MilaNLP@Multilingual Counterspeech Generation: Evaluating Translation and Background Knowledge Filtering

Emanuele Moscato, Arianna Muti, Debora Nozza
Bocconi University, Milan, Italy
{emanuele.moscato2, arianna.muti, debora.nozza}@unibocconi.it

Abstract

We describe our participation in the Multilingual Counterspeech Generation shared task, which aims to generate a counternarrative to counteract hate speech, given a hateful sentence and relevant background knowledge. Our team tested two different aspects: *(i)* translating outputs from English vs generating outputs in the original languages and *(ii)* filtering pieces of the background knowledge provided vs including all the background knowledge. Our experiments show that filtering the background knowledge in the same prompt and leaving data in the original languages leads to more adherent counternarrative generations, except for Basque, where translating the output from English and filtering the background knowledge in a separate prompt yields better results. Our system ranked first in English, Italian, and Spanish and fourth in Basque.

1 Introduction

Hate speech (HS) poses a significant challenge in online spaces, fostering division and perpetuating discrimination. The need for effective interventions becomes increasingly urgent. Among the various strategies for countering hate speech, counternarrative generation (CNG) has emerged as a promising approach (Bonaldi et al., 2024a). Rather than simply removing harmful content, counternarratives aim to actively challenge hate speech by offering constructive, persuasive and non-polarized discourse, which might offer alternative standpoint both to the author of the hate speech message and to users navigating the online web and running into hateful comments. The Multilingual Counterspeech Generation shared task proposes to address this problem by asking participants to generate counterspeech for multiple targets (Jews, LGBT+, migrants, people of color, and women) and languages (Basque, English, Italian, and Spanish), with texts in languages other than English being translations from their English counterparts. The shared task data also comprises *background knowledge* (BK) sentences, which may be helpful to generate the counternarratives. This system paper describes our approach to the shared task.

During a preliminary manual evaluation of LLMs' outputs, we observed two issues that could potentially compromise the quality of counternarrative generation. First, the models produced low-quality text in languages other than English, inventing non-existent words (e.g., the nonexisting Italian word "contini") or generating ungrammatical sentences (e.g., the incorrect Italian article in "Non c'è posto per *la* odio"). Second, the background knowledge included in the data was often not only unhelpful but also interfered with the logical flow of the generated counternarratives. For instance, the model confused the figurative meaning of "iron fist" (i.e., exercising power in an oppressive or ruthless manner) with its literal meaning (i.e., a punch).

For this reason, our system submission focused on two key questions: *(i)* For languages other than English, is it better to ask the model to generate responses in that language, or should it generate them in English and then be translated? *(ii)* Is it better to filter the background knowledge sentences (in one or two separate steps), or should all of them be used?

Our results demonstrate that the optimal approach involves: *(i)* providing the model with input data in its original language and generating responses in that same language, and *(ii)* filtering the background knowledge in a single step within the same prompt rather than in two different steps. The best performance is still achieved by models that generate counternarratives directly in the target language, regardless of potential grammatical issues, likely because the content is more important than grammatical accuracy. **Our system achieved first place in three out of the four languages** in the shared task: English, Spanish, and Italian.

Proceedings of the First Workshop on Multilingual Counterspeech Generation (MCG), pages 56–64
January 19th, 2025.

2 Related Work

Counterspeech or counternarrative (the terms are used interchangeably in the NLP community) is the strategic response to a hate speech message that provides an opposing stance, aiming at changing the hate-related viewpoint, by not attacking the interlocutor but the content of the message (Bonaldi et al., 2024a). Countering hate speech through the generation of counternarratives provides a constructive and pro-active approach to hate speech that goes beyond mere detection. To do so, several datasets have been developed. 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 only on Islamophobia. Moreover, they introduce a taxonomy for the following types of CNs: Presentation of Facts, Pointing out Hypocrisy Or Contradiction, Warning Of Consequences, Affiliation, Positive Tone, Negative Tone, Humor, Counter-Questions, Other. Then, with MultiTarget CONAN (MT-CONAN), Fanton et al. (2021b) expand on the previous dataset by creating 5000 HS/CN pairs in English Language, covering multiple hate targets, in terms of race, religion, country of origin, sexual orientation, disability, or gender.

