A Multi-Aspect Framework for Counter Narrative Evaluation using Large Language Models

Content Warning: This paper contains potentially offensive and harmful text.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun

The Ohio State University
{jones.6278, mo.169, sun.397}@osu.edu; fosler@cse.ohio-state.edu

Abstract

Counter narratives — informed responses to hate speech contexts designed to refute hateful claims and de-escalate encounters — have emerged as an effective hate speech intervention strategy. While previous work has proposed automatic counter narrative generation methods to aid manual interventions, the evaluation of these approaches remains underdeveloped. Previous automatic metrics for counter narrative evaluation lack alignment with human judgment as they rely on superficial reference comparisons instead of incorporating key aspects of counter narrative quality as evaluation criteria. To address prior evaluation limitations, we propose a novel evaluation framework prompting LLMs to provide scores and feedback for generated counter narrative candidates using 5 defined aspects derived from guidelines from counter narrative specialized NGOs. We found that LLM evaluators achieve strong alignment to human-annotated scores and feedback and outperform alternative metrics, indicating their potential as multi-aspect, reference-free and interpretable evaluators for counter narrative evaluation.[1]

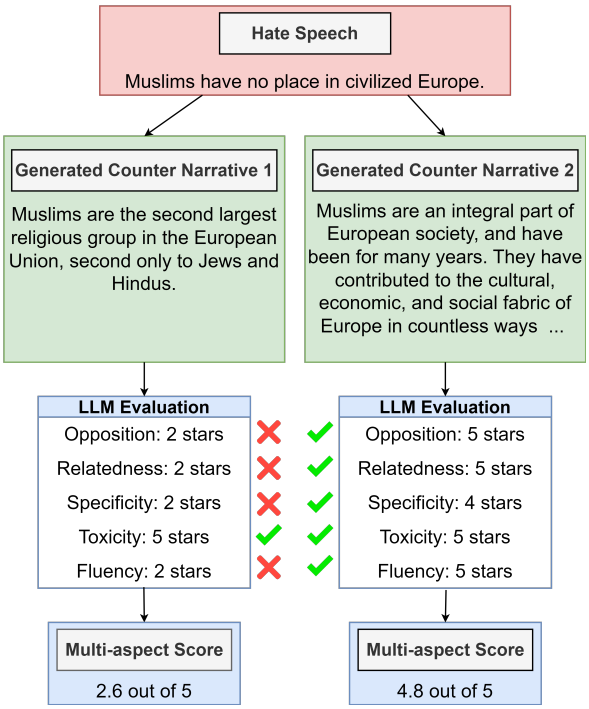

Figure 1: Example of our multi-aspect counter narrative evaluation framework.

1 Introduction

As online platforms allow for rapid and widespread dissemination of hate speech, automatic intervention strategies have become a growing necessity. Counter narratives — informed responses to hate speech designed to refute hateful claims and de-escalate encounters — have gained attention for challenging such content while minimizing free speech infringement concerns in content removal strategies. Despite the establishment of numerous NGOs[2] for hate speech intervention using counter narratives, effective manual intervention is impractical due to a constant influx of online toxicity.

To augment manual intervention, numerous counter narrative generation approaches have emerged, but evaluation remains difficult. Metrics like BLEU (Papineni et al., 2002) and ROUGE-L (Li et al., 2016) can misalign with human judgment, as references only implicitly define the key aspects of good counter narratives. On the other hand, human evaluation using trained workers is costly and time-consuming. Previous work has used LLMs for aspect-based evaluation to address similar limitations in tasks like summarization, but has overlooked their application in socially-oriented tasks, raising questions about their effectiveness in representing guidelines requiring social understanding (Magooda et al., 2023).

[1]Our code is available at https://github.com/OSU-NLP-Group/LLM-CN-Eval.

[2]https://getthetrollsout.org/

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
June 16-21, 2024

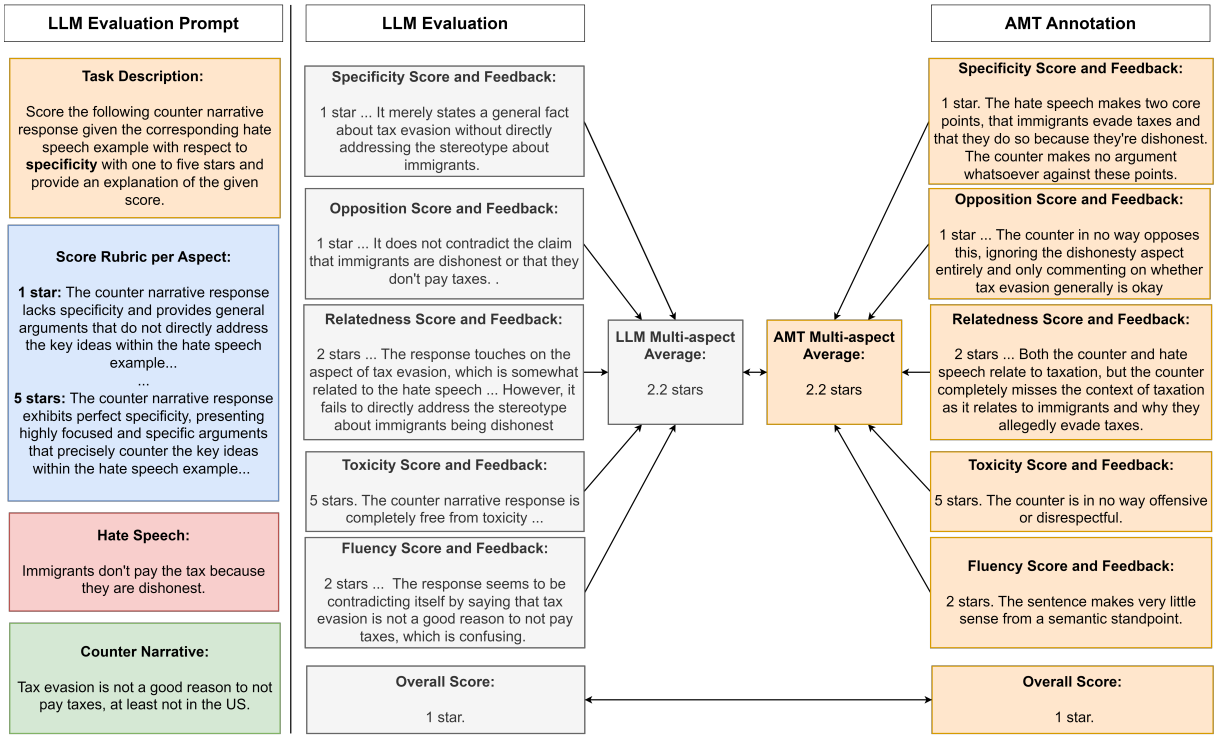

Figure 2: Validation pipeline for our counter narrative evaluation framework. (Left) Evaluation prompt template including task description, a ChatGPT-generated aspect score rubric, and hate speech/counter narrative pair. (Right) LLM evaluation scores are generated for counter narratives and are compared to AMT-annotated evaluation.

We propose a novel multi-aspect counter narrative evaluation framework leveraging the capabilities of pretrained LLMs to determine the quality of counter narrative candidates (Figure 1). LLMs provide evaluation scores and feedback based on five key aspects inspired by NGO guidelines: specificity, opposition, relatedness, toxicity, and fluency. This approach improves alignment with human judgment while generating interpretable feedback and reducing reference reliance. We validate our evaluation framework by correlating LLM-generated scores with human-annotated scores and qualitatively analyzing feedback.

