to detect counterspeech attributes that are useful for evaluation (Chung et al., 2021; Mun et al., 2023; Sougata and Rohini, 2023; Poudhar et al., 2024), however, judging by traditional metrics like accuracy and F1, existing classifiers face significant limitations. For instance, the classifier for "hypocrisy" in Saha et al. (2024) achieved the lowest F1 score of 0.59 among all attributes. Similarly, the classifier for "hypocrisy" in Chung et al. (2021) achieved a macro-F1 of 0.49 under a multilingual setting, while the classifier for "humour" achieved an even lower F1 of 0.44. The few-shot GPT-4 model in Mun et al. (2023) achieved an average F1 score of 0.6 on six attributes. In Poudhar et al. (2024)'s work, the macro-averaged F1 of the few-shot GPT-3.5 classifier across all attributes is 0.62. Gupta et al. (2023) and Hengle et al. (2024) use a multi-class classifier, with an accuracy of 0.75. Chung et al. (2021) observed that certain counterspeech attributes may share similarities or intersect with each other, making it challenging for classifiers to distinguish their differences which leads to limited performance. It seems challenging to build a classifier that achieves top performance for all counterspeech attributes. Given these shortcomings, it is crucial to investigate whether classifiers are, despite their flaws, trustworthy enough to be used for automatic evaluation of generation models.

3 Data

To investigate in-depth automatic counterspeech generation, we construct a dataset that is made publicly available to the community[2].

3.1 Counterspeech Attributes

In constructing our dataset, we adopt the counterspeech taxonomy from Benesch et al. (2016), which has been widely adopted by research on controllable counterspeech generation (Gupta et al., 2023; Hengle et al., 2024; Saha et al., 2024; Poudhar et al., 2024; Wang et al., 2024; Kumar et al., 2025). We took their taxonomy of common counterspeech strategies and added a separate label about the counter-speaker's `Stance` towards the hate speaker. The counterspeech strategies attributes include: `Factual`, `Denouncing`, `Empathy` and `Hypocrisy`, with definitions in Appendix A. A counterspeech might employ multiple or no strategies, but it must have one label for `Stance`. In the first example in Table 2, the counter-speaker uses two strategies in their response. In the second example, the counter-speaker does not use any strategies in their response.

3.2 Dataset Construction

We construct an English counterspeech dataset based on the publicly available hate speech dataset from Fanton et al. (2021). We use their dataset consisting of 880 hate speech (H_S) sentences with their corresponding 2,728 counterspeech produced in the first human-in-the-loop session (C_S). We hired five native English speaker annotators (named $A1 - A5$) to label the 2,728 H_S/C_S pairs according to the strategies used in the counterspeech. For the guidelines, see Appendix A. The annotators are five undergraduate/PhD students in computer science with age ranging between 21-30. All annotators took part in two training sessions. They

[2] Link: https://osf.io/azb63/?view_only=3e05bd61045b4a2cb5ec3581caec1368

#	Example	Attributes
1	Hate Speech: *Must be hard being gay. They have to do everything we do but also find the time to molest children.* Counterspeech: *Child molesters are not all gay, just as men are not all rapists.*	Factual: ✓ Denouncing: ✓ Stance: *neutral*
2	Hate Speech: *White immigrants are welcome.* Counterspeech: *What about when they come from other races or religions?*	Stance: *antagonistic*

Table 2: Examples of counterspeech to hate speech comments and the attributes present in each.

Attribute	End of training	Final
Factual	0.479	0.438
Denouncing	0.475	0.432
Empathy	0.407	0.350
Hypocrisy	0.317	0.280
Stance	0.382	0.354

Table 3: The inter-annotator agreement scores of each attribute calculated with Fleiss' Kappa.

Attributes per C_S	
Counts	total
0	317
1	1768
2	626
3	27
4	0

Attribute	Counts
Factual	1472
Denouncing	551
Empathy	680
Hypocrisy	378
Neutral	1459
Positive	260
Antagonistic	1009

Table 4: (Left) The total number of counterspeech that has 0, 1, 2, 3 and 4 strategies labels. (Right) The counts of each attribute in the dataset.

were shown 20 randomly selected H_S/C_S pairs, discussed the annotation guideline and labelled the 20 samples together, then discussed more counterspeech attribute examples and separately worked on another randomly selected 50 pairs before discussing those as a group. All annotators had access to support resources since viewing hateful and offensive content can be distressing. Table 3 (middle column) shows the Fleiss' κ (Fleiss, 1971) for inter-annotator agreement (IAA) for each of the attributes annotated after the second training session. Detecting the presence or absence of attributes in counterspeech is an inherently hard and subjective task as counterspeech attributes may intersect with each other, as observed by Chung et al. (2021).

2,628 H_S/C_S instances were split between annotators, while 100 pairs were annotated by all five annotators to measure IAA (Table 3, right column). Results show a moderate degree of agreement between annotators on the attributes `Factual` and `Denouncing`, while there is only fair agreement with the remaining attributes, demonstrating variations in interpretations of attributes and the subjectivity of the task. For the 100 samples annotated by five annotators, we use a majority vote to decide the final label. We also refer to these 100 samples and their attribute labels as the ground-truth data throughout the rest of the paper.

Table 4 presents the number of attributes found in each of the counterspeech samples and the number of occurrences of each attribute in the dataset. Most counterspeech only has one strategy attribute (1768) and some have two (626). Other counterspeech did not have any specific strategy in the response (317). There are 27 counterspeech with three attributes, while no counterspeech has four attributes. With regard to the counts of attributes in counterspeech, `Hypocrisy` and `Stance-positive` are the least used in the data.

