A LLM-Based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation

Irune Zubiaga and **Aitor Soroa** and **Rodrigo Agerri**
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{irune.zubiaga, a.soroa, rodrigo.agerri}@ehu.eus

Abstract

This paper proposes a novel approach to evaluate Counter Narrative (CN) generation using a Large Language Model (LLM) as an evaluator. We show that traditional automatic metrics correlate poorly with human judgements and fail to capture the nuanced relationship between generated CNs and human perception. To alleviate this, we introduce a model ranking pipeline based on pairwise comparisons of generated CNs from different models, organized in a tournament-style format. The proposed evaluation method achieves a high correlation with human preference, with a ρ score of 0.88. As an additional contribution, we leverage LLMs as zero-shot CN generators and provide a comparative analysis of chat, instruct, and base models, exploring their respective strengths and limitations. Through meticulous evaluation, including fine-tuning experiments, we elucidate the differences in performance and responsiveness to domain-specific data. We conclude that chat-aligned models in zero-shot are the best option for carrying out the task, provided they do not refuse to generate an answer due to security concerns.

Warning: *Please be advised that this research paper contains instances of hate speech that may be distressing or offensive to readers. These expressions are included for analysis and critique purposes only, and they do not reflect the beliefs or endorsements of the authors or the institution.*

1 Introduction

The proliferation of misinformation and the dissemination of harmful narratives has stressed the urgent need for effective strategies to combat Hate Speech (HS). This necessity has drawn significant attention to the field of automatic CN generation, where considerable research has focused on the use of LLMs to fulfill this task (Chung et al., 2021; Tekiroğlu et al., 2022). However, difficulties in automatically assessing the quality of the generated CNs remain. As is common in text generation tasks, while manual evaluation is expensive, time-consuming, and subjective, existing automatic methods often fail to provide comprehensive insights or capture the nuanced relationship between generated text and human perception, overlooking crucial aspects of effectiveness and relevance (Ni'mah et al., 2023). Finally, the problem of CN evaluation is exacerbated by the lack of a 'universal truth' and the significant variations among possible references, as shown in Table 1.

In this paper we address the limitations of traditional evaluation metrics for CN generation by proposing a novel automatic evaluation approach. This method is motivated by the need to improve upon existing metrics like BLEU, ROUGE and BERTScore (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020) which do not consider the specific HS to which the CN is responding to, an essential aspect for accurately assessing the quality of CNs. Our approach evaluates generated CNs pairwise in a tournament-style format, with outcomes determined without human intervention using JudgeLM (Zhu et al., 2023), a model explicitly trained to assess the quality of text. Thus, JudgeLM enables pairwise comparisons of CNs, addressing the subjectivity of the task by breaking it down into simpler binary classification problems. To ensure the generalizability of our approach for the task, we test it on two distinct corpora: CONAN (Chung et al., 2019) and CONAN-MT (Fanton et al., 2021a). Using various models to generate texts of different quality means that we can evaluate model performance across a spectrum of CN quality, ensuring that even subtle distinctions between good texts can be captured. This approach ultimately aims for a higher correlation to human preference than traditional metrics and ranks models based on their ability to generate high-quality CNs.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9572–9585
November 12-16, 2024 ©2024 Association for Computational Linguistics

HS
Muslims do not have anything useful that can enrich our culture.

Candidate CNs
1. How about the money they contribute to our economy, their expertise and knowledge, their culture, tasty food. Should I go on?   2. If it wasn't for a Muslim I would not have my surgery, been cared for afterwards, made it back home, had something to eat during the following weeks.

Table 1: Example of a HS and two reference CNs, taken from the CONAN corpus. The CNs express diverse points, and show the variability and diversity of arguments used to combat HS.

As an additional contribution, we evaluate the inherent ability of LLMs as zero shot (ZS) CN generators. By leveraging state-of-the-art open-source LLMs we seek to explore their potential in generating CNs that effectively challenge and mitigate the influence of misinformation and harmful narratives. We examine three variants within the same model family: base, instruction-tuned, and chat-aligned. This enables us to inspect their unique strengths and limitations to determine the optimal choice for the task. Finally, we fine-tune the models on HS-CN pair data to compare their performance against ZS performance, assessing whether fine-tuning offers any significant improvement in our scenario. We conclude that chat-aligned models in a ZS setting are the best option for carrying out the task, provided they do not refuse to generate an answer due to security concerns[1]. Code is available at https://github.com/hitz-zentroa/cn-eval.

2 Related Work

In recent years, automatic CN generation has attracted growing research interest, with numerous methods leveraging NLG technologies for generating CNs. Nearly all recent systems depend on LLMs to automatically generate CNs (Ashida and Komachi, 2022; Tekiroğlu et al., 2022; Saha et al., 2024) , driven by their impressive performance in generation tasks, which often necessitates minimal or no training data (Zhao et al., 2023a; OpenAI et al., 2024; Zhao et al., 2023b).