Research on counternarrative generation (CNG) has increased due to LLMs' impressive performance in generating text (Zubiaga et al., 2024). However, often the generated CN is beautifully written but generic, repetitive and poor in terms of content, which should show credible evidence, factual arguments and alternative viewpoints by adopting an empathetic, polite and constructive tone (Fanton et al., 2021a; Chung et al., 2021; Bonaldi et al., 2024a). Generating effective counternarratives necessitates a deep understanding of cultural, historical and social factors mentioned in the hateful instances. For this reason, CNG benefits from the use of background knowledge or knowledge retrieval to generate text, which makes it a close task to counter-argumentation and misinformation countering (Bonaldi et al., 2024a). Therefore, the CNG task should foresee two steps: first the extraction of relevant knowledge from an external source, and secondly the generation of knowledge-augmented counterspeech. This approach has been proposed by Chung et al. (2021) through extracted and generated keyphrases and by Jiang et al. (2023b), who extract background knowledge relevant to hate speech with an opposite stance in an unsupervised fashion. They retrieve and filter information from multiple perspectives of stance consistency, semantic overlap rate between the knowledge retrieved and the hateful message, and fitness for hate speech. Bonaldi et al. (2024b) show that the presence of safety guardrails in LLMs hinders the quality of the generations. Moreover, since hate speech is often expressed through implicit arguments (Muti et al., 2024a), Bonaldi et al. (2024b) decompose the hate speech into premises and conclusion, showing that attacking a specific component of the hate speech, in particular its implied statement, leads to richer argumentative generations.

3 Data

The data consists of 596 hateful messages, each appearing in four languages (English, Spanish, Basque, and Italian), for a total of 2384 datapoints across all languages.

The dataset is divided into 3 splits: development (400 instances across all languages), train (1584), and test (400). Each instance presents the following features:

- **HS**: a Hate Speech sentence, taken from the MTCONAN dataset (Fanton et al., 2021b).
- **BK**: up to 5 separate pieces of background knowledge (textual) that could be used to generate the counternarrative to the Hate Speech sentence.
- **CN**: a ground-truth counternarrative, generated by humans and present only in the development and train splits of the dataset.
- **LANG**: the language of the Hate Speech sentence, background knowledge and counternarrative (if present).
- **TARGET**: the social or ethnic group targeted by the Hate Speech sentence.
- **SPLIT**: the split of the dataset the datapoint belonged to.
- **MTCONAN_ID**: the ID of the datapoint in the MTCONAN dataset the Hate Speech sentence was taken from.
- **PAIR_ID**: an ID identifying the same datapoint **across all languages** (non-unique across the dataset, i.e. each value appeared four times, once for each language).

- **ID**: a concatenation of a string identifying the language and the **PAIR_ID** field, resulting in an identifier that is unique across the dataset.

Although the shared task permits the use of external data as background knowledge, we rely exclusively on the knowledge provided.

3.1 Metrics

Teams were asked to automatically generate counternarratives for the test split, which is then evaluated with several metrics, both automatic and LLM-based. For the automatic scores, organizers chose BERTscore (Zhang et al., 2019), BLEU (Papineni et al., 2002; Post, 2018), Rouge-L (Lin, 2004), and novelty (Tekiroglu et al., 2022). They also report the generation length. For the LLM-based, they opted for the "LLM as a judge" framework (*JudgeLM*) (Zubiaga et al., 2024). This framework evaluates generated CNs pairwise in a tournament-style format, assessing the quality of the generated counternarrative.

4 System Description

We develop an LLM-based pipeline for automatic counterspeech generation without fine-tuning. In particular, we compare the performance of Llama3-8B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a) and Zephyr-7B-beta (Tunstall et al., 2023), with Mistral emerging as the overall best-performing one from preliminary manual evaluation on ten instances. Moreover, Mistral shows the least refusal to answer, which makes it a good candidate since safety guardrails have proved to be detrimental to the generation of counternarratives (Bonaldi et al., 2024b). All models are provided via the Hugging Face model hub[1].