3.3 Attribute Analysis

To provide deeper insight into the relationships between labels, we calculated attribute co-occurrence statistics. The analysis reveals that counterspeech strategies often intersect (also observed by Chung et al. (2021)), which helps explain the inherent difficulty of the annotation task and it may be the reason for low to moderate IAA scores. For example, Table 8 shows that 39% of `Empathy` counterspeech instances and 33% of `Denouncing` instances also contain `Factual` content. Such co-occurrence often reflects multi-strategy counterspeech in which a single reply both corrects misinformation and criticises the hate speaker's words. As observed in Table 9, the choice of `Stance` is strongly correlated with the strategies of the response, particularly for antagonistic counterspeech. For `Denouncing` (297 out of 551) or `Hypocrisy` (223 out of 378) counterspeech, over half were delivered with an antagonistic stance, while `Factual` counterspeech is the least likely to be delivered with an antagonistic stance.

4 Automatic Evaluation of Controllable Counterspeech Generation

We outline the three-step process for the automatic evaluation of controllable counterspeech generation (Figure 1) in this section. The goal of this evaluation is to determine how well a set of generation models ($M_1, ..., M_m$) produce outputs that express intended attributes ($Att_1, ..., Att_n$). An automatic method, such as a fine-tuned classifier C, produces performance scores for each model, which are then used to derive a ranking, which we denote as R. By incorporating and improving classifier validation, automatic evaluation becomes more trustworthy.

4.1 Detecting Attributes with Classifiers

For our experiment, we fine-tune classifiers separately for each attribute. Classifier for `Stance` is three-class while classifiers for other attributes are binary. The text classifiers are based on the RoBERTa-large model (Liu et al., 2019) from Huggingface. The classifiers make predictions from both the hate speech and the counterspeech, with input formatted as $[BOS]\, H\_S\, [SEP]\, C\_S\, [EOS]$. The 100 samples annotated by all five annotators is used in testing. The rest of the dataset is divided into two subsets: 90% for training (hyperparameters reported in Appendix B) and 10% for validation. The classifier fine-tuned on all annotations is named C_{All}, which learnt from a mix of subjective annotator viewpoints. Furthermore, we fine-tune classifiers on individual annotator labels (performance reported in Table 11). These annotator-specific classifiers are named C_{A1} to C_{A5}.

4.2 Validation of the Classifiers

4.2.1 Performance on Test Data

Table 5 reports the macro-F1 scores of C_{All} over each attribute on the test set, with classification performance consistent with the classifiers in other studies (Chung et al., 2021; Saha et al., 2024; Poudhar et al., 2024). We calculate the average human performance by averaging the macro-F1 scores of individual annotators. The attributes `Hypocrisy` and `Stance` have the lowest macro-F1 for C_{All} and the average human annotator. This indicates they are the most difficult attributes for both C_{All} and humans to align with the group consensus, likely due to its high subjectivity. With regard to the remaining attributes, although the macro-F1 scores of the classifiers are not perfect, they demonstrate a comparable level of performance with average human. Although some of the existing research would assume these results are qualified for evaluation (Table 1), we argue that the F1 alone is not a sufficient indicator to trust the classifier.

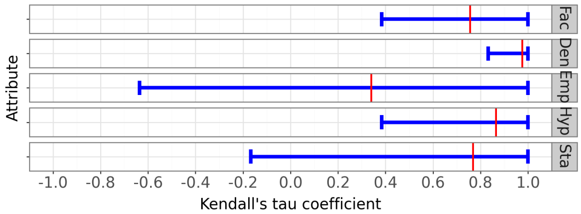

Figure 2: Bootstrapped 90% confidence intervals (blue lines) of the Kendall's τ coefficient (red lines) between C_{All} and human rankings of generation models across five counterspeech attributes. Results vary by attribute: near-perfect for `Denouncing`, strong for `Factual` and `Hypocrisy`, but untrustworthy for `Empathy` and `Stance`.

4.2.2 Performance on Ranking Generation Models

While macro-F1 scores confirm that the majority of the classifiers have comparable performance to humans on test data, the following experiment validates their performance in ranking generation models consistently with human judgments, addressing *RQ 1*: How critical is high classifier performance for automatic evaluation? Can imperfect classifiers still reliably rank LLMs?

Procedure: For the experiment, we use six LLMs for zero-shot generation: Llama-3-chat-8B (Grattafiori et al., 2024), Claude-3-opus[3], Gemma-9B-Instruct (Team et al., 2024), Phi-4 (Abdin et al., 2024) and DeepSeek-V3 (Liu et al., 2024). The outputs are out-of-distribution with respect to the classifier training data. The LLMs' controllability for each attribute is evaluated separately using macro-F1 as the criterion. Every LLM generates 1,000 counterspeech conditioned on 1,000 hate speech randomly chosen from our dataset with the prompt in Appendix D. The generations are balanced to mitigate bias in classifier evaluation: for binary attributes, 500 contain the target attribute (e.g., `Factual`), and 500 do not; for `Stance`, generations are evenly split across classes.