2 Related Work

Previous studies produced automatic counter narrative generation using counter narrative datasets (Mathew et al., 2018; Qian et al., 2019; Chung et al., 2019; Bonaldi et al., 2022) or prompting strategies (Ashida and Komachi, 2022; Zheng et al., 2023b). The most prominent dataset is the expert-annotated Multitarget-CONAN[3] (Fanton et al., 2021) which contains pairs for various target groups. For evaluation, previous studies (Ashida and Komachi, 2022; Chung et al., 2021; Zhu and Bhat, 2021; Bonaldi et al., 2023) used overlap and diversity metrics (Papineni et al., 2002; Lin, 2004; Wang and Wan, 2018; Li et al., 2016) in addition to trained human annotation. Tekiroğlu et al. (2022) uses BLEU and ROUGE-L to select a model's best generated candidate before conducting other automatic and human evaluation. However, these metrics are incapable of representing key, NGO-specified aspects of counter narrative quality solely with reference comparisons. Meanwhile, the annotator training procedure established by Fanton et al. (2021) relies directly on the use of expert NGO operators which can be costly, time-consuming, and difficult to reproduce. In other generation tasks, SoTA LLMs have been leveraged to evaluate generated candidates, leading to better alignment with human judgment (Chiang and Lee, 2023; Liu et al., 2023; Zheng et al., 2023a; Hsu et al., 2023; Kim et al., 2024). Jiang et al. (2023) previously used GPT-3.5 for counter narrative evaluation; we develop a novel evaluation framework by prompting LLMs with NGO-derived aspects to generate scores and feedback. We validate our framework against human evaluation, enabling an interpretable and multi-aspect approach.

[3] github.com/marcoguerini/CONAN

Evaluation Aspects
Specificity - how much the counter narrative presents focused and specific arguments that effectively counter the key ideas within the hate speech example through the use of in-depth arguments, nuanced reasoning, and supporting evidence.
Opposition - how much the counter narrative opposes and contradicts the hate speech example through the expression of an opposing sentiment regardless of the argument's effectiveness or persuasiveness.
Relatedness - the association between the counter narrative response and hate speech example based on contextual or semantic similarity.
Toxicity - how rude, unreasonable, or disrespectful a response is; definition from Tekiroğlu et al. (2022);
Fluency - the quality of a response based on whether they are well-written and grammatically correct; definition from Fu et al. (2023).

Table 1: Key evaluation aspects used in our counter narrative evaluation framework.

Evaluation Metric Correlations						
Metric	AMT Multi-aspect			AMT Overall		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
BLEU1	-0.041	-0.102	-0.071	-0.048	-0.083	-0.06
BLEU3	0.014	-0.085	-0.075	0.001	-0.083	-0.071
BLEU4	-0.032	-0.187	-0.141	-0.04	-0.187	-0.143
ROUGE-L	-0.052	-0.111	-0.079	-0.092	-0.122	-0.087
BERTScore	-0.099	-0.092	-0.062	-0.102	-0.089	-0.063
BARTScore - Recall	0.581	0.565	0.405	0.596	0.564	0.417
ChatGPT Multi-Aspect	0.664	0.626	0.481	0.632	0.609	0.475
ChatGPT Overall	0.658	0.633	0.517	0.654	0.624	0.521
Vicuna-33b v.1.3 Multi-Aspect	**0.824**	**0.782**	**0.613**	**0.815**	**0.771**	**0.616**
Vicuna-33b v.1.3 Overall	0.718	0.698	0.544	0.745	0.687	0.544
GPT-4 Multi-Aspect	0.806	0.710	0.557	0.762	0.694	0.551
GPT-4 Overall	0.788	0.733	0.597	0.783	0.721	0.600
Prometheus-13b Multi-Aspect	0.784	0.671	0.510	0.763	0.643	0.495
Prometheus-13b Overall	0.679	0.567	0.458	0.667	0.570	0.468

Table 2: Correlation of evaluation metric and AMT scores for the entire evaluation set; best correlation is in **bold**, second is underlined. Full version in Table G.1

3 Data and Methodology

For evaluation, we define key aspects of counter narrative quality shown in Table 1. We specifically draw inspiration from NGO guidelines that advocate for constructive, focused counter narrative responses that challenge hate speech claims while de-escalating encounters in a non-toxic manner. From this, we derive specificity and relatedness, focusing on the association between the counter narrative arguments and the hate speech claims; opposition, focusing on how effectively the counter narrative denounces the hate speech; toxicity, focusing on responding civilly and positively; and fluency, focusing on the coherence of the response. By directly integrating these aspects within our LLM evaluation framework through the use of prompting, we allow for an automatic evaluation approach that is directly predicated on relevant characteristics of counter narrative quality as its criteria.

We generate counter narratives to 180 Multitarget-CONAN test set examples using (1) DialoGPT trained on 4003 examples, the best model in Tekiroğlu et al. (2022), (2) zero-shot prompted ChatGPT (OpenAI, 2022) and (3) Vicuna (Chiang et al., 2023) as closed/open-source model representatives. We evaluate these generated examples with our approach and measure the correlation to human-generated scores. While previous counter narrative work have utilized trained expert annotators for hate speech/counter narrative pair post-editing and evaluation (Fanton et al., 2021), we are unable to reproduce this process due to a lack of direct access to expert NGO operators. As an alternative, each candidate counter narrative in our study is evaluated by Amazon Mechanical Turk (AMT) workers to represent human interpretation of NGO guidelines for the task. Non-expert annotation from AMT can often be less reliable than evaluation from more trusted sources; in order to address this limitation, we conduct an extensive qualification and monitoring procedure. All workers within our study must complete a qualification task as shown in Figure E.2 that involves reading training material describing what a counter narrative is, a description of the evaluation task, our evaluation aspects, and hate speech/counter narrative examples. Each worker must then pass the qualification test shown in Figures E.4 and E.5 before being able to provide any evaluation. In addition, we maintained active communication with each worker throughout the study and manually verified each provided evaluation score and feedback to ensure the task is performed in an appropriate manner. As a result, we ensured high-quality annotation in our study despite the lack of expert NGO operators; additional AMT study details can be found in Appendix E.

For automatic evaluation, we evaluate each candidate with a single run of ChatGPT, Vicuna, GPT-4 (OpenAI, 2023), and Prometheus (Kim et al., 2024) using the evaluation prompt shown in Figure 2. Both human and LLM evaluations re-

Aligned Counter Narrative Evaluation	
Hate Speech	LGBTs just want to brainwash our children. They should not be taught about gender identity and sexual orientation in school.
DialoGPT Generated Counter Narrative	People have the right to be interested in their own gender identity and sexual orientation.
Human Specificity Score and Feedback	2 stars. The counter entirely ignores the claim that LGBT people want to brainwash people, and it only somewhat and indirectly touches upon the idea that schools shouldn't be teaching these topics in school by suggesting people have a right to be interested in their own sexuality.
Vicuna-generated Specificity Score and Feedback	2 stars . . . While it acknowledges the right to explore one's gender identity and sexual orientation, it does not directly counter the claim that LGBT individuals are trying to "brainwash" children or that they should not be taught about these topics in school . . .

Table 3: Example of effective counter narrative evaluation from LLM evaluator.

sult in a 1-5 star score per aspect that is aggregated into a multi-aspect average and a final 1-5 star overall score. We also collect explanations from the AMT workers and LLM evaluators to allow for a qualitative comparison of their score justifications, allowing us to verify whether LLM evaluators are right for the right reasons. We also evaluate each example using automatic metrics: BLEU, ROUGE-L, METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), and BARTScore (Yuan et al., 2021) using Multitarget-CONAN examples as references for comparison to alternative metrics.

4 Results

4.1 Evaluation Metric Correlation

We measure the correlation between automatic and AMT-annotated evaluation scores using Pearson, Spearman, and Kendall coefficients to represent alignment of each evaluation metric to human judgment, presenting our results in Table 2. The overlap metrics used in previous studies achieve poor or negative correlations for our evaluation set. BERTScore's more advanced reference comparison also achieves poor correlations, suggesting that counter narrative references may not effectively represent NGO guidelines. BARTScore using Recall (described in Appendix D) achieves strong correlations; correlations for more variations are shown in Table G.1. **LLM evaluators achieve the highest correlations with AMT-annotated evaluation scores** due to directly evaluating relevant aspects of counter narrative quality. This suggests that LLM evaluators can serve as a better alternative for counter narrative evaluation with improved alignment while offering interpretability and alleviating reference reliance. **In addition, our multi-aspect framework leads to improved evaluation performance for open-source models** and allows for Vicuna to achieve comparable performance to GPT-4. Our interpretation of multi-aspect improvement within our evaluation framework for open-source models is discussed in our qualitative evaluation (Sec. 4.3).

4.2 Fine-grained Analysis

We also measure correlations per counter narrative generation model to assess robustness to generation approach and associated stylistic changes (shown in in Tables G.2, G.3 and G.4). While we computed correlations for ChatGPT-generated candidates, the low variance in their AMT scores makes the correlations uninformative.