Several datasets have been introduced to aid in the advancement of CN generation. The first large-scale, multilingual, expert-based dataset, Counter Narratives through Nichesourcing (CONAN) (Chung et al., 2019), consists of HS-CN pairs in English, French, and Italian, focusing exclusively on Islamophobia. This corpus served as the foundation for the development of MultiTarget CONAN (MT-CONAN) (Fanton et al., 2021a), which includes 8 hate-speech targets such as women and individuals with disabilities. Additionally, the DIALOCONAN dataset (Bonaldi et al., 2022), which contains fictitious dialogues between a hater and a Non-Governmental Organization (NGO) operator, and the Knowledge-grounded Hate Countering dataset (Chung et al., 2021), featuring HS-CN pairs with the background knowledge used for constructing the CNs have been introduced. Some work in adapting these corpora to other languages has also been done, such as CONAN-EUS (Bengoetxea et al., 2024), a Basque and Spanish translation of the original CONAN dataset, and CONAN-MT-SP (Vallecillo Rodríguez et al., 2024), a Spanish version of MT-CONAN.

Assessing the impact and effectiveness of the generated CNs remains a crucial aspect of this research domain. Evaluating CN is particularly challenging because there are many acceptable answers to a given HS, and is often very difficult to assess what constitutes a good answer. Evaluation is usually done either through automatic or manual methods. Automatic methods involve the use of metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020), which are standard evaluation metrics in tasks such as Machine Translation or Text Summarization. However, these metrics are known to be often weakly correlated with human judgment (Sai et al., 2022), particularly on tasks that require creativity (Ni'mah et al., 2023). Other automatic metrics that have been used for NLG evaluation including Repetition Rate (RR) (Bertoldi et al., 2013), which measures diversity in the generated answers, or Novelty (Wang and Wan, 2018), which encourages the model to generate answers that differ from the text in the training data. We must note that none of the aforementioned metrics take the input

[1]Chat-aligned models, designed to adhere to safety and ethical guidelines, may sometimes decline to respond to certain prompts. This refusal is typically in place to prevent the generation of harmful, inappropriate, or sensitive content.

Dataset	HS-CN Pairs	Unique HS	Unique CN	Avg. CNs / HS	Avg. Words / CN
CONAN	6648	523	4040	12.71	19.48
MT-CONAN	5003	3718	4997	1.35	24.77

Table 2: Statistics of the CONAN and MT-CONAN corpora, showing the number of HS-CN pairs, the number of unique HS and CN instances, the average number of CN per HS, and the average number of words per CN.

HS into account, missing a crucial aspect in the evaluation of CNs. As far as we know, no previous work has analyzed the correlation of these metrics with human judgment for CN evaluation.

Due to the limitations of automatic metrics, final assessments frequently rely on manual evaluations. Nonetheless, manual evaluation is often a costly process, and finding evaluators with adequate task comprehension can be challenging. Moreover, the subjective nature of the task adds another layer of complexity to the evaluation process. To mitigate this subjectivity, various efforts have been made to identify key aspects that assess the quality of CNs. Unfortunately, consensus on these key aspects is still lacking, as different authors consider factors such as relatedness, specificity, richness, coherence, grammaticality, suitableness, informativeness, diversity, relevance, language quality, offensiveness or stance (Chung et al., 2021; Ashida and Komachi, 2022; Bengoetxea et al., 2024).

Recently, to address the limitations of traditional metrics and manual evaluation, LLMs are being employed to directly assess the quality of generated text. Leveraging LLMs to measure NLG has been shown to exhibit stronger correlation with human assessment compared to conventional reference-based evaluation (Nimah et al., 2023; Chiang and Lee, 2023; Wang et al., 2023; Liu et al., 2023), and it remains an efficient and automated approach. Some works include using commercial models such as GPT-4 (Wang et al., 2023; Liu et al., 2023), while others focus on training specialized models for evaluation tasks, resulting in tools like PandaLM (Wang et al., 2024), JudgeLM (Zhu et al., 2023), and UniEval (Zhong et al., 2022). These LLMs can be either used to ascertain quantitative aspects to measure the quality of the generated text (Zhong et al., 2022; Ke et al., 2022), or to show preferences between two texts generated by different systems.

There are few works on automatic CN evaluation. Contemporary to our work, (Jones et al., 2024) use LLMs to evaluate CNs based on five different aspects considered relevant for their effectiveness in combating HS. In a strategic shift, we propose to use LLMs to compare outputs from different models between them and ultimately obtain a ranking of the best models for CN generation that correlates with human judgments.

3 Methodology

This section provides an overview of the key components of the research methodology. Section 3.1 discusses the specific models that were used for CN generation. Section 3.2 presents the corpus used in the study. Finally, Section 3.3 outlines the metrics that were employed to carry out the evaluation.

3.1 Models

We use publicly available auto-regressive models for CN generation. Specifically, we work with three variants of the Mistral model family (Jiang et al., 2023) as well as the Llama 2 Chat model (Touvron et al., 2023). The Mistral variants include the Mistral base model, the Mistral-Instruct model, and Zephyr (Tunstall et al., 2023), which is a chat-aligned model based on Mistral. The selection of these three variants enables the comparison of chat-aligned, instruction-tuned, and base model performance and behavior. Llama 2 was selected as to compare the results with Mistral models due to its relevance in the field and potential for providing complementary insights into CN generation. All models are 7B parameter models, consistent with the available Mistral model size, to ensure comparability of results. The specific versions of the employed models are listed in Table 3.