Additionally, for each attribute, two experts (PhD students working on hate speech detection) evaluated 100 generations from each of the six LLMs. The process began with a calibration phase where both experts annotated and discussed 10%

[3] https://www.anthropic.com/claude

Model \ Attribute	Factual	Denouncing	Empathy	Hypocrisy	Stance
RoBERTa	0.82	0.85	0.81	0.64	0.75
Human average	0.84	0.83	0.81	0.76	0.67

Table 5: The macro-F1 scores of the attribute classifiers and the average of human annotators on the 100 samples annotated by five annotators taking the majority vote as the correct label.

of the data together to align their understanding of the guidelines. Then they worked independently in a blind setting on the remaining data. The Cohen's κ IAA scores between their annotations are: Factual ($\kappa = 0.82$), Denouncing ($\kappa = 0.90$), Empathy ($\kappa = 0.76$), Hypocrisy ($\kappa = 0.73$), and Stance ($\kappa = 0.83$). To create the final benchmark, the experts discussed disagreement cases to reach a consensus. We also conducted an intra-annotator consistency check by randomly sampling 20 instances per attribute and asked both experts to re-annotate them after eight weeks. We then compared each expert's labels against the gold labels using Cohen's κ. Agreement between each expert's annotations and the gold labels was substantial across all attributes (κ ranging from 0.8 to 1.00), suggesting that expert judgments are reasonably stable.

We validate C_{All} and C_{A1-A5} by measuring the correlation between classifier rankings of LLMs (R_{auto}) and human rankings of LLMs (R_{human}) instead of only comparing the macro-F1 scores or comparing the best LLM. This is because, in some generation systems, such as those based on ensemble learning, it may be desirable to select several of the top models for further development. We use the two-tailed Kendall's τ rank correlation test (Sen, 1968). We also calculate the 90% confidence interval for the τ coefficient ($-1 \leq \tau \leq 1$) via bootstrapping (Haukoos and Lewis, 2005). The τ value reflects the strength of correlation between R_{auto} and R_{human}, with higher values indicating greater alignment in evaluation results.

Results: The validation results of C_{All} (Figure 2) show significant variation in the degree to which R_{auto} correlates with R_{human} across different attributes. We observe a statistically significant and near-perfect correlation for Denouncing ($\tau = 0.97, p < 0.05$), indicating the classifier is fully trustworthy for evaluation. Factual ($\tau = 0.87, p < 0.05$) and Hypocrisy ($\tau = 0.87, p < 0.05$) show statistically significant and strong correlations with slightly wider confidence intervals. In contrast, the correlations for Empathy and Stance were positive but not statistically significant ($p > 0.05$), showing that the classifiers only ranked some models in the correct order. The τ value of 0 (discordance) is within the confidence interval, indicating the classifiers can produce completely incorrect model rankings. In summary, setting a $\tau = 0$ as the threshold, Factual, Denouncing and Hypocrisy classifiers are trustworthy for evaluation; while Empathy and Stance classifiers are not as trustworthy. Validation results for the annotator-specific classifiers C_{A1-A5} are shown in Figure 4.

4.3 Automatic Evaluation: Identifying the Best Models

To demonstrate the practical value of classifier validation, we illustrate how validated classifiers can support reliable model comparison. In this case study, we identify the optimal model for zero-shot controllable counterspeech generation by evaluating 15 LLMs (listed in Appendix E). Each LLM generates 1,000 counterspeech using 1,000 hate speech inputs and attribute labels from our dataset, following the prompt structure in Appendix D. We use the five attribute classifiers C_{All} for automatic evaluation (Table 12). It shows that DeepSeek-R1 achieves the top performance on Factual and Stance; GPT-4o on Denouncing; Llama-3.3-70B-Instruct on Empathy and Hypocrisy.

As C_{All} for Empathy and Stance failed validation, we instead use annotator-specific classifiers C_{A1-A5} to evaluate all 15 LLMs. To assess agreement among $A1 - A5$ on LLM rankings, we compute the weighted Kendall's tau (Vigna, 2015) coefficient between the rankings produced by C_{All} and C_{A1-A5} (Table 6). This weighted variant of Kendall's tau emphasises higher-ranked LLMs, which suits our purpose of model comparison. We classify coefficients following the interpretation of Wicklin. Overall, the results demonstrate strong correlations in model rankings among trustworthy classifiers, showing consistent evaluation results of LLM controllability between all annotators. However, C_{A4} and C_{A5} for Hypocrisy make an exception with below moderate correlation ($\tau = 0.248$).

	Factual				
	C_{All}	C_{A1}	C_{A2}	C_{A3}	C_{A4}
C_{A1}	0.674				
C_{A2}	0.605	0.456			
C_{A3}	0.407	0.250	0.642		
C_{A4}	0.660	0.641	0.345	0.281	
C_{A5}	0.389	0.305	0.534	0.612	0.390
	Denouncing				
	C_{All}	C_{A1}	C_{A2}	C_{A3}	C_{A4}
C_{A1}	0.711				
C_{A2}	0.020	0.397			
C_{A3}	0.861	0.639	-0.027		
C_{A4}	0.696	0.439	-0.225	0.829	
C_{A5}	0.771	0.719	0.075	0.786	0.706
	Empathy				
	C_{All}	C_{A1}	C_{A2}	C_{A3}	C_{A4}
C_{A1}	0.812				
C_{A2}	0.473	0.366			
C_{A3}	0.576	0.716	0.429		
C_{A4}	0.763	0.693	0.471	0.601	
C_{A5}	0.756	0.800	0.616	0.680	0.603
	Hypocrisy				
	C_{All}	C_{A1}	C_{A2}	C_{A3}	C_{A4}
C_{A1}	0.177				
C_{A2}	0.780	0.353			
C_{A3}	0.514	-0.326	0.384		
C_{A4}	0.652	-0.024	0.626	0.684	
C_{A5}	0.422	0.490	0.543	0.219	0.248
	Stance				
	C_{All}	C_{A1}	C_{A2}	C_{A3}	C_{A4}
C_{A1}	0.780				
C_{A2}	0.535	0.627			
C_{A3}	0.580	0.619	0.475		
C_{A4}	0.698	0.618	0.295	0.353	
C_{A5}	0.790	0.707	0.397	0.457	0.886

Table 6: Using C_{All} and C_{A1-A5} (highlighted in blue are validated as trustworthy) to evaluate 15 LLMs, the table shows coefficients of weighted Kendall's τ correlation test between model rankings across five attributes. **Very strong** ($\tau \geq 0.71$) **Strong** ($0.49 \leq \tau < 0.71$) Moderate ($0.26 \leq \tau < 0.49$)

The correlation between C_{A5} and other annotators is on average lower than the correlation between C_{A4} and other annotators. This suggests variations in opinions among annotators for the attribute Hypocrisy, where models performing well on $A5$'s criteria may not meet other annotators' standards. However, variation for the rest of the counterspeech attributes does not significantly influence evaluation results.