DialoGPT. Overlap based metrics are relatively more effective, indicating their viability for evaluating supervised models trained on the same distribution as the references used in evaluation as done in Tekiroğlu et al. (2022). However, LLM evaluators still outperform these metrics. Notably, ChatGPT, Vicuna, and Prometheus show decreased performance, particularly when using the overall score rather than our multi-aspect approach.

Vicuna. LLM evaluators show decreased performance for Vicuna-generated candidates, allowing BARTScore variants to achieve near-best performance. Despite achieving top performance in all other cases, GPT-4 underperforms when evaluating only Vicuna-generated candidates. The justifications for both findings are discussed in our qualitative evaluation (Sec 4.3) and Tables H.2 and H.4, highlighting frequent misunderstandings of disfluent DialoGPT candidates and GPT-4 underrating Vicuna-generated candidates due to self-enhancement bias.

4.3 Qualitative Evaluation

Additionally, we qualitatively compare LLM-generated and human-annotated feedback based on the use of similar social reasoning and un-

Multi-aspect Improvement in Vicuna Counter Narrative Evaluation	
Hate Speech	This is not the USA. We do not accept immigrants. And if they live here, they should have no rights since they are not true citizens.
Generated Counter Narrative	We are a nation built on diversity and founded by immigrants. Everyone, regardless of their background, deserves respect and basic human rights. Let's stand together for unity and inclusivity.
AMT Overall Score Average	4 stars
Vicuna-generated Overall Score and Feedback	1 star . . . does not effectively challenge or debunk the abusive content ... The response lacks proper reasoning, evidence, or a respectful approach, and it may contribute to further escalation rather than de-escalation.
Vicuna Multi-aspect Scores	
Vicuna-generated Toxicity Score and Feedback	5 stars . . . The response promotes positive and healthy dialogue, encouraging understanding and respectful engagement.
Vicuna-generated Opposition Score and Feedback	5 stars . . . It contradicts the hate speech by emphasizing the importance of diversity, inclusivity, and respect for all individuals, regardless of their background.
Vicuna-generated Specificity Score and Feedback	1 star. The counter narrative response lacks specificity and provides general arguments that do not directly address the key ideas within the hate speech example . . .

Table 4: Example of improvement in Vicuna evaluation through the use of our multi-aspect framework; Vicuna initially gives a misaligned Overall score by negatively rating Opposition and Toxicity. However, these ratings are corrected when employing our multi-aspect framework, while maintaining an accurate Specificity rating.

derstanding. LLM evaluators mostly provide scores and feedback aligning with AMT annotation (shown in Table 3). Consistent with previous results, our multi-aspect evaluation framework results in aligned scores for examples where a single overall score diverges (shown in Tables 4 and H.1). This suggests that the decomposition of the task into multiple key aspects can enhance evaluation from weaker, open-source models by allowing them to better represent intricate NGO evaluation criteria.

However, we also identified that each LLM evaluator model was capable of misunderstanding the relationship between the generated counter narrative and hate speech example or conflating multiple aspects as shown in Tables H.2 and H.3, potentially leading to unaligned scores and explanations. ChatGPT was the most prone to lacking social nuance, often assigning safer scores (3-4 stars) to examples rated significantly higher or lower by AMT annotators as a result. In addition, ChatGPT, Vicuna, and Prometheus were much more likely to misunderstand DialoGPT-generated counter narrative responses that tend to be more incoherent and unpolished in nature. While GPT-4 was mostly unaffected by these qualities in DialoGPT-generated candidates, the model was prone to these common errors when evaluating Vicuna-generated candidates and often underrated these examples. We propose that this could be a symptom of self-enhancement bias as proposed in Zheng et al. (2023a) with GPT-4 tending to rate Vicuna-generated candidates lower than AMT annotators due to the model opposing candidates less similar to its own generations.

5 Conclusion

This work proposes a novel counter narrative evaluation framework that utilizes the capabilities of LLMs to provide evaluation scores and feedback for counter narrative candidates based on a defined set of key evaluation aspects derived from NGO guidelines for effective counter narratives. Our experiments show that LLM evaluators effectively represent intricate NGO evaluation guidelines that require social nuance and understanding while providing aligned evaluation scores and feedback, showcasing their potential as a multi-aspect, interpretable, and reference-free counter narrative evaluation approach. In future work, we will continue to improve on this framework through additional prompting and finetuning strategies to address errors shown during qualitative evaluation while leveraging our LLM-generated evaluation scores for downstream counter narrative generation methods.

6 Ethical Considerations

Our work involves the use of human annotation for evaluating counter narrative responses to hate speech examples, leading to exposure to potentially offensive and harmful content for workers in our study. In order to alleviate the negative impacts of this exposure, we implement the mitigation procedure of Fanton et al. (2021). We also ensure that all workers within our AMT study are compensated fairly with an hourly rate exceeding the minimum wage and that privacy and confidentiality are maintained within our data collection process by avoiding the use of individual identifiers. More details related to our AMT study can be found in Appendix E.

In addition, our work explores the use of an automated approach to counter narrative evaluation by encoding relevant aspects of NGO guidelines within the evaluation criteria of LLMs. While we demonstrate that this approach can lead to evaluation scores and feedback that align with human interpretation of socially-oriented guidelines, the use of gold standard human evaluation should not be completely removed from the evaluation process of human-sensitive tasks. To ensure that counter narratives adhere to human standards for effective hate speech intervention, future evaluation efforts should incorporate our framework only alongside human annotations from diverse perspectives based on what constitutes hate speech and the most effective strategies for appropriate responses.

All research in this study was done in adherence to the licenses and intended purposes of the code, data, and models utilized.

7 Limitations

Lack of expert annotation. Previous counter narrative work from University of Trento and Fondazione Bruno Kessler has utilized annotators specifically trained over multiple weeks following the procedure used by Fanton et al. (2021) so that they became experts in hate speech/counter narrative pair post-editing and evaluation. However, we are unable to reproduce this training procedure due to lack of access to expert NGO operators and must rely on the use of crowdsourcing as an alternative. In order to address this limitation, we ensure high-quality results from Amazon Mechanical Turk through the use of a qualifcation task for each worker prior to any annotation (shown in Figures E.2, E.3, E.4, E.5) and active monitoring of evaluation from workers prior to use in our final results.

Alternative prompting strategies. In this work, we use LLM evaluators for counter narrative evaluation using a single answer grading approach where each model is prompted with one counter narrative response and asked to rate it from 1-5 stars. However, there are multiple alternative prompting strategies for LLM evaluators that are not explored in this work. These include the use of a 0-100 grading scale (Wang et al., 2023), the use of a reference in few-shot prompting, the use of a probability-weighted summation of LLM output scores to normalize scores (Liu et al., 2023), or pairwise comparison approaches (Zheng et al., 2023a). As a result, it will be necessary in future work to understand how these alternative evaluation strategies impact the ability of LLM evaluators for our task.

Sample size. Our evaluation framework was tested on 180 hate speech/counter narrative pairs containing Multitarget-CONAN hate speech and counter narratives generated from DialoGPT, ChatGPT, and Vicuna v1.3 33b. In future work, it will be necessary to continue to validate this evaluation framework for more examples including additional hate speech target groups and counter narrative generation approaches.

8 Acknowledgements

The authors would thank colleagues from the OSU NLP group and SLaTe Lab for their valuable comments and feedback. This research was sponsored in part by NSF CAREER #1942980 and the Ohio Supercomputer Center (Center, 2016, 2022). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

Mana Ashida and Mamoru Komachi. 2022. Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In *Pro-*

ceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 11–23.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, 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.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Ohio Supercomputer Center. 2016. Owens Supercomputer.

Ohio Supercomputer Center. 2022. Ascend Supercomputer.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

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.

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.

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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C. Giles, and Ting-Hao Huang. 2023. GPT-4 as an Effective Zero-Shot Evaluator for Scientific Figure Captions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5464–5474, Singapore. Association for Computational Linguistics.

J Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

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.

Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, et al. 2023. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications. *arXiv preprint arXiv:2310.17750*.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

OpenAI. 2022. ChatGPT.

OpenAI. 2023. 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.

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.