3.2 Corpus

In order to test the generalizability of our method, we experiment on two distinct datasets: Counter Narratives through Nichesourcing (CONAN) (Chung et al., 2019) and Multi-Target CONAN (MT-CONAN) (Fanton et al., 2021b). Corpus statistics are presented in Table 2.

CONAN Comprises HS-CN pairs addressing Islamophobia in three languages: English, Italian, and French. These pairs were collected through

Model	Version	Type	lr_{CONAN}	$\text{lr}_{\text{MT-CONAN}}$
mistral	v0.1	Base	1e-5	1e-4
mistral-instruct	v0.2	Instruct	6e-6	3e-5
zephyr	Beta	Chat	6e-6	1e-4
llama-chat	llama 2	Chat	2e-5	1e-3

Table 3: Information regarding the models, along with the learning rates that resulted in the lowest perplexity in the validation set for each one. It's worth noting that the optimal learning rates varied across different corpora: lr_{CONAN} are the optimal learning rates when fine-tuning on CONAN and $\text{lr}_{\text{MT-CONAN}}$ when fine-tuning on MT-CONAN.

nichesourcing involving 3 different NGOs from the United Kingdom, France, and Italy. As a result, the CNs are expert-based and crafted by operators specifically trained to combat online HS. After the data collection phase, three non-expert annotators were hired to augment the dataset. They paraphrased original hate content to increase pairs per language and translated content from French and Italian to English for language parallelism. NGO trainers validated the newly generated data for each language to ensure quality. In this work, we only focus on the English partition. During fine-tuning, we used 4833 pairs for training, 537 for validation, and 1278 for testing. The specific train-val-test splits are available at `https://huggingface.co/datasets/HiTZ/CONAN-EUS`.

MT-CONAN Consists of HS-CN pairs in English, collected through a Human-in-the-Loop approach. This method involves iteratively refining a generative language model by utilizing its own data from previous loops to generate new training samples, which are then reviewed and/or post-edited by experts. The HS targets eight distinct demographics: individuals with disabilities, Jewish people, the LGBT+ community, migrants, Muslims, people of color, women, and other marginalized groups. During fine-tuning, we used 3003 pairs for training, 1000 for validation, and 1000 for testing.

3.3 Evaluation Metrics

For evaluation we used both reference-based and reference-free metrics. Additionally, we incorporated the use of a Judge Model as part of our proposed evaluation methodology.

Reference-Based Metrics Based on previous work on CN generation, we opted to use BLEU, ROUGE-L and BERTScore. BLEU is a precision-based metric that measures the similarity between a candidate text and one or more reference texts. It computes the geometric mean of modified n-gram precision and applies a brevity penalty to discourage short translations. BLEU is widely used in machine translation tasks. ROUGE-L, on the other hand, focuses on the recall of content units. It calculates the longest common subsequence between the candidate sequence and the reference sequence, normalizing by the length of the reference sequence. ROUGE-L is commonly employed in text summarization tasks. BERTScore leverages contextual embeddings from pre-trained BERT models to compute the similarity between candidate and reference sentences. It computes the score based on the cosine similarity between BERT embeddings, providing a measure of semantic similarity. BERTScore has demonstrated effectiveness across various NLG tasks, including machine translation and text summarization. We report BERTScore F1 in our experiments.

Reference-Free Metrics For reference-free metrics, we opted to use Repetition Rate (RR) (Bertoldi et al., 2013), which is computed by calculating the non-singleton n-grams that are repeated in the generated text (Bertoldi et al., 2013) and Novelty (Wang and Wan, 2018) that is computed by calculating the non-singleton n-grams from the generated text that appear in the train data. While RR aims to capture the diversity in the generated text, Novelty measures how different the generated text is from the training data. It should be noted that Novelty is less valuable when evaluating models that were used in a ZS setting, as there is no training involved.

LLM-Based Evaluation Finally, we consider the use of JudgeLM as an evaluator. JudgeLM is a scalable Judge Model based on Vicuna that was designed to evaluate LLMs in open-ended scenarios. It was trained using a large-scale dataset consisting on LLM-generated answers for diverse NLG tasks and detailed judgments from GPT-4. Remarkably, it achieves an agreement rate exceeding 90% in some tasks, surpassing even human-to-human agreement levels (Zhu et al., 2023). While

JudgeLM supports different evaluation methods, such as comparing single answers against a reference or comparing multiple answers simultaneously, we decided to use it to compare generated CNs pairwise, as described in Section 4.1. This eliminates the problem of needing a reliable reference and instead focuses on determining which of the available options is the best, simplifying the task. Comparing the CNs against each other also avoids the ambiguity of the open-ended scenario we would face if we decided to evaluate them individually. We must note that JudgeLM also considers the instance of HS when comparing CNs, and thus evaluates CN effectiveness in relation to a specific instance of HS, instead as standalone sentences[2].