To account for variation in annotator opinions, evaluation should rely on classifiers that align with other annotators. For Factual, Denouncing and Hypocrisy, C_{All} shows strong alignment with most of the trustworthy annotator-specific classifiers, meaning its evaluation results may be recognised by most annotators. As C_{All} for Empathy and Stance failed the validation, we sought reliable alternatives among the annotator-specific classifiers that were validated as trustworthy. We selected the classifier whose model rankings were most consistently aligned with the rankings of the other classifiers. For Empathy, C_{A5} demonstrated consistently strong or very strong correlations with all other classifiers, including the other trustworthy ones, making it the most broadly representative choice. For Stance, C_{A1} showed consistently strong correlations with the majority of the other annotators' classifiers, indicating its judgments were the most representative of the consensus. As detailed in Appendix B, Llama-3.3-70B-Instruct achieves the top performance on Empathy while DeepSeek-R1 on Stance. In summary, leveraging C_{All} for Factual, Denouncing and Hypocrisy, C_{A5} for Empathy and C_{A1} for Stance, DeepSeek-R1, GPT-4o and Llama-3.3-70B-Instruct are the top LLMs for controllable counterspeech generation.

5 Analysis

RQ1: How critical is high classifier performance for automatic evaluation? Can imperfect classifiers still reliably rank LLMs?

First, we assess C_{All}'s performance on the 100 samples annotated by the five annotators (taking majority vote as the correct label), finding that all except for Hypocrisy demonstrate agreement with average human judgements. We then validate the classifiers' ability to compare models by comparing classifier rankings against expert rankings. Results show that C_{All} for Factual, Denouncing and Hypocrisy produce trustworthy model rankings, with the Hypocrisy classifier demonstrating surprising alignment with expert rankings. In contrast, while Empathy and Stance C_{All} achieve good F1 scores, they prove untrustworthy for model ranking, as they can produce incorrect evaluations. These findings highlight the distinction between instance-level classification performance and trustworthiness for model comparison. Therefore, we argue that classification scores (F1/accuracy) are insufficient for assessing if a classifier is fit for evaluation. Instead, the correlation with human judgments on ranking models provides a more direct and meaningful measure.

Empathy, Hypocrisy, and Stance show lower IAA in the dataset, reflecting their subjectivity. However, only Empathy and Stance fail ranking validation. The co-occurrence statistics suggest that Hypocrisy, while subjective, is less entangled with other attributes than Empathy: Empathy cooccurs more frequently with Factual and Denouncing, making its signal more overlapped and ambiguous. As a result, the Hypocrisy classifier can still support stable model rankings, whereas for Empathy and Stance, heavy overlap and class imbalance lead to a larger gap between macro-F1 and ranking alignment– good F1 does not necessarily translate into good ranking agreement with humans.

RQ2: How does the variation in annotator opinions affect the results of automatic evaluation?

Despite variations in opinions on counterspeech attributes, our findings demonstrate that trustworthy classifiers are mostly robust for comparative evaluation, where most correlations between classifier ranking of models are at least moderate except for $A4$ and $A5$ in Hypocrisy. The low IAA between them suggests variation between these annotators (in Table 7), which may have led to different rankings of generation models in evaluation. Interestingly, the IAA between $A3$ and $A4$ shows slight agreement ($\kappa < 0.2$) while the model rankings produced by classifiers (C_{A3} and C_{A4}) trained on their labels are strongly correlated. This suggests that even when annotators disagree on individual instances (low IAA), their overall judgment patterns of counterspeech attributes may still lead to consistent model rankings. Nonetheless, in cases of extreme annotator divergence, we should be careful when using the classifier for evaluation.

6 Discussion and Implications

The advancement of LLMs such as DeepSeek-R1 (Guo et al., 2025) has fundamentally shifted the challenges in controllable counterspeech generation from fluency to fine-grained attribute control. This has made model comparison a central task, requiring researchers and practitioners to reliably determine which of the many available models, prompts, or training configurations is most effective. Current reference-free automatic evaluation approaches often employ ad hoc text classifiers to detect attributes, with examples in Hengle et al. (2024); Kumar et al. (2025). However, they typically trust these classifiers based on their F1 or accuracy scores–though some conduct no validation at all–without validating if they are truly fit for model comparison. This is particularly problematic when classifiers are zero-shot LLMs or fine-tuned on different datasets (e.g. Saha et al. (2022) and Wang et al. (2024)). Our findings for **RQ1** address this gap. We argue that for comparative evaluation purposes, **a classifier's ability to rank generation models correctly is more important than its absolute classification performance**. The discrepancy we observed–where a low-F1 classifier produced rankings that align with human rankings better than a high-F1 classifier proves that F1 and ranking trustworthiness are not interchangeable. Therefore, we recommend that researchers use ranking correlation with human judgments as the primary metric for validating an evaluation tool intended for model comparison, not just classification metrics (Liu et al., 2014).