The prompt for counternarrative generation (see Appendix A) includes the following information:

- Hate speech statement
- Background knowledge sentences
- Targeted social/ethnic group
- Language of the provided text and language in which to generate the counternarrative.

Furthermore, we explicitly instruct the model to avoid using any information beyond the provided background knowledge, assuming that stricter adherence results in more factual counternarratives.

[1] https://huggingface.co/models

Runs	Translation	Filtering
1	Y	Y*
2	N	Y*
3	N	Y
4	Y	Y
5	N	N

Table 1: Summary of the conducted experiments. The Y* label denotes the separate-prompt filtering process.

Multilingual generation VS translation The complete dataset comprises four languages, with Basque being a low-resource language. Although the chosen LLM is able to generate text in all four languages, we expect that the quality may vary (and it can do so in ways that are hard to evaluate), especially for low-resource languages. Nozza (2021) and Muti and Barrón-Cedeño (2022) have exposed the limits on zero-shot classification of different forms of hate speech across languages on encoder-based models, due to the language- and culture-specific lexical variation of hate speech. Furthermore, during a preliminary manual evaluation, we identified certain challenges in generating text in languages other than English. These issues included the production of non-existent words and ungrammatical sentences.

To address this, we experimented with two approaches for generating text in languages other than English:

- generation directly in the target language;
- generation in English, with a subsequent translation in Spanish, Basque, and Italian.

The machine translation task is performed using the NLLB model (NLLBTeam et al., 2022).

These experiments were feasible because each hate speech sentence and background knowledge text in the dataset is available in all four languages.

Background knowledge filtering Upon examining a sample of the development and training data, we observed that some of the provided background knowledge sentences are not relevant to generating the corresponding counternarratives. We therefore experiment with:

- providing the LLM with all the background knowledge points, asking the model to choose which ones to use at inference time (same-prompt filtering),

Run	BERTScore				BLEU				Rouge-L				Novelty			
	EN	ES	EU	IT	EN	ES	EU	IT	EN	ES	EU	IT	EN	ES	EU	IT
1	0.710	0.716	0.692	0.710	0.049	0.055	0.016	0.046	0.187	0.203	0.110	0.171	0.805	0.781	0.873	0.803
2	0.711	0.734	0.708	0.722	0.047	0.087	0.072	0.075	0.189	0.239	0.183	0.206	0.804	0.755	0.831	0.781
3	0.706	0.733	0.712	0.726	0.044	0.088	0.072	0.075	0.179	0.233	0.190	0.207	0.813	0.761	0.833	0.785
4	0.708	0.714	0.689	0.708	0.045	0.049	0.014	0.044	0.181	0.197	0.108	0.170	0.814	0.792	0.880	0.809
5	0.715	0.738	0.719	0.734	0.059	0.097	0.081	0.092	0.200	0.246	0.204	0.229	0.810	0.757	0.828	0.776

Table 2: Results on the development set. The higher the better.

- first filtering the background knowledge points and then feeding the resulting subset to the LLM to generate the counternarrative (separate-prompt filtering) (see Appendix A for the prompt),

- avoiding any kind of filtering and just asking the model to generate a CN using the available BK.

A schema of the experiments can be found in Table 1.

5 Results

The results of our experiments on the development and train splits of the dataset are presented in Table 2. The best performance is achieved by run 5, which involves neither translation nor filtering of the background knowledge (BK). These results suggest that Mistral performs well in a simpler setup. However, upon closer inspection, the counternarratives generated in run 5 are of low quality, replicating the issues observed during the preliminary manual evaluation of a small subset. For this reason, we have decided to exclude this run from the final submission. Therefore, *the runs submitted to the shared task are 1, 2, and 3*, which according to a preliminary observation perform the best. The manual evaluation has been chosen over traditional metrics because the latter have been shown not to correlate well with human preferences when evaluating generation (Nimah et al., 2023). Table 3 shows the results on the test set with the JudgeLM metric used for the official ranking provided by the organizers, along with the position in the leaderboard. Run 3 - no translation, same-prompt filtering, performs the best, with respect to our other runs and other participants, in English, Italian, and Spanish. In Basque, on the other hand, our run 1, which employs translation and separate-prompt filtering, obtains the best score among our runs and fourth place in the leaderboard.