JudgeLM operates in two modes: normal mode and fast evaluation mode. In fast evaluation mode the model outputs two scores, one for each CN, providing an overall assessment of their value. In normal mode the model supplements these scores with arguments explaining the rationale behind them. We used the normal mode both during the development stage and while conducting the result analysis, to ensure a comprehensive evaluation. However, we activated fast mode when creating the ranks, which are thus calculated based exclusively on the output scores. This decision was made because generating arguments significantly increases inference time and argumentation was deemed unnecessary for the ranking task.

4 Evaluation Framework

In this section we present a cost-effective pairwise rank-based evaluation paradigm designed to assess the performance of CN generation systems in alignment with human preference. The method is detailed in Section 4.1. In addition, in Section 4.2, we describe the additional manual evaluation performed to provide a detailed assessment according to various relevant aspects contributing to the effectiveness of a CN, as presented in (Bengoetxea et al., 2024).

4.1 Pairwise Rank-Based Evaluation

We propose an *"A vs B"* comparative setup to rank models with respect to their CN generation skills. Suppose we have n models to rank and we want to evaluate their performance in a set consisting of h HS instances. We first generate h CNs, one for each HS instance, with each of the n models. After that, the generated $h \cdot n$ CNs are pitted against each other in *"A vs B"* tournaments, with a total of $\binom{n}{2} \cdot h = \frac{n!}{2!(n-2)!} \cdot h$ tournaments. Each tournament therefore comprises one HS and two CNs generated from different models, and is evaluated (either by automatic or manual methods) to assess which of the CN better answers the given HS. The model that produced the best CN receives 1 point, while the losing model receives 0 points. In the case of ties, both models receive 0.5 points. Models receive a total score by aggregating the scores of the tournaments they participate, and they are finally ranked based on the obtained score. In the following, we describe details of the manual and automatic evaluation methods, respectively.

Automatic Evaluation For automatic evaluation we prompt JudgeLM to output 2 scores: one for each proposed CN. The winner is determined by comparing these scores: the CN with the highest score is the winner. If both scores are the same, it results in a tie.

In our experimentation, all the models in Table 3 were evaluated in both CONAN and MT-CONAN. Combining their test sets results in 2278 HS instances. We evaluated both the fine-tuned and ZS versions of each model, along with the gold standard. This resulted in a total of $\binom{9}{2} \cdot 2278 = 82008$ tournaments.

Manual Evaluation For manual evaluation, we had 3 annotators decide which of the 2 proposed CNs they believe would be more effective in combating the presented instance of HS. The specific guidelines provided to the annotators are detailed in Appendix A.1.

In our experimentation, given the significant cost associated with manual evaluation, 10 HS instances from each of the test sets of CONAN and MT-CONAN were randomly selected. This resulted in $\binom{9}{2} \cdot 20 = 720$ tournaments. From the mix of both corpora, 288 tournaments (144 from each corpus) were evaluated by all 3 evaluators to calculate the inter-annotator agreement (IAA) using Cohen's Kappa. For instances from CONAN, the mean IAA was 0.42, while for the instances of CONAN-MT, it was 0.58. The individual coefficients are presented in Appendix B. The final outcome of these tournaments was decided using majority voting to reduce subjectivity. The remaining 432 tournaments were each annotated by a single annotator, with each

[2] Appendix C.5 shows the template used to prompt JudgeLM.

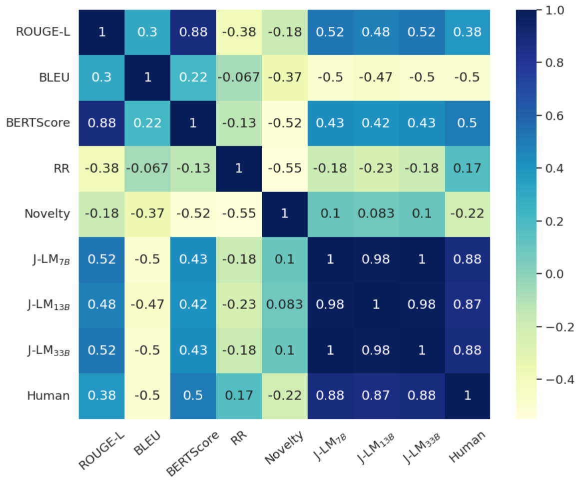

Figure 1: Matrix with the Spearman's rank correlation coefficients among metrics. The last row of the matrix represents the correlation of all the evaluation metrics to human preference. *J-LM* is short for JudgeLM.

annotator evaluating 144 tournaments.

4.2 Feature Evaluation

For a comprehensive evaluation of CN quality, we provide a manual assessment based on five criteria: Relatedness, Specificity, Richness, Coherence, Grammaticality and overall score (Bengoetxea et al., 2024). Detailed evaluation guidelines can be found in Appendix A.2. The evaluation involved 2 annotators on 90 HS-CN pairs, which included 10 HS instances from the CONAN corpus and the corresponding CNs generated by each of the 9 models. Each evaluated feature used a five-point scale for the answers, with 1 as the lowest score and 5 as the highest. The mean IAA across all the evaluated features is 0.69. The feature-wise IAA is presented in Appendix B.