Annotating counterspeech attributes is a challenging task for humans, and variations in annotator opinions are common (for example, IAA in Mathew et al. (2019); Chung et al. (2019); Poudhar et al. (2024)). Our findings for **RQ2** reveal that **variation in annotator opinions on counterspeech attributes has minimal impact on model comparison outcomes**. Even a classifier trained only on a single annotator's labels can align very well with a classifier based on many annotators, as observed in Empathy and Stance. However, there are exceptions in the cases of extreme disagreement, as seen with the Hypocrisy attribute, where classifiers trained on different annotators produced very different rankings. This underscores the need to account for annotator variation in model evaluation. For studies that fine-tune classifiers based on their own datasets (e.g., Gupta et al. (2023); Saha et al. (2024); Hengle et al. (2024)), our results highlight the importance of carefully selecting annotations for training. In the cases where even classifiers fine-tuned with all annotations are not validated as trustworthy for evaluation, the strong performance of certain annotator-specific classifiers suggests that individual expert judgments may sometimes be *more* suitable for evaluation. The evaluation results from these trustworthy annotator-specific classifiers can align across all annotators.

Furthermore, the challenge with using third-party classifiers (Saha et al., 2022) or zero-shot LLMs (Wang et al., 2024) is that the researcher training the classifier often lacks insight into whether their classifier's outputs align with the intended goals of model evaluation. Our experiments illustrate the necessity of classifier validation to determine trustworthy model comparison. For example, in evaluating Factual with annotator-specific classifiers, all classifiers may agree with each other on model rankings, but only two (C_{A1}, C_{A4}) passed validation. This highlights the importance of performing classifier validation to decide which classifier to use for the automatic evaluation.

7 Conclusion

This paper examines the automatic evaluation of controllability in counterspeech generation with a focus on model comparison by answering the overarching *RQ*: When can we trust the automatic evaluation of controllable counterspeech generation? We argue that classifiers for attribute detection should be validated to understand the trustworthiness of the evaluation results. Through validation, we have demonstrated that traditional classification metrics (accuracy/F1) and ranking reliability capture distinct dimensions of evaluation quality. Our findings reveal that ranking generation models should be the primary criterion for selecting classifiers for model comparison.

Limitations

This paper focuses specifically on evaluating attribute control in controllable counterspeech generation, without considering other important metrics such as text fluency, toxicity, or appropriateness. Though we do not claim that these metrics are less important than attribute accuracy, we emphasise that counterspeech attributes present unique challenges for automatic evaluation. In contrast, metrics like fluency and toxicity have been extensively studied in other controllable text generation tasks, while appropriateness is better captured in human evaluation. Furthermore, our work frames automatic evaluation primarily through the lens of model comparison. While this is a critical task, we acknowledge that automatic evaluation can serve other purposes, such as instance-level error analysis. Our central argument–that ranking correlation is a more suitable validation metric than F1–is most relevant for comparative evaluation, and its applicability to these other scenarios is an area for future work. While this study focuses on English, future work could involve annotating and analysing counterspeech attributes in other languages. Our proposed validation procedure is designed to assess classifiers for attribute detection, and future work could test it on other reference-free metrics. Besides, for classifier validation, we picked six LLMs to generate counterspeech. Each LLM zero-shot generates 1,000 counterspeech samples. To achieve more robust validation results, it would be beneficial to have more generation models and diverse counterspeech outputs.

Ethical Considerations

This paper is driven by the goal of addressing online hate speech through the use of counterspeech to challenge harmful content and promote inclusive discourse. As online hate becomes increasingly prevalent due to AI with malicious intents and algorithmic personalisation (Gorwa et al., 2020; Hinduja; Rughiniș et al., 2024), automatic generation of counterspeech presents a promising solution empowering users and moderators to combat online hate. As Zheng et al. (2023) suggests, the success of counterspeech relies critically on its quality, so it is important to reliably and ethically evaluate counterspeech generation systems. Existing studies rely heavily on automatic evaluation, and poor automatic evaluation practices risk promoting models that generate inappropriate or even counterproductive responses. For instance, a model that fails to express `Empathy` or the intended `Stance` might cause further trauma to victims of hate speech or escalate tensions. As a result, improving automatic evaluation standards for counterspeech generation is not just a technical necessity but an ethical imperative. This paper highlights the importance of validating classifiers before using them in automatic evaluation, we aim to support the development of effective controllable counterspeech generation systems. This contributes to making controllable counterspeech generation a more viable and ethically sound tool in the fight against online hate speech.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for successful counterspeech. *Dangerous speech project*.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, 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 Tekiroğlu, 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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Jason S Haukoos and Roger J Lewis. 2005. Advanced statistics: bootstrapping confidence intervals for statistics with “difficult” distributions. *Academic emergency medicine*, 12(4):360–365.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with rlaif. 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 6716–6733.

Sameer Hinduja. Generative ai as a vector for harassment and harm. https://cyberbullying.org/generative-ai-as-a-vector-for-harassment- and-harm. Accessed: 2025-04-10.

Aswini Kumar, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. Counterspeech the ultimate shield! multi-conditioned counterspeech generation through attributed prefix learning. *arXiv preprint arXiv:2505.11958*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yangguang Liu, Yangming Zhou, Shiting Wen, and Chaogang Tang. 2014. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, 6(4):20–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777.

Antonio Nucci. 2025. Llm evaluation: Key metrics and frameworks. Accessed: 2025-Sep.

Aashima Poudhar, Ioannis Konstas, and Gavin Abercrombie. 2024. A strategy labelled dataset of counterspeech. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 256–265.