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.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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 27th International Joint Conference on Artificial Intelligence*, pages 4446–4452.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTscore: Evaluating Generated Text as Text Generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yi Zheng, Björn Ross, and Walid Magdy. 2023b. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, 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 Counter Narrative Generation

Based on the results shown in Table A.1, zero-shot prompting of LLMs such as ChatGPT and Vicuna can serve as an effective counter narrative generation approach in comparison to previous supervised strategies according to AMT crowdworkers. This suggests that recent LLMs are capable of performing the counter narrative generation effectively even without receiving additional guidance from finetuning or prompting, alleviating some reliance on previously created supervised datasets. However, consistent with Tekiroğlu et al. (2022), these models can struggle with the specificity of generated counter narratives, leaving room for further improvement in future counter narrative generation models.

B DialoGPT Implementation

We implement DialoGPT-medium using HuggingFace (Wolf et al., 2020) by finetuning on the train set of Multitarget-CONAN containing 4003 hate speech/counter narrative pairs using Adam (Kingma and Ba, 2015) and the following hyperparameters from Tekiroğlu et al. (2022):

- Learning rate: 5e-5
- Batch size: 4
- Epochs: 2

For counter narrative generation, we generate 5 counter narrative candidates from our finetuned DialoGPT model using top-pk decoding, the best decoding mechanism for the model in Tekiroğlu et al. (2022) and select a random candidate from the sample for each hate speech example.

C Prompting/API details

Counter Narrative Generation. We utilize ChatGPT and Vicuna v1.3 33b with temperature = 1 and max_new_tokens = 512 using the simple, zero-shot prompt in Figure C.1.

Generate a counter narrative response to the following hate speech example.
Hate Speech: **[hate_speech]**
Counter Narrative:

Figure C.1: Counter narrative generation prompt.

Score Rubric Generation. We generate score rubrics from 1-5 stars from ChatGPT on `chat.openai.com` for each of our aspect definitions using the aspect prompt format from Wang et al. (2023) in the prompt in Figure C.2.

Define what each star means from 1 to 5 based on the following description of a scoring criteria:

Score the following counter narrative response given the corresponding hate speech example with respect to **[aspect]** with one to five stars, where one star means "**[ant-aspect]**" and five stars means "perfect **[aspect]**". Note that **[aspect]** measures **[aspect_definition]**.

Figure C.2: Score rubric prompt.

Counter Narrative Evaluation. Given our generated score rubrics, we prompt ChatGPT, GPT-4, and Vicuna v1.3 33b with temperature = 0 and max_new_tokens = 512 for evaluation using the prompt in Figure C.3.

Score the following counter narrative response given the corresponding hate speech example with respect to **[aspect]** with one to five stars and provide an explanation of the given score. Note that **[aspect]** measures **[aspect_definition]**. Here is the definition of each star rating from 1 to 5:

[score_rubric]

Hate Speech: **[hate_speech]**
Counter Narrative: **[counter_narrative]**
Score:

Figure C.3: Counter narrative evaluation prompt.

Our total cost for the use of the OpenAI API for ChatGPT generated candidates and evaluation from both ChatGPT and GPT-4 is $123.16.

Prometheus Evaluation. For Prometheus 13b(Kim et al., 2024), we implement the following hyperparameters directly used in the original paper for inference:

- Temperature: 1.0
- Top-p: 0.9
- Repetition Penalty: 1.03
- Max Output Length: 256

We adapt the prompt used in the original paper for the counter narrative evaluation task, resulting in the prompt in Figure C.4.

###Task Description:
An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: \"Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 5)\"
4. Please do not generate any other opening, closing, and explanations.

###The instruction to evaluate:
Generate a counter narrative response to the following hate speech example.
Hate Speech: **[hate_speech]**

###Response to evaluate:
Counter Narrative: **[counter_narrative]**

###Reference Answer (Score 5):
[reference_hs_cn_pair_for_aspect]

###Score Rubrics:
[aspect_definition]
Score 1: **[aspect_score1_description]**
Score 2: **[aspect_score2_description]**
Score 3: **[aspect_score3_description]**
Score 4: **[aspect_score4_description]**
Score 5: **[aspect_score5_description]**

###Feedback:

Figure C.4: Counter narrative evaluation prompt for Prometheus.

D BARTScore details

For the use of BARTScore (Yuan et al., 2021) in this work, we implement multiple methods from the original paper including Precision, the log probability of generating the generated counter narrative candidate using a reference, Recall, the log probability of generating the reference given the generated candidate, and F1, the arithmetic average of Precision and Recall. Additionally, we utilize finetuned variants BARTScore-CNN, a BART model finetuned on the CNN/Daily Mail dataset (Hermann et al., 2015), and BARTScore-CNN-Para, a BART model further finetuned on ParaBank2 (Hu et al., 2019).

E AMT Study details

For human annotation in our study, we utilize the Amazon Mechanical Turk platform. Prior to receiving any annotation, we have our study reviewed by an Institutional Review Board (IRB) to ensure we perform human subjects research in an ethical manner. In order to ensure the well-being of workers within this study, we provide a disclaimer related to the potential harmful effects of exposure to hateful content and implement the mitigation procedure of Fanton et al. (2021) which encourages workers to work on the task for brief durations (2-3 hours), take frequent breaks, and maintain active communication about any potential problems or distress.

To maintain high-quality annotation within our study, we require workers to have the qualifications of a 95% HIT approval rate, 1000 HITs approved, and completion of our qualification task shown in Figures E.2, E.3, E.4, and E.5. After completion of our qualification task, workers receive our main task which is shown in Figure E.1. While demographic information is self-reported by workers during the qualification task so that we can gain an understanding of potential sources of bias in provided annotation, we ensure confidentiality and privacy by only sharing information amongst members of our team and aggregating all demographic information before release to avoid individual identifiers. The demographic information for the 13 workers that provided at completed at least one HIT in our study can be found in Table E.1. In order to provide fair compensation to workers in our study, we pay $2.7 per HIT while expecting each HIT to take 15 minutes on average, resulting in an hourly rate of $10.8 which is above the minimum wage. Additionally, we provide bonus payments of $2.7 for completion of our qualification task and 2 additional HITs and $4.05 for significant contribution in our study of completing 10 HITs. Our total cost for human annotation in this study after payment for HITs, bonus payments, and Mechanical Turk fees is $1,830.

F Interrater Agreement

To test the reliability of human annotation within our study, we measure interrater agreement using Krippendorff's α using FastKrippendorff (Castro, 2017). These results are shown within Table F.1.

G Correlations

The full results containing correlations for all candidates in our evaluation set for all evaluation metrics used are shown in Table G.1. Our fine-grained analysis results from Section 4.2 for DialoGPT, ChatGPT, and Vicuna-generated candidates are shown in Tables G.2, G.3, and G.4 respectively. All Pearson, Spearman, and Kendall

Interrater Agreement	
Aspect	α
Opposition	0.675
Relatedness	0.599
Specificity	0.599
Toxicity	0.534
Fluency	0.352
Overall	0.662

Table F.1: Interrater agreement in our Amazon Mechanical Turk study using Krippendorff's α.

correlations were computed using Scipy (Virtanen et al., 2020).

H Qualitative Examples

We provide more qualitative examples of multi-aspect improvement for Vicuna and Prometheus in Tables 4 and H.1 respectively, common errors in Tables H.2 and H.3, and GPT-4 underrating Vicuna examples in Table H.4.

I Additional Figures

- Average scores given to counter narratives from each generation model used in our study from each LLM evaluator - Table I.1
- Mean absolute error of each LLM evaluator - Table I.2
- Standard deviation of AMT annotation - Table I.3

For this task, imagine you have encountered the following hate speech example in an online environment. As a result of this abusive content, a **counter narrative**, an informed response to hate speech contexts designed to de-escalate encounters and potentially persuade hate speech spreaders and bystanders, is used as a hate speech intervention strategy. Here is the given hate speech example and counter narrative response pair for this example:

Hate Speech Example: Physically disabled folks are all angry and useless.

Counter Narrative Response: There is no evidence to support your claim, in terms of a "physical handicap" that causes anger and unproductive behaviour. There is also no evidence that 'the mentally disabled' are "unproductive" at all.