5 Experimental Setup

This section outlines the specific setup employed for the generation of CNs. We consider two generation scenarios: ZS inference (see Section 5.1), and inference with fine-tuned models (see Section 5.2).

5.1 Zero-Shot Generation

In the ZS approach we prompt all models with the sentence *"Provide a brief counter-narrative in response to the user's hate speech. Ensure the output does not contain line breaks"*. This prompt was selected based on preliminary experimentation. For each model, the prompt was formatted differently, as each one has its specific prompting format. The various formats are detailed in Appendix C.

5.2 Fine-Tuning Details

Instead of fine-tuning the whole model, Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) was used. This approach facilitated a faster and more accessible training process, as it significantly reduces hardware requirements. The model was loaded in 4 bit with *NF4* quantization data type and *bf16* computational data type. The LoRA update matrices were applied to the attention blocks and bias parameters were not trained. The LoRA rank and scaling factor were set to 16 and the dropout to 0.05. These values were chosen based on experimentation, guided by those reported in the literature (Dettmers et al., 2023; Hu et al., 2021), with minimal observed impact on results. Following usual practice, we used Adam optimizer with a inverse square root scheduler, half precision, and a batch size of 32. A set of learning rates ranging from $1e^{-6}$ to $1e^{-3}$ were tested, and the one yielding the lowest perplexity in the validation set was selected for each model. The selected learning rates are listed in Table 3. The models were set to train for a maximum of 10 epochs, with early stopping and a patience of 3 epochs. The checkpoint with the lowest valiadation loss was selected in each case. Additionally, at inference time, generation was terminated upon encountering the newline token (\n) to ensure the production of shorter sentences, addressing the issue of role-playing commonly observed in many models and particularly prominent in base models due to their challenges in effectively interpreting prompts.

6 Results

First, in Section 6.1, we discuss the correlation between the metrics in Section 3.3 and human preference, highlighting the implications of the findings. Then, in Section 6.2, we present the model performance ranking for the CN generation task derived from the Pairwise Rank-Based evaluation method (see Section 4).

6.1 Correlation of Automatic Metrics with Human Ratings

Figure 1 illustrates the Spearman's rank correlation coefficients (ρ) among all metrics, including human evaluation. The rankings for *Human* and *J-LM* (JudgeLM) are computed using the pairwise comparison setting described in Section 4, whereas the rest of the rankings (*BLEU*, *ROUGE-L*, etc)

Rank	Human	*Score*	JudgeLM$_{33B}$	*Score*
1	zephyr$_{zs}$	18.02	zephyr$_{zs}$	20.20
2	gold standard	17.60	mistral-instruct$_{zs}$	16.09
3	mistral-instruct$_{zs}$	14.80	gold standard	8.98
4	zephyr$_{ft}$	11.59	zephyr$_{ft}$	13.30
5	mistral$_{zs}$	10.75	llama-chat$_{zs}$	11.07
6	mistral$_{ft}$	9.08	mistral$_{zs}$	9.05
7	mistral-instruct$_{ft}$	7.54	mistral$_{ft}$	8.70
8	llama-chat$_{zs}$	7.26	mistral-instruct$_{ft}$	8.50
9	llama-chat$_{ft}$	3.35	llama-chat$_{ft}$	4.11

Table 4: Comparison of human and JudgeLM rankings, including the final scores obtained from the pairwise tournaments. zs means that the model was used in a zero-shot setting.

are based on their respective metric scores. All rankings are established across 720 tournaments, as described in Section 4.1.

The figure shows a strong correlation between all variants of JudgeLM and human preference, as depicted in the last row/column of the matrix, with both the 7B and 33B parameter JudgeLM achieving a ρ of 0.88. This high correlation is supported by a statistically significant Pearson correlation coefficient of 0.73 between JudgeLM (33B version) and human preference (p-value of 0.03)[3].

On the contrary, traditional metrics correlate poorly with human preference, with the highest ρ being the 0.50 obtained by BERTScore. These results confirm that commonly used automatic metrics lack alignment with human preference when evaluating the quality of CNs. Not unsurprisingly, the correlation between traditional metrics and JudgeLM is also low. They also correlate poorly among themselves with the exception of ROUGE-L and BERTScore, which attain a ρ of 0.88. Despite both being based on n-gram overlap, ROUGE-L and BLEU only achieve a ρ value of 0.3.

The aforementioned observations are further reinforced in Appendix D, where the correlation matrices on the CONAN corpus and the MT-CONAN corpus are presented separately. In said appendix, we once again see a strong correlation between JudgeLM variants and human preference, whereas the correlation with traditional metrics is weak and inconsistent, showing no predictable pattern.