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.

Răzvan Rughiniș, Cosima Rughiniș, and Emanuela Bran. 2024. Generative ai and social engines of hate. In *Regulating Hate Speech Created by Generative AI*, pages 1–18. Auerbach Publications.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. 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, 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. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389.

Saha Sougata and Srihari Rohini. 2023. Consolidating strategies for countering hate speech using persuasive dialogues. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 378–392.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, pages 1166–1176.

Stephen M. Walker. 2025. Llm benchmarks. Accessed: 2025-Sep.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9131–9142.

Rick Wicklin. Weak or strong? how to interpret a spearman or kendall correlation. `https://blogs.sas.com/content/iml/2023/04/05/interpret-spearman-kendall-corr.html`. Accessed: 2025-03-15.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, Prague, Czechia. Association for Computational Linguistics.

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.

A Our Dataset

The hate speech and counterspeech instances in our dataset are from the work of Fanton et al. (2021), which is for research purposes only. The dataset in Fanton et al. (2021) does not contain any personally identifiable information and we do not add such information in the annotation. Our dataset is for academic use only.

The five annotators took part in training sessions and had access to support resources in case of any negative effect from viewing hateful and offensive content. Following the institution's policies where this research took place, we set the remuneration of annotators as £13 per hour, over minimum wage. The total amount spent on participant compensation is around £800.

A.1 Annotation guideline

During annotation, we provide an annotation guideline as well as examples for each counterspeech attribute to the annotators.

1. Strategies
    - (`Factual`) Presentation of factual claims: In this strategy, the counterspeech tries to persuade by correcting misstatements/misinformation. An obvious example would be using statistics to correct misinformation. The counterspeech could also use a general and harmless claim to reduce hate. This type of counterspeech must address the substance of the claim.
    - (`Denouncing`) Denouncing hate and Warning of online or offline consequences: The counterspeech could be denouncing the message as being hateful, or warning the hate speaker of possible online/offline consequences of their actions.
    - (`Empathy`) Empathy-based: An empathy-based counterspeech might have "affiliation". The empathy-based counterspeech should show empathy for the target group. The counter-speaker sometimes discloses their affiliation (either with the target group or with the hate speaker) in efforts to counter hate speech. Some

counter-speakers take it upon themselves to apologise on behalf of identities they presumed they had in common with hate speakers. The counterspeech can also express support to the target groups by acknowledging their difficulties or achievements.

- (Hypocrisy) Pointing out hypocrisy or contradiction: In this strategy, the counterspeech points out the hypocrisy in the hate speaker's statements. In other words, "pointing out when someone's actions or statements do not align with their stated beliefs or principles, which is used to hold someone accountable for their actions or words". Pointing out hypocrisy many times does not address the substance of the claim. An example would be counterspeech that uses "whataboutism".

2. (Stance) Stance to the hate speaker

- Positive: We consider empathic, kind, polite, or civil counterspeech to be positive.
- Antagonistic: An antagonistic counterspeech might use abusive, hostile, or obscene comments in response to the hate speech.
- Neutral

A.2 IAA between annotators on the 100 samples annotated by five annotators

Table 7 shows the IAA between annotators on the 100 samples annotated by five annotators.

	Factual				
	All	*A1*	*A2*	*A3*	*A4*
A1	0.879				
A2	0.702	0.587			
A3	0.599	0.516	0.583		
A4	0.589	0.615	0.375	0.218	
A5	0.626	0.519	0.462	0.312	0.309
	Denouncing				
	All	*A1*	*A2*	*A3*	*A4*
A1	1.000				
A2	0.397	0.397			
A3	0.627	0.627	0.306		
A4	0.917	0.917	0.378	0.549	
A5	0.434	0.434	0.232	0.292	0.414
	Empathy				
	All	*A1*	*A2*	*A3*	*A4*
A1	0.806				
A2	0.355	0.272			
A3	0.554	0.466	0.107		
A4	0.752	0.709	0.306	0.355	
A5	0.600	0.451	0.030	0.486	0.315
	Hypocrisy				
	All	*A1*	*A2*	*A3*	*A4*
A1	0.798				
A2	0.532	0.393			
A3	0.348	0.259	0.382		
A4	0.616	0.654	0.274	0.112	
A5	0.358	0.170	0.318	0.089	0.122
	Stance				
	All	*A1*	*A2*	*A3*	*A4*
A1	0.731				
A2	0.632	0.474			
A3	0.499	0.368	0.306		
A4	0.700	0.588	0.410	0.280	
A5	0.472	0.285	0.407	0.189	0.299

Table 7: The table shows the IAA among *All*, *A1*, *A2*, *A3*, *A4*, *A5* on the 100 samples annotated by five annotators, measured by Cohen's κ across five attributes (annotator's classifier validated as trustworthy is highlighted in blue). **Substantial** ($\tau > 0.6$) **Moderate** ($0.4 < \tau \leq 0.6$) **Fair** ($0.2 < \tau \leq 0.4$)

A.3 Analysis of counterspeech attributes in the dataset

Table 8 and Table 9 show the co-occurrence statistics among the five counterspeech attributes of our dataset.

	Total	Fac	Den	Emp	Hyp
Fac	1472	–	–	–	–
Den	551	181	–	–	–
Emp	680	263	92	–	–
Hyp	378	93	35	33	–

Table 8: Co-occurrence statistics among the four counterspeech strategies attributes of our dataset.

Stance	Total	Fac	Den	Emp	Hyp
Neu	1459	926	185	384	133
Pos	260	150	69	73	22
Ant	1009	396	297	223	223

Table 9: Co-occurrence statistics among the stance attribute and four counterspeech strategies attributes of our dataset.