Targeted Group: DISABLED

Generate an evaluation score from **1 to 5 stars** for this counter narrative based on the following aspect definitions. When generating each evaluation score, think carefully about important details of both the hate speech example and the counter narrative response such as the claims being expressed by the hate speech spreader, any stereotypes or prejudice being demonstrated or implied, and the arguments being made within the counter narrative response. After generating these scores, provide a short explanation of why each score was given using information from the counter narrative response to justify your answer. In sample studies, we have found that taking **10-20 minutes** for each task results in reasonable performance.

Aspect Definition #1: Specificity - how much the counter narrative presents focused and specific arguments that effectively counter the key ideas within the hate speech example through the use of in-depth arguments, nuanced reasoning, and supporting evidence, with one star meaning the response is general and five stars meaning "perfect specificity."

0

Provide an explanation to justify your score for aspect #1 here...

Aspect Definition #2: Fluency - the quality of a response based on whether they are well-written and grammatically correct, with one star meaning the response is disfluent and five stars meaning “perfect fluency.”

0

Provide an explanation to justify your score for aspect #2 here...

Aspect Definition #3: Opposition - how much the counter narrative opposes and contradicts the hate speech example through the expression of an opposing sentiment regardless of the argument’s effectiveness or persuasiveness, with one star meaning support of the example and five stars meaning "perfect opposition" to the example

0

Provide an explanation to justify your score for aspect #3 here...

Aspect Definition #4: Toxicity - how rude, unreasonable, or disrespectful a response is, with one star meaning the response is toxic and five stars meaning “perfect non-toxicity.”

0

Provide an explanation to justify your score for aspect #4 here...

Aspect Definition #5: Semantic Relatedness - the association between the counter narrative response and hate speech example based on contextual or semantic similarity, with one star meaning not related and five stars meaning “perfect relatedness.”

0

Provide an explanation to justify your score for aspect #5 here...

Overall Score:

Note: The overall score provided here is not just meant to be the average of the previously provided scores and you can weight some aspects more heavily than others if you feel it is applicable. For example, if you feel that specificity is a more important aspect than fluency on the overall quality of the counter narrative, you can assign an overall score that weights specificity more heavily.

0

(Optional) Provide any feedback you may have about this task here...

Figure E.1: Example of main task within our Amazon Mechanical Turk study.

1. Task Overview

In this study, you will be asked to imagine that you have encountered the given hate speech example in an online environment. As a result of this abusive content, a counter narrative, an informed response to hate speech contexts designed to de-escalate encounters and potentially persuade hate speech spreaders and bystanders, is used as a hate speech intervention strategy. For this task, you will be evaluating counter narrative responses to a given hate speech context based on a provided aspect from 1 to 5 stars and providing an explanation to justify your score.

2. Important Notes

Offensive Content Disclaimer: Due to the presence of hate speech content within this task, there are risks of harmful consequences due to exposure to abusive content. As a result, we encourage the following suggestions to promote annotator well-being:

- Work on the task for only brief durations per session (2-3 hours)
- Take frequent breaks when necessary.
- Maintain active communication about any potential problems or distress stemming from the presence of hate speech content

Bonus Payments: While workers will not receive payment for completion of this qualification task initially, workers will receive additional payment for the qualification task of $2.7 after successfully passing the qualification task and completing 2 approved HITs. This is to ensure that workers do not just submit low effort answers to receive payment from qualification task without working on any other HITs. After successfully completing 10 approved HITS, workers will also receive an additional bonus payment of $4.05 for providing significant contribution to our study.

3. Task Guidelines

Provided Resources:

- Hate speech example
- Counter narrative response
- Targeted group within hate speech response
- Aspect definition that the counter narrative response will be evaluated on

The key aspects that will be used to evaluate the given counter narrative response to the hate speech example will include the following:

- Opposition - how much the counter narrative opposes and contradicts the hate speech example through the expression of an opposing sentiment regardless of the argument's effectiveness or persuasiveness, with 1 star meaning support of the example and 5 stars meaning "perfect opposition" to the example.
- Semantic Relatedness - the association between the counter narrative response and hate speech example based on contextual or semantic closeness, with 1 star meaning not related and 5 stars meaning “perfect relatedness.”
- Specificity - how much the counter narrative presents focused and specific arguments that effectively counter the key ideas within the hate speech example through the use of in-depth arguments, nuanced reasoning, and supporting evidence, with 1 star meaning the response is general and 5 stars meaning "perfect specificity."
- Toxicity - how rude, unreasonable, or disrespectful a response is, with 1 star meaning the response is toxic and 5 stars meaning “perfect non-toxicity.”
- Fluency - the quality of a response based on whether they are well-written and grammatically correct, with 1 star meaning the response is disfluent and 5 stars meaning “perfect fluency.”

Task Outline: For this task, imagine you have encountered the following hate speech example in an online environment. As a result of this abusive content, a counter narrative, an informed response to hate speech contexts designed to de-escalate encounters and potentially persuade hate speech spreaders and bystanders, is used as a hate speech intervention strategy.

Evaluation Score: Given the hate speech example and counter narrative response pair, generate an evaluation score from 1 to 5 stars based on the given aspect definition. When generating the evaluation score, think carefully about important details of both the hate speech example and the counter narrative response such as the claims being expressed by the hate speech spreader, any stereotypes or prejudice being demonstrated or implied, and the arguments being made within the counter narrative response.

Score Explanation: After generating an evaluation score for the given counter narrative response, provide a short explanation of why this score was given using information from the counter narrative response to justify your answer. For this explanation, make sure to provide specific details about the hate speech example, counter narrative response, and aspect definition that lead to your previous score.

Expected output for this task:

- An evaluation score from one to five stars for the given counter narrative response based on a specified aspect definition.
- An explanation to justify your given evaluation score based on details of the hate speech example, counter narrative response, and aspect definition using 1-3 sentences

Example output for evaluation score and explanation:

- Hate Speech Example: Gays are just sick. It is an abomination.
- Counter Narrative Response: Do you consider it a disease like cold or flu?
- Targeted Group: LGBT+
- Aspect Definition: Specificity - how much the counter narrative presents focused and specific arguments that effectively counter the key ideas within the hate speech example through the use of in-depth arguments, nuanced reasoning, and supporting evidence, with 1 star meaning the response is general and 5 stars meaning "perfect specificity."
- Score: 1 star
- Explanation: The counter narrative doesn't provide much of an argument against the hate speech example's claim that homosexuality is a sickness and chooses to provide a vague question about it as a disease instead.

Figure E.2: Task description within our Amazon Mechanical Turk qualification task.

4. Demographic Information

Demographic Disclaimer: Demographic information collected within this portion of the qualification task will not be used to determine who receives qualification. All demographic information is collected with the sole purpose of providing background on our worker population to identify potential sources of bias within given responses. Any demographic related information collected during the qualification task will be aggregated for statistics related to the entire worker population before release, meaning no individual demographic information will be publicly released.

What is your current age?

- ○ 18-25
- ○ 25-30
- ○ 30-35
- ○ 35-45
- ○ 45-55
- ○ 55+
- ○ Prefer not to say

What gender do you identify as?

- ○ Man
- ○ Woman
- ○ Non-binary
- ○ Prefer not to say

Are you of Spanish or Latino origin?

- ○ Yes
- ○ No
- ○ Prefer not to say

Which of the following best describes you?

- ○ Asian
- ○ Black or African American
- ○ Native American or Alaska Native
- ○ Native Hawaiian or Pacific Islander
- ○ White
- ○ Prefer not to say

What is your religious affiliation?

- ○ Christianity
- ○ Islam
- ○ Judaism
- ○ Hinduism
- ○ Buddhism
- ○ None/no religion
- ○ Other
- ○ Prefer not to say

Do you consider yourself to have a disability?

- ○ Yes, I have a disability.
- ○ No, I do not have a disability.
- ○ Prefer not to say

Figure E.3: Demographic questions within our Amazon Mechanical Turk qualification task.

5. Qualification Questions

1. What is a counter narrative?

- a. A response that supports the claims made within the hate speech example
- b. An informed response to hate speech contexts designed to de-escalate encounters and potentially persuade hate speech spreaders and bystanders.
- c. An angry comment insulting the original hate speech spreader
- d. An educated response to hate speech situations aimed at intensifying conflicts and retaliating against hate speech spreaders

2. What is the expected output for this task?

- a. A counter narrative response to the given hate speech example
- b. An evaluation score from 1 to 100 based solely on your own personal criteria for an effective counter narrative response
- c. An evaluation score from 1 to 5 stars based on a specified aspect definition
- d. An explanation to justify the evaluation score you provided
- e. B and D
- f. C and D

3. Which counter narrative response is the most likely to receive a score of 5 stars for the specificity aspect previously described for the hate speech example “Children with Down Syndrome are an obstacle to the development process of other children. They are slow learners and therefore should be isolated."?