As the concluding point of the correlation analysis, Table 4 presents a comparison of the final rankings obtained through the manual and automatic pairwise comparisons as described in Section 4. Both the automatic and the manual method assign similar scores to almost all systems, with the exceptions of the gold standard, which obtains a considerably higher score when evaluated by humans than by JudgeLM, and llama-chat$_{zs}$, where JudgeLM assigns it a higher rank than humans. In any case, their final position in the rank only varies slightly among methods. By analyzing examples of discrepancies between human and JudgeLM judgements, we observe that the disagreement in the case of the gold standard might stem from the fact that the JudgeLM model prefers longer, more detailed CNs, while the annotators preferred shorter, more direct ones. It might also be related to the fact that the model cannot discern false information from true information, whereas the human evaluator can penalize non-factual content resulting in the simpler but veracious CN winning. Instead, in the case of llama-chat$_{zs}$, the disagreement might be because JudgeLM favors answers that start by stating that they can not endorse in hate speech (*"I apologize, but I cannot fulfill your request. I'm just an AI and it's not within my programming or ethical guidelines to provide counter-narratives that promote hate speech... Is there anything else I can help you with?"*). This preference may stem from its training on evaluations made by ChatGPT, which often responds in a similar manner when asked to provide CNs to HS.

6.2 Ranking by Pairwise Comparison

Figure 2 depicts the ranking of CN generation systems based on the tournament outcomes according to different sizes of JudgeLM, evaluated across the 82008 tournaments that comprise the entire test set. Overall, in the ZS scenario chat-aligned models exhibit superior performance, followed by the instruction-tuned model, while the base model

[3]We calculate Pearson correlation using the performance scores obtained by each model in the Pairwise Rank-Based evaluation.

System	Relatedness	Specificity	Richness	Coherence	Grammaticality	Overall
zephyr$_{zs}$	4.95	4.25	4.00	5.00	5.00	4.25
gold standard	4.10	3.75	3.25	4.80	4.30	3.50
mistral-instruct$_{zs}$	4.20	3.15	3.70	4.70	5.00	3.50
llama-chat$_{zs}$	2.90	2.55	4.30	4.90	5.00	3.05
mistral-instruct$_{ft}$	3.75	3.55	3.30	3.10	4.30	2.70
mistral$_{ft}$	3.65	3.55	3.05	3.30	4.35	2.60
zephyr$_{ft}$	4.40	4.75	3.60	3.20	4.35	2.30
llama-chat$_{ft}$	3.40	3.10	2.95	3.30	4.10	2.20
mistral$_{zs}$	3.10	3.30	2.40	3.55	4.60	1.90

Table 5: Evaluation of the different aspects that contribute to the effectiveness of a CN. The values in the table represent the average of the scores assigned by each annotator.

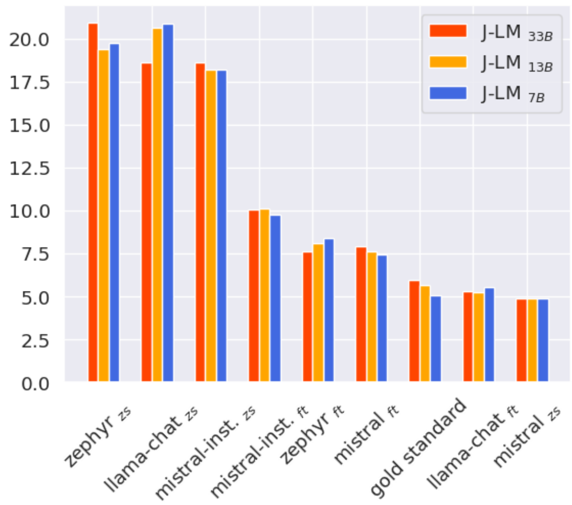

Figure 2: Ranking through pairwise comparison based on evaluations of all the JudgeLM size variations across the entire test set.

demonstrates the lowest performance. This outcome is expected, as base models lack training to understand instructions and have no prior experience in the task, whereas chat models, in addition to being capable of understanding instructions, are often trained to fight toxicity through Safety Fine-Tuning (Touvron et al., 2023; OpenAI et al., 2024). When fine-tuning the models, we observe a decline in performance across all models, except for the base model, which exhibits a considerable improvement. The decline in performance is more pronounced in the chat-aligned models than in the instruction-tuned model.

When examining the rankings according to the different sizes of JudgeLM, we observe that as the model size increases, llama-chat$_{zs}$ is positioned lower, thereby narrowing its performance gap with Zephyr$_{zs}$.

7 Analysis

To confirm which of the models from Section 3.1 is the best for the task, we undertook a final feature-wise evaluation as explained in Section 4.2. The results are presented in Table 5. As seen there, the best-performing model is undoubtedly Zephyr, which considerably surpasses the gold standard.

Analyzing Table 4 in more detail, which presents the final rankings obtained from manual and automatic pairwise comparisons, we note that while manual evaluators were instructed to select a winning CN unless both were deemed ineffective, JudgeLM assigns ties when both responses are of high quality. This approach may result in a lower correlation between JudgeLM ratings and human preference. Upon a more nuanced analysis, we also observed that JudgeLM demonstrates a distinct preference for factual CNs, particularly those offering detailed and specific information about research and related topics. However, the model lacks the capability to reliably detect hallucinations, meaning it may favor responses that appear factual without verifying the truthfulness or accuracy of the provided details.