B Fine-tuning Attribute Classifiers

Hyperparameter	Bounds
Epochs	50, 100
Learning rate	2e-6, 5e-6, 1e-5
Batch size	8, 16, 32
Optimizer	AdamW
Weight decay	0.01
Warm up	0.1

Table 10: For hyperparameter search, the bounds for each hyperparameter.

For the purpose of this study, we fine-tune multi-class text classifiers based on the RoBERTa-large model (Liu et al., 2019) from Huggingface[4]. The RoBERTa-large model consists of 355 million parameters. The distribution of each attribute among training, validation and testing sets are balanced. However, we acknowledge that the training set exhibits class imbalance, as detailed in Table 4, with attributes like `Hypocrisy` being less frequent. Class imbalance can potentially bias a classifier towards the majority class during training. To mitigate this during model selection, we used macro-F1 as the primary metric for hyperparameter tuning, because it gives equal weight to each class's performance regardless of its frequency. While other techniques like data resampling or class re-weighting could further improve performance, we opted to follow standard fine-tuning procedures. Our goal was not to develop a state-of-the-art attribute detection classifier, but rather to create and validate classifiers that are representative of those commonly used in practice for automatic evaluation.

	All	Annotator				
		1	2	3	4	5
Fac	0.82	0.80	0.79	0.74	0.77	0.78
Den	0.85	0.88	0.65	0.82	0.78	0.73
Emp	0.81	0.74	0.73	0.82	0.57	0.77
Hyp	0.64	0.70	0.63	0.62	0.60	0.61
Sta	0.75	0.67	0.62	0.64	0.40	0.49

Table 11: Macro-F1 scores of annotator-specific classifiers C_{A1} to C_{A5} when fine-tuned and tested on individual annotators' labels.

We use a single NVIDIA RTX3090 for model training. We report the hyperparameters in classifier manual fine-tuning and the bounds for each hyperparameter in Table 10 (training time is around 1 to 3 hours). Hyperparameter configuration for the best classifier for `Factual` and `Hypocrisy` is epochs of 50, learning rate of 5e-6 and batch size of 16; for `Denouncing`, `Empathy` and `Stance` is epochs of 50, learning rate of 1e-5 and batch size of 32. Furthermore, we trained dedicated classifiers for each annotator's judgements (five annotators total). Table 11 reports the macro-F1 scores.

C Validating Attribute Classifiers

In Section 4.2.2, we conducted human evaluation to validate classifiers' trustworthiness to evaluate generation models. The classifier's ranking of generation models is compared to expert rankings. Although larger sets of human annotations would improve the validation, there are practical resource limitations. To estimate how many samples are needed for human annotation, we randomly select sub-groups of size 10, 25, 50, 75 or 90 from the 100 expert annotations, then follow the validation procedure in Section 4.2.2. This process is repeated 1,000 times. We plot the range of τ in Figure 3. It can be observed that the τ coefficient on larger subgroup sizes (75 or 90) is more concentrated with fewer outliers. This shows that the classifiers can be validated with a limited number of annotated generations (in our case, 100 test samples for each attribute).

We further validated C_{A1} to C_{A5}'s ability to rank counterspeech generation models using the procedure outlined in Section 4.2.2, results shown in Figure 4.

[4] https://huggingface.co/FacebookAI/roberta-large

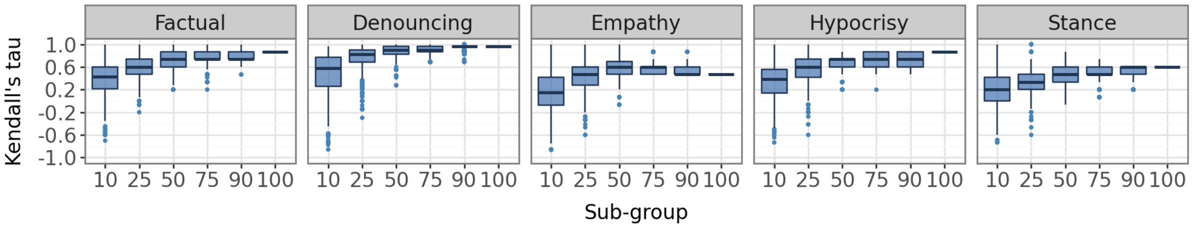

Figure 3: Performing automatic evaluation on 1,000 test samples from each of six generation models and randomly selecting sub-groups of size 10, 25, 50, 75, 90 and 100 for human evaluation (repeated 1,000 times), the boxplot shows the range of τ between classifier rankings and human rankings. A larger sub-group size shows more consistent evaluation results.