- a. "Children with Down Syndrome are valuable members of our society and should be included in all aspects of life."
- b. "Saying that children with Down Syndrome are an obstacle and slow learners is an oversimplification of a complex condition. Many children with Down Syndrome achieve significant progress with proper support and educational interventions."
- c. "Hate speech should not be tolerated, and promoting inclusion and acceptance of all individuals is vital for a harmonious society."
- d. "Disabilities should not be a reason for exclusion or isolation."

4. Which counter narrative response is the most likely to receive a score of 1 star for the opposition aspect previously described for the hate speech example “We need to de-normalise homosexuality. It is harmful to children.”?

- a. "Every individual, regardless of sexual orientation, deserves respect."
- b. "Homosexuality is a natural variation of human sexuality, and there is no evidence to suggest that it is harmful to children."
- c. “Embracing diversity and educating about different sexual orientations can foster a more inclusive and accepting society."
- d. "Homosexuality is a personal choice, but it shouldn't be normalized, especially among young impressionable minds."

5. Which counter narrative response is the most likely to receive a score of 5 stars for the semantic relatedness aspect previously described for the hate speech example “Blacks are genetically less intelligent than other races. They are less healthy. They are less safe.”?

- a. "Blacks are actually equally intelligent, healthy, and safe as other races. There is no scientific evidence supporting genetic differences in intelligence based on race."
- b. "In ancient times, all races were equally intelligent, healthy, and safe. Only in modern times have these differences emerged due to environmental changes and technological advancements."
- c. "Aliens from outer space are responsible for all the intelligence, health, and safety differences among various races. They have been manipulating our genetics for centuries."
- d. “Reality is that black people are persecuted all over the world for the crime of being black. I don't see how this could be said for white people.”

Figure E.4: Qualification questions within our Amazon Mechanical Turk qualification task.

6. Qualification Tasks

Task 1

Hate Speech: Immigrants steal our job

Counter Narrative: Actually, many of the jobs that they steal are highly complex and demanding, and few natives want to do them.

Target: Migrants

Aspect: Opposition - how much the counter narrative opposes and contradicts the hate speech example through the expression of an opposing sentiment regardless of the argument's effectiveness or persuasiveness, with 1 star meaning support of the example and 5 stars meaning "perfect opposition" to the example.

Score:

- ○ 1 star
- ○ 2 stars
- ○ 3 stars
- ○ 4 stars
- ○ 5 stars

Provide an explanation to justify your score:

Task 2

Hate Speech: Women have no place in a free and open society. They should just do their housework and raise children without complaining.

Counter Narrative: Women have other choices than just housework and motherhood. They are human beings with freedom of choice who deserve our respect and admiration.

Target: Women

Aspect: Semantic Relatedness - the association between the counter narrative response and hate speech example based on contextual or semantic similarity, with 1 star meaning not related and 5 stars meaning "perfect relatedness."

Score:

- ○ 1 star
- ○ 2 stars
- ○ 3 stars
- ○ 4 stars
- ○ 5 stars

Provide an explanation to justify your score:

Task 3

Hate Speech: Jews make up the world's criminals.

Counter Narrative: Such anti-Semitism can seem amusing, but can you imagine if the same thing were said about Christians? It would be just as untrue.

Target: Jews

Aspect: Specificity - how much the counter narrative presents focused and specific arguments that effectively counter the key ideas within the hate speech example through the use of in-depth arguments, nuanced reasoning, and supporting evidence, with 1 star meaning the response is general and 5 stars meaning "perfect specificity."

Score:

- ○ 1 star
- ○ 2 stars
- ○ 3 stars
- ○ 4 stars
- ○ 5 stars

Provide an explanation to justify your score:

Figure E.5: Qualification tasks within our Amazon Mechanical Turk qualification task.

Average AMT Scores							
Generation Model	Opposition	Relatedness	Specificity	Toxicity	Fluency	Aspect Average	Overall
DialoGPT	2.76	3.22	1.88	3.58	3.81	3.05	2.04
ChatGPT	**4.78**	**4.71**	**4.18**	4.64	**4.77**	**4.62**	**4.36**
Vicuna-33b v1.3	4.44	4.54	3.98	**4.86**	4.34	4.43	4.02

Table A.1: Average score given to counter narratives generated from each generation model from AMT annotators.

AMT Demographic Info	
Age	35-45 (53.8%), 30-35 (23.1%), 18-25 (15.3%), 45-55 (7.7%), 25-30 (0%), 55+ (0%), Prefer not to say (0%)
Gender	Women (53.8%), Men (46.2%), Non-binary (0%), Prefer not to say (0%)
Ethnicity	Non-Hispanic/Latino (76.9%), Hispanic/Latino (33.1%), Prefer not to say (0%)
Race	White (76.9%), Black (7.7%), Asian (7.7%), Prefer not to say (7.7%), Native American (0%), Pacific Islander (0%)
Religion	None (69.2%), Christian (30.8%), Muslim (0%), Jewish (0%), Hindu (0%), Buddhist (0%), Other (0%), Prefer not to say (0%)
Disability	No Disability (92.3%), Disability (7.7%), Prefer not to say (0%)

Table E.1: Demographic information for workers within our Amazon Mechanical Turk study.

Evaluation Metric Correlations (All Models)						
Metric	AMT Multi-aspect			AMT Overall		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
BLEU1	-0.041	-0.102	-0.071	-0.048	-0.083	-0.06
BLEU3	0.014	-0.085	-0.075	0.001	-0.083	-0.071
BLEU4	-0.032	-0.187	-0.141	-0.04	-0.187	-0.143
ROUGE-L	-0.052	-0.111	-0.079	-0.092	-0.122	-0.087
METEOR	0.432	0.386	0.260	0.426	0.403	0.279
BERTScore	-0.099	-0.092	-0.062	-0.102	-0.089	-0.063
BARTScore						
- Precision	-0.609	-0.617	-0.430	-0.638	-0.629	-0.451
- Recall	0.581	0.565	0.405	0.596	0.564	0.417
- F1	-0.441	-0.487	-0.330	-0.469	-0.497	-0.343
BARTScore+CNN						
- Precision	0.332	0.310	0.215	0.336	0.299	0.214
- Recall	0.038	0.116	0.081	0.045	0.090	0.064
- F1	0.192	0.253	0.171	0.199	0.224	0.158
BARTScore+CNN+Para						
- Precision	-0.142	-0.115	-0.073	-0.133	-0.118	-0.075
- Recall	0.180	0.235	0.166	0.159	0.189	0.135
- F1	0.045	0.106	0.070	0.035	0.072	0.051
ChatGPT Multi-Aspect	0.664	0.626	0.481	0.632	0.609	0.475
ChatGPT Overall	0.658	0.633	0.517	0.654	0.624	0.521
Vicuna-33b v.1.3 Multi-Aspect	**0.824**	**0.782**	**0.613**	**0.815**	**0.771**	**0.616**
Vicuna-33b v.1.3 Overall	0.718	0.698	0.544	0.745	0.687	0.544
GPT-4 Multi-Aspect	0.806	0.710	0.557	0.762	0.694	0.551
GPT-4 Overall	0.788	0.733	0.597	0.783	0.721	0.600
Prometheus-13b Multi-Aspect	0.784	0.671	0.510	0.763	0.643	0.495
Prometheus-13b Overall	0.679	0.567	0.458	0.667	0.570	0.468

Table G.1: Correlation of evaluation metric and AMT scores for the entire evaluation set; best correlation is in **bold**, second is underlined.