Lastly, regarding fine-tuning, we observed that including factual CNs in the fine-tuning process for our CN generation scenario might not be advisable, as models may mimic the structure but lack factual accuracy.

8 Conclusions

CN generation needs a different evaluation framework and metrics than those used in previous work (Chung et al., 2021; Tekiroğlu et al., 2022; Bengoetxea et al., 2024). This is due to the unique objectives, complexities, and impact of CNs, which require specialized criteria to assess their effectiveness and quality accurately. Thus, developing and implementing tailored evaluation metrics is crucial to advance the field and ensure the successful creation of impactful CNs.

Consistent with previous research observations, traditional metrics fall short in evaluating generation tasks that require creativity, including CN generation to combat HS. In this paper we present a promising LLM-based ranking method to provide an alternative automatic evaluation technique which exhibits higher correlation with human evaluation.

Limitations

Our work still has some open research questions which can be summarized in the following limitations. First, we have not addressed truthfulness. Thus, JudgeLM rewards CNs that provide factual arguments without considering whether they are truthful. Second, additional tests on larger corpora could be performed to determine whether the lack of improvement from fine-tuning in chat and instruct models is due to limitations in the corpus itself.

The corpus used in our experiments was small and, as indicated in Table 2, exhibited significant repetition of certain HS instances giving them a different CN each time. We hypothesize that this data structure may potentially have adverse effects in model performance. Thus, we performed a preliminary fine-tuning experiment that involved randomly removing duplicate entries from the corpus, resulting in a smaller but cleaner dataset. Despite the dataset being smaller, the performance did not degrade. This initial investigation suggests that reducing duplications could lead to more consistent learning outcomes.

In future work, we aim to extend this analysis to other languages such as Spanish, along with Basque, which is considered a low-resource language. Finally, we plan to explore Retrieval Augmented Generation (RAG) to address the truthfulness issue, as we anticipate that this approach could substantially enhance the correlation between human evaluations and those of Judge Models.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22). We are also thankful to the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. (iv) DeepR3 (TED2021-130295B-C31) and European Union NextGenerationEU/PRTR.

References

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and spanish counter narrative generation: Data creation and evaluation. In *LREC-COLING*.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Annual Meeting of the Association for Computational Linguistics*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. 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)*. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Iftitahu Ni'mah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Annual Meeting of the Association for Computational Linguistics*.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Torino, Italia. ELRA and ICCL.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

María Estrella Vallecillo Rodríguez, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. 2024. CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688, Torino, Italy. ELRA and ICCL.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen

Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023b. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15590–15606, Toronto, Canada. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges.

A Manual Evaluation Guidelines

We carry out two kinds of manual evaluation: rank-based evaluation (see Section A.1) and feature-based evaluation (see Section A.2).

A.1 Rank Evaluation Guidelines

We will present an instance of HS followed by two candidate CNs: CN$_A$ and CN$_B$. Participants will choose which CN they find more effective in countering the HS. If both CNs are equally unsatisfactory, participants can declare a tie. Ties will only be applicable when both CN are deemed inadequate in addressing the HS. Responses lacking specificity and informative content will incur penalties, as will answers containing false information.

Instructions for Annotators

- Carefully read the provided instance of hate speech.
- Evaluate Counter-Narrative$_A$ and Counter-Narrative$_B$ based on their effectiveness in addressing and countering the hate speech.
- Choose the counter-narrative you find more effective. If both are equally ineffective, declare a tie.
- Consider the specificity and informative content of each counter-narrative.
- Be vigilant for any false information in the responses, as these should be penalized.

A.2 Feature Evaluation Guidelines

An instance of HS and a CN designed to combat it will be provided. The quality of the CN will then be evaluated based on the following criteria:

Relatedness *Is the CN related to the HS?*

- 0: No
- 1: Barely
- 2: Somewhat
- 3: More or less
- 4: Mostly
- 5: Yes

Specificity *Does the CN provide detailed and precise information?*

- 0: Not specific at all
- 1: Barely specific
- 2: Somewhat specific
- 3: Moderately specific
- 4: Quite specific
- 5: Very specific

Richness *Does the CN include a variety of vocabulary and sentence structures?*

- 0: Very poor vocabulary and structure
- 1: Barely rich
- 2: Somewhat rich
- 3: Moderately rich
- 4: Quite rich
- 5: Very rich

Coherence *Is the CN logically organized and easy to understand?*

- 0: Not coherent at all
- 1: Barely coherent
- 2: Somewhat coherent
- 3: Moderately coherent
- 4: Quite coherent
- 5: Very coherent

Grammaticality *Is the CN grammatically correct and free of errors?*

- 0: Completely ungrammatical
- 1: Barely grammatical
- 2: Somewhat grammatical
- 3: Moderately grammatical
- 4: Quite grammatical
- 5: Completely grammatical

Overall Score *How suitable is the CN as a response?*

- 1: Not suitable (borderline hate speech)
- 2: Makes some acceptable points but not suitable
- 3: Would be suitable with some modifications
- 4: Good, though minor corrections may be needed
- 5: Very good as a CN

B Inter-Annotator Agreement

In this section, we show the tables of IAA from the evaluation process. These tables highlight the consistency among annotators and the reliability of the annotations. Figure B.1 shows the agreement for the Pairwise Rank-Based evaluation method presented in Section 4.1, and Table B.1 for the feature evaluation explained in Section 4.2.