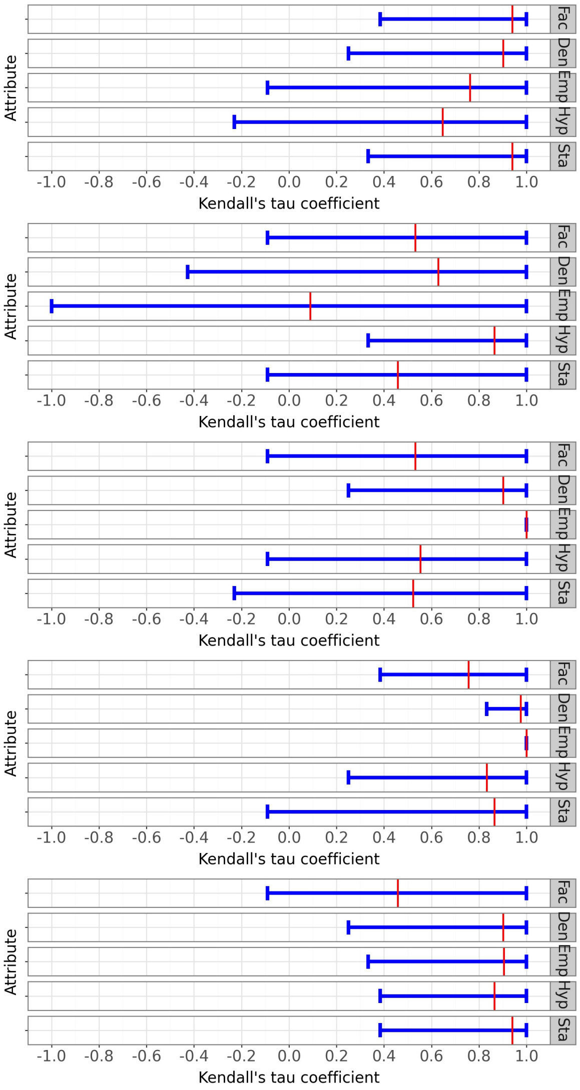

Figure 4: Bootstrapped 90% confidence intervals (blue lines) of the Kendall's τ coefficient (red lines) between classifier (C_{A1} to C_{A5} from top to down) and human rankings of generation models across five counterspeech attributes.

D Zero-shot Generation Prompt

The structure of the prompt used for zero-shot generation is (the prompt is adjusted based on the attributes required to generate):

> *Generate a concise counterspeech (under 50 words) responding to the hate speech below. Your response must strictly include the following attributes:*
>
> - *Factual: presentation of factual claims;*
> - *Hypocrisy: pointing out hypocrisy or contradictions in the hate speech;*
> - *Empathy: showing empathy to the target group;*
> - *Denouncing: denouncing speech as hateful or warn the hate speaker of possible consequences;*
> - *Tone-positive: showing positive tone to hate speaker;*
> - *Tone-neutral: showing neutral tone to hate speaker;*
> - *Tone-antagonistic: showing antagonistic tone to hate speaker.*
>
> *Do NOT include additional elements or exceed 50 words.*
>
> *Hate Speech: <Put Hate Speech>*
>
> *Counterspeech:*

E Zero-shot Generation LLMs

A list of the 15 models that have been used in Sec 4.3 for zero-shot generation of counterspeech. The total cost of running the LLMs was £50. Table 13 and Table 14 show automatic evaluation results by C_{A1} and C_{A5}.

#	Model	Description
1	G4o	GPT-4o (Achiam et al., 2023)
2	G4m	GPT-4o-mini
3	G4T	GPT-4 Turbo
4	G3T	GPT-3.5 Turbo
5	L3.3	Llama-3.3-70B-Instruct
6	L3.1	Llama-3.1-405B-Instruct
7	C	Claude-3.7-sonnet
8	Gb	Grok-beta
9	DS	DeepSeek-R1
10	W	WizardLM-2-8x22B
11	M	Mixtral-8x22B-Instruct-v0.1
12	G	gemma-2-27b-instruct
13	Q	Qwen2-72B-Instruct
14	P3	phi-3.5-MoE-instruct
15	P4	phi-4

LLM Attribute	G4o	G4m	G4T	G3T	L3.3	L3.1	C	Gb	DS	W	M	G2	Q	P3	P4
Factual	0.86	0.83	0.84	0.84	0.87	0.88	0.79	0.83	**0.89**	0.83	0.80	0.80	0.85	0.80	0.78
Denouncing	**0.70**	0.58	0.68	0.59	0.65	0.61	0.66	0.65	0.65	0.64	0.64	0.60	0.64	0.69	0.70
Empathy	0.55	0.46	0.59	0.48	**0.70**	0.58	0.61	0.62	0.63	0.57	0.60	0.61	0.62	0.59	0.53
Hypocrisy	0.52	0.50	0.55	0.51	**0.65**	0.54	0.50	0.58	0.54	0.50	0.51	0.53	0.54	0.47	0.48
Stance	0.49	0.44	0.42	0.56	0.61	0.55	0.64	0.62	**0.74**	0.52	0.51	0.66	0.59	0.44	0.34

Table 12: Automatic evaluation results with C_{All} about attribute controllability, the macro-F1 scores of GPT-4o (G4o), GPT-4o-mini (G4m), GPT-4 Turbo (G4T), GPT-3.5 Turbo (G3T), Llama-3.3-70B-Instruct (L3.3), Llama-3.1-405B-Instruct (L3.1), Claude-3.7-sonnet (C), Grok-beta (Gb), DeepSeek-R1 (DS), WizardLM-2-8x22B (W), Mixtral-8x22B-Instruct-v0.1 (M), gemma-2-27b-instruct (G2), Qwen2-72B-Instruct (Q), phi-3.5-MoE-instruct (P3), phi-4 (P4) controlling each attribute.

LLM Attribute	G4o	G4m	G4T	G3T	L3.3	L3.1	C	Gb	DS	W	M	G2	Q	P3	P4
Stance	0.41	0.37	0.38	0.39	0.44	0.36	0.38	0.48	**0.52**	0.37	0.37	0.50	0.43	0.36	0.30

Table 13: Automatic evaluation results with the `Stance` classifier C_{A1}.

LLM Attribute	G4o	G4m	G4T	G3T	L3.3	L3.1	C	Gb	DS	W	M	G2	Q	P3	P4
Empathy	0.63	0.61	0.70	0.63	**0.78**	0.75	0.73	0.71	0.76	0.64	0.72	0.66	0.71	0.66	0.59

Table 14: Automatic evaluation results with the `Empathy` classifier C_{A5}.