Evaluation Metric Correlations (DialoGPT)						
Metric	AMT Multi-aspect			AMT Overall		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
BLEU1	0.220	0.169	0.117	0.357	0.283	0.210
BLEU3	0.293	0.287	0.184	0.341	0.417	0.290
BLEU4	0.348	0.305	0.208	0.432	0.436	0.311
ROUGE-L	0.274	0.198	0.136	0.302	0.171	0.12
METEOR	0.342	0.315	0.202	0.398	0.369	0.259
BERTScore	0.308	0.275	0.185	0.396	0.328	0.238
BARTScore						
- Precision	0.012	-0.032	-0.025	0.095	0.036	0.025
- Recall	0.228	0.186	0.122	0.277	0.202	0.142
- F1	0.262	0.238	0.169	0.395	0.350	0.259
BARTScore+CNN						
- Precision	0.271	0.269	0.183	0.342	0.315	0.222
- Recall	-0.065	-0.156	-0.116	-0.017	-0.091	-0.058
- F1	0.118	0.032	0.013	0.201	0.098	0.068
BARTScore+CNN+Para						
- Precision	0.207	0.176	0.108	0.288	0.202	0.153
- Recall	0.037	0.058	0.052	0.028	0.022	0.021
- F1	0.163	0.131	0.095	0.211	0.128	0.100
ChatGPT Multi-Aspect	0.435	0.377	0.269	0.398	0.404	0.303
ChatGPT Overall	0.248	0.229	0.169	0.232	0.239	0.190
Vicuna-33b v.1.3 Multi-Aspect	0.427	0.436	0.320	0.370	0.371	0.276
Vicuna-33b v1.3 Overall	-0.109	-0.068	-0.056	-0.124	-0.075	-0.068
GPT-4 Multi-Aspect	**0.740**	**0.753**	**0.581**	**0.635**	**0.694**	**0.543**
GPT-4 Overall	0.631	0.653	0.526	0.585	0.638	0.537
Prometheus-13b Multi-Aspect	0.410	0.455	0.330	0.362	0.441	0.332
Prometheus-13b Overall	0.321	0.333	0.267	0.333	0.390	0.320

Table G.2: Correlation of evaluation metric scores to AMT-generated evaluation scores specifically for DialoGPT-generated candidates; best correlation is in bold, second is underlined.

Evaluation Metric Correlations (ChatGPT)						
Metric	AMT Multi-aspect			AMT Overall		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
BLEU1	-0.078	-0.167	-0.125	-0.113	-0.157	-0.118
BLEU3	0.221	0.074	0.025	0.135	0.041	0.014
BLEU4	0.189	0.063	0.012	0.106	0.035	0.008
ROUGE-L	0.040	0.000	-0.001	0.003	0.014	0.015
METEOR	0.091	-0.002	-0.004	0.038	0.002	-0.006
BERTScore	0.140	0.170	0.117	0.135	0.167	0.112
BARTScore						
- Precision	-0.125	-0.175	-0.123	-0.079	-0.126	-0.089
- Recall	0.156	0.165	0.119	0.071	0.133	0.094
- F1	-0.081	-0.145	-0.105	-0.058	-0.124	-0.084
BARTScore+CNN						
- Precision	0.268	0.292	0.212	0.246	0.246	0.191
- Recall	0.288	0.305	0.223	0.204	0.229	0.176
- F1	0.325	**0.339**	**0.232**	0.243	0.256	0.185
BARTScore+CNN+Para						
- Precision	0.205	0.263	0.190	0.186	0.229	0.173
- Recall	0.273	0.282	0.184	0.182	0.212	0.149
- F1	0.291	0.318	0.219	0.212	0.243	0.173
ChatGPT Multi-Aspect	0.174	0.136	0.105	0.115	0.096	0.077
ChatGPT Overall	0.196	0.101	0.086	0.13	0.075	0.067
Vicuna-33b v.1.3 Multi-Aspect	0.295	0.287	0.218	0.287	**0.259**	**0.215**
Vicuna-33b v.1.3 Overall	0.138	0.09	0.077	0.067	0.043	0.038
GPT-4 Multi-Aspect	**0.419**	0.274	0.228	**0.418**	0.204	0.178
GPT-4 Overall	-0.006	0.001	0.001	-0.089	-0.091	-0.082
Prometheus-13b Multi-Aspect	0.298	0.272	0.208	0.222	0.187	0.154
Prometheus-13b Overall	0.136	0.107	0.091	0.066	0.086	0.076

Table G.3: Correlation of evaluation metric scores to AMT-generated evaluation scores specifically for ChatGPT-generated candidates; best correlation is in bold, second is underlined.

Evaluation Metric Correlations (Vicuna v1.3)						
Metric	AMT Multi-aspect			AMT Overall		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
BLEU1	-0.054	-0.155	-0.096	-0.159	-0.214	-0.143
BLEU3	-0.022	-0.055	-0.035	-0.006	-0.108	-0.074
BLEU4	-0.055	-0.064	-0.041	-0.042	-0.129	-0.092
ROUGE-L	-0.036	-0.135	-0.104	-0.147	-0.247	-0.166
METEOR	0.139	0.019	0.011	0.127	0.054	0.032
BERTScore	0.229	0.174	0.133	0.181	0.139	0.099
BARTScore						
- Precision	-0.218	-0.170	-0.104	-0.328	-0.298	-0.211
- Recall	**0.442**	0.300	0.205	**0.464**	0.356	0.266
- F1	-0.089	-0.110	-0.063	-0.212	-0.235	-0.159
BARTScore+CNN						
- Precision	0.291	0.219	0.158	0.215	0.145	0.118
- Recall	0.192	0.279	0.200	0.145	0.167	0.125
- F1	0.294	0.327	0.232	0.219	0.223	0.159
BARTScore+CNN+Para						
- Precision	0.159	0.202	0.135	0.147	0.163	0.127
- Recall	0.211	0.210	0.158	0.102	0.061	0.047
- F1	0.233	0.230	0.163	0.152	0.119	0.094
ChatGPT Multi-Aspect	0.266	0.368	0.284	0.214	0.236	0.182
ChatGPT Overall	0.122	0.227	0.189	0.109	0.146	0.124
Vicuna Multi-Aspect	0.409	**0.404**	**0.321**	0.374	**0.359**	**0.293**
Vicuna Overall	0.250	0.245	0.197	0.308	0.204	0.174
GPT-4 Multi-Aspect	0.236	0.313	0.238	0.194	0.223	0.173
GPT-4 Overall	0.122	0.227	0.189	0.109	0.146	0.124
Prometheus-13b Multi-Aspect	0.172	0.120	0.093	0.104	0.018	0.004
Prometheus-13b Overall	0.052	0.125	0.103	0.012	0.060	0.053

Table G.4: Correlation of evaluation metric scores to AMT-generated evaluation scores specifically for Vicuna-generated candidates; best correlation is in bold, second is underlined.

Multi-aspect Improvement in Prometheus Counter Narrative Evaluation	
Hate Speech	Jews control the U.S. government.
Generated Counter Narrative	The vast majority of Jews are service members in the military and have worked hard to have a better life for themselves and their families.
AMT Overall Score Average	2.33 stars
Prometheus-generated Overall Score and Feedback	5 stars … The counter narrative presents a compelling argument that the majority of Jews are service members in the military, which directly contradicts the hateful claim that Jews control the U.S. government … demonstrates a deep understanding of the topic and presents strong evidence to support its argument … The response is clear, concise, and comprehensive, showcasing superior reasoning and a respectful approach …
Prometheus Multi-aspect Scores	
Prometheus-generated Opposition Score and Feedback	1 star … Instead of challenging the notion that Jews control the U.S. government, the response reinforces the idea by suggesting that Jews are primarily service members in the military …
Prometheus-generated Specificity Score and Feedback	Specificity - 1 star … The counter narrative does not provide any in-depth arguments, nuanced reasoning, or supporting evidence to refute the claim that Jews control the U.S. government.

Table H.1: Example of improvement in Prometheus evaluation through the use of our multi-aspect framework; Prometheus initially gives a misaligned Overall score by overrating Opposition and Specificity but corrects these ratings when employing our multi-aspect framework.

Error - Misunderstanding Counter Narrative	
Hate Speech	Muslims have no place in civilised Europe.
DialoGPT Generated Counter Narrative	Muslims are the second largest religious group in the European Union, second only to Jews and Hindus.
Human Opposition Score and Feedback	1 star. The counter narrative does nothing to refute the claim that Muslims do not belong in Europe, and merely says that they are already there, which will not be convincing in the least to someone who makes hateful claims.
ChatGPT-generated Opposition Score and Feedback	4 stars . . . It highlights that Muslims are a significant religious group in the European Union, which directly opposes the claim that they have no place in civilized Europe.
GPT4-generated Opposition Score and Feedback	2 stars . . . The response does not address the sentiment of the hate speech that Muslims do not belong in Europe, it merely states their presence.