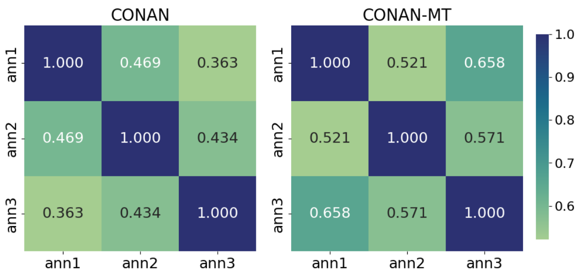

Figure B.1: IAA of the Pairwise Rank-Based evaluation.

C Prompting

This section presents the exact formats used to prompt each of the models in Section 3.1.

Feature	IAA
Relatedness	0.6397
Specificity	0.5947
Richness	0.7806
Coherence	0.8270
Grammaticality	0.7087
Overall	0.6183

Table B.1: IAA of the feature-wise evaluation.

C.1 Mistral Prompt Format

Provide a brief counter-narrative in response to the user's hate speech. Ensure the output does not contain line breaks.
###Input:
Hate Speech
###Output:
Counter-Narrative

C.2 Mistral-Instruct Prompt Format

<s>[INST] Provide a brief counter-narrative in response to the user's hate speech. Ensure the output does not contain line breaks. *Hate Speech* [/INST] *Counter-Narrative*</s>

C.3 Zephyr Prompt Format

<|system|>
Provide a brief counter-narrative in response to the user's hate speech. Ensure the output does not contain line breaks.</s>
<|user|>
Hate Speech</s>
<|assistant|>
Counter-Narrative</s>

C.4 Llama-Chat Prompt Format

<s>[INST] «SYS»
Provide a brief counter-narrative in response to the user's hate speech. Ensure the output does not contain line breaks.
«/SYS»*Hate Speech* [/INST] *Counter-Narrative* </s>

C.5 JudgeLM Prompt Format

You are a helpful and precise assistant for checking the quality of the answer.
[Question]
Provide a brief counter-narrative in response to the

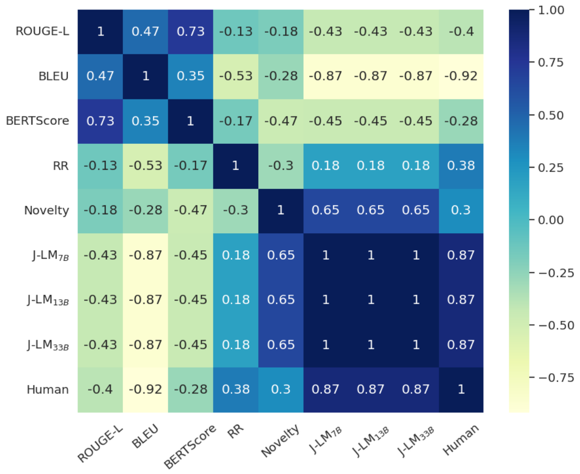

Figure D.1: Matrix with the Spearman's rank correlation coefficients among metrics, created using 360 tournaments from CONAN. The last row of the matrix represents the correlation of all the evaluation methods to human preference.

following hate speech: *Hate Speech* It is very important for the counter-narrative to be relevant to the hate speech.
[The Start of Assistant 1's Answer]
Counter-Narrative$_A$
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
Counter-Narrative$_B$
[The End of Assistant 2's Answer]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

D Correlation Matrix

Correlation between the metrics presented in Section 3.3 and human preference in the CONAN corpus (Table D.1) and in the MT-CONAN corpus (Table D.2).

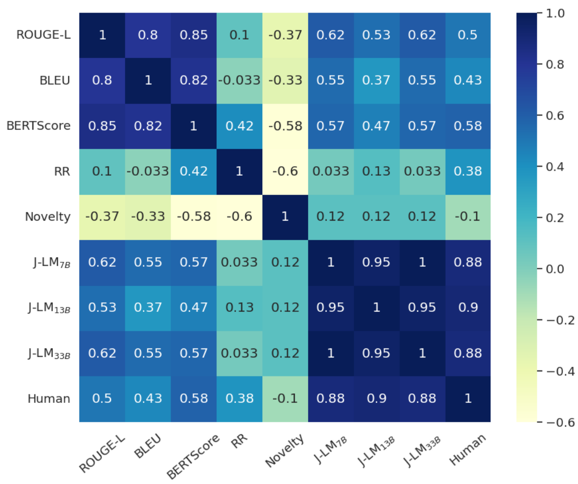

Figure D.2: Matrix with the Spearman's rank correlation coefficients among metrics, created using 360 tournaments from MT-CONAN. The last row of the matrix represents the correlation of all the evaluation methods to human preference.