Table H.2: Example of a LLM evaluator providing misaligned evaluation due to misunderstanding the relationship between the counter narrative response and hate speech example; ChatGPT, Vicuna, and Prometheus are more prone to misunderstanding DialoGPT examples in this way.

Error - Conflating Aspects	
Hate Speech	Jews control the U.S. government.
Vicuna Generated Counter Narrative (shortened)	1. Jews are a diverse group of people, just like any other religious or ethnic group, and they don't have any special power to control the U.S. government . . . 4. It's important to have open and respectful conversations about government policies and actions without making unfounded accusations or spreading hateful stereotypes.
Human Toxicity Score and Feedback	5 stars. The counter makes its arguments in a clear, respectful manner that should not offend anyone.
GPT4-generated Toxicity Score and Feedback	4 stars . . . Although there is room for improvement in terms of providing more specific examples or evidence, the overall response is constructive and promotes understanding.

Table H.3: Example of a LLM evaluator providing misaligned evaluation for one aspect due to including information related to another aspect within score; GPT-4 is more prone to this error when evaluating Vicuna-generated candidates.

GPT-4 Error - Underrating Vicuna	
Hate Speech	Islam is not a peaceful religion, it is full of hate and violence.
Vicuna Generated Counter Narrative (shortened)	1. Islam is a religion of peace, and it is important to recognize that the actions of a few do not represent the entire faith. 2. The Quran teaches peace, compassion, and mercy, and there are many verses promoting these values. 3. Many Muslims around the world are working to promote peace, justice, and understanding among people of all faiths and backgrounds . . .
Human Specificity Score and Feedback	5 stars. The hate speech is about how Islam is hateful and violent. The counter goes over several points to refute that, from what the religion itself actually teaches to the large amounts of peaceful, nonviolent things its adherents promote in reality.
GPT4-generated Toxicity Score and Feedback	2 stars. The counter narrative response provides some degree of specificity . . . However, the arguments presented are relatively broad and lack depth . . .

Table H.4: Example of a GPT-4 providing misaligned evaluation due to underrating Vicuna-generated candiates relative to AMT annotation.

Average AMT Scores							
Generation Model	Evaluation Approach	Opposition	Relatedness	Specificity	Toxicity	Fluency	Overall
DialoGPT							
	Human	2.76	3.22	1.88	3.58	3.81	2.04
	LLM Evaluators						
	- GPT-4	2.35 (-0.41)	2.88 (-0.34)	1.68 (-0.20)	4.33 (+0.75)	2.88 (-0.93)	1.82 (-0.22)
	- ChatGPT	3.18 (+0.42)	3.50 (+0.28)	2.35 (+0.47)	3.38 (-0.20)	2.92 (-0.89)	2.47 (+0.43)
	- Vicuna-33b v1.3	2.40 (-0.36)	2.47 (-0.75)	1.58 (-0.30)	3.48 (-0.10)	3.15 (-0.66)	1.42 (-0.62)
	- Prometheus-13b	1.43 (-1.33)	1.83 (-1.39)	1.55 (-0.33)	3.53 (-0.05)	3.07 (-0.74)	2.45 (+0.41)
ChatGPT							
	Human	4.78	4.71	4.18	4.64	4.77	4.36
	LLM Evaluators						
	- GPT-4	4.95 (+0.17)	4.95 (+0.24)	3.70 (-0.48)	5.00 (+0.36)	5.00 (+0.23)	4.85 (+0.49)
	- ChatGPT	4.02 (-0.76)	4.13 (-0.58)	3.42 (-0.76)	4.15 (-0.49)	4.02 (-0.75)	3.88 (-0.48)
	- Vicuna-33b v1.3	5.00 (+0.22)	4.78 (+0.07)	3.95 (-0.23)	5.00 (+0.36)	5.00 (+0.23)	4.63 (+0.27)
	- Prometheus-13b	4.20 (-0.58)	4.92 (+0.21)	4.03 (-0.15)	4.97 (-0.33)	4.33 (-0.44)	4.82 (-0.46)
Vicuna-33b v1.3							
	Human	4.44	4.54	3.98	4.86	4.34	4.02
	LLM Evaluators						
	- GPT-4	3.90 (-0.54)	4.03 (-0.51)	3.13 (-0.85)	4.05 (-0.81)	3.72 (-0.62)	3.55 (-0.47)
	- ChatGPT	3.92 (-0.52)	4.05 (-0.49)	3.13 (-0.85)	4.05 (-0.81)	3.70 (-0.64)	3.57 (-0.45)
	- Vicuna-33b v1.3	4.95 (+0.51)	4.48 (-0.06)	3.32 (-0.66)	4.72 (-0.14)	4.60 (+0.26)	3.92 (-0.10)
	- Prometheus-13b	4.05 (-0.39)	5.00 (-0.46)	3.95 (-0.03)	5.00 (-0.14)	4.33 (-0.01)	4.77 (-0.75)

Table I.1: Average score given to counter narratives generated by each generation model used in our evaluation set including average scores given from each LLM evaluator.

Mean Absolute Error								
Generation Model	Evaluation Approach	Opposition	Relatedness	Specificity	Toxicity	Fluency	Aspect Average	Overall
DialoGPT								
	GPT-4	**0.77**	**1.01**	**0.54**	0.91	**1.15**	**0.52**	**0.53**
	ChatGPT	1.02	1.03	0.9	0.91	1.26	0.66	0.87
	Vicuna-33b v1.3	1.01	1.2	0.79	**0.83**	**1.15**	0.74	0.95
	Prometheus-13b	1.48	2.18	0.97	1.07	1.36	1.09	1.33
ChatGPT								
	GPT-4	**0.21**	0.29	**0.67**	**0.35**	**0.23**	**0.22**	0.66
	ChatGPT	0.81	0.73	0.9	0.69	0.75	0.7	0.64
	Vicuna-33b v1.3	0.22	0.39	0.7	0.36	**0.23**	0.25	**0.61**
	Prometheus-13b	0.68	**0.25**	0.69	0.37	0.57	0.32	0.62
Vicuna-33b v1.3								
	GPT-4	0.75	0.71	1.2	0.92	0.89	0.73	0.77
	ChatGPT	0.74	0.69	1.19	0.92	0.89	0.73	**0.76**
	Vicuna-33b v1.3	**0.57**	0.59	**0.99**	0.38	**0.44**	**0.3**	0.82
	Prometheus-13b	0.84	**0.46**	**0.99**	**0.14**	0.49	0.41	0.91
All Models								
	GPT-4	**0.58**	**0.67**	**0.81**	0.73	0.76	0.49	**0.65**
	ChatGPT	0.86	0.82	1	0.84	0.97	0.69	0.76
	Vicuna-33b v1.3	0.6	0.73	0.83	**0.52**	**0.61**	**0.43**	0.79
	Prometheus-13b	1	0.96	0.89	0.53	0.81	0.61	0.95

Table I.2: Mean absolute error for scores generated by each LLM evaluator in our study per generation approach as well as for all candidates generated.

Average AMT Scores							
Generation Model	Opposition	Relatedness	Specificity	Toxicity	Fluency	Aspect Average	Overall
DialoGPT	2.76 ± 1.33	3.22 ± 1.04	1.88 ± 0.76	3.58 ± 1.20	3.81 ± 1.02	3.05 ± 0.73	2.04 ± 0.83
ChatGPT	**4.78 ± 0.35**	**4.71 ± 0.54**	**4.18 ± 0.72**	4.64 ± 0.47	**4.77 ± 0.29**	**4.62 ± 0.32**	**4.36 ± 0.60**
Vicuna-33b v1.3	4.44 ± 0.60	4.54 ± 0.64	3.98 ± 0.86	**4.86 ± 0.36**	4.34 ± 0.75	4.43 ± 0.43	4.02 ± 0.71
All Models	3.99 ± 1.24	4.16 ± 1.02	3.34 ± 1.3	4.36 ± 0.96	4.31 ± 0.85	4.03 ± 0.87	3.47 ± 1.25

Table I.3: Average score given from AMT workers to counter narratives generated by each generation model used in our evaluation set including standard deviation.