Run	JudgeLLM				Rank			
	EN	ES	EU	IT	EN	ES	EU	IT
1	2327	1853	2243	1824	8	5	**4**	4
2	2358	1942	430	1912	2	2	25	2
3	2523	2002	423	1986	**1**	**1**	26	**1**
Ground truth	1176	899	1535	930	-	-	-	-

Table 3: Results on the test set and scores of the ground truth counternarrative. We only report the metrics on which the ranking was based.

5.1 Error Analysis

In order to assess the two aspects investigated (translation vs original; same-prompt vs separate-prompts filtering vs no filtering), we perform a human evaluation on a random subset of 50 Italian instances, 10 for each target.

How does translation affect generation? We compare run 1 with run 2 to evaluate the differences between generating responses first in English and translating them into the target language versus generating responses directly in the target language. Although, in both runs, filtering has been done in a separate prompt, the BK chosen can be different for the same instances. This affects the final generation. Therefore, we do not evaluate the relevance to the ground truth, which highly depends on selected pieces of BK. We focus only on the quality of the language. In English, the responses are much more concise and less generic. Many instances generated in the original language begin with a vague structure, such as “Si tratta di un’affermazione ingiusta e offensiva nei confronti di un gruppo di persone sulla base della loro etnia", translated as “It is an unjust and offensive statement towards a group of people based on their ethnicity", or “Nessuno deve essere discriminato per la propria orientazione sessuale", translated as “nobody should be discriminated for their sexual orientation", or “Tutti gli esseri umani hanno il potenziale di raggiungere grandi successi, indipendentemente dalla loro origine etnica o colore della pelle", translated as “All human beings have the potential to achieve great success, regardless of their

HS	CN	BK	CN_run3	CN_run5
Jews are nothing but criminals.	Jews are not criminals, but Nazi were. The memory of the Jews murdered—murdered because they were Jews — must be recognized with historical accuracy.	Chodakiewicz noted that after World War II, the Jews were not only victims, but also aggressors.	It is important to remember that **while some Jews participated in the aggression after World War II, they were also victims**.	Moreover, after World War II, **Jews were not only victims, but also aggressors**, as noted by Chodakiewicz.

Table 4: Dataset example and generated counternarratives. Both run 3 and run 5 select that BK to be used in the prompt. However, while run 3 transforms it to generate counterspeech, run 5 tends just to repeat BK points as they are, resulting in producing hate speech.

ethnic origin or skin color". Responses generated in the original language rarely show errors. Here is an error example: *Molto spiace, ma sono sbagliati quelle affermazioni* instead of the correct form *Mi dispiace molto, ma quelle affermazioni sono sbagliate*. This links us to the third aspect: politeness. Responses generated in the original language tend to be more polite than those generated in English.

On the other hand, responses translated from English contain more grammatical errors, including hallucinations of invented words that are similar to the original ones, such as *progreditiamo, ritrarrele* instead of *progrediamo, ritrarle* (*en:* we progress, depict them). Other categories of errors are wrong syntactic structures: *Il femminismo non è quello di svalutare* instead of *Il femminismo non svaluta* (*en:*feminism does not devalue); wrong coordination of gender: *Le donne non sono bambini* instead of *bambine*; lack of figurative language understanding: *non sono bambini lavati al cervello* instead of *non sono bambine a cui è stato fatto il lavaggio di cervello* (*en:* they are not brainwashed little girls); the figurative meaning of “iron fist” (i.e., exercising power in an oppressive or ruthless manner) confused with its literal meaning (i.e., a punch).

How does filtering BK affect generation? To answer this question, our manual evaluation focuses on the relevance of the different BK selected with respect to the ground truth CN. We compare the best-performing model, which uses filtering, run 3, with run 5, which does not filter BK. We investigate whether the LM in run 5 considers all the BK points by adding irrelevant information or whether it filters information deliberately. We notice that run 5 generations are much longer, and they contain all BK sentences. Therefore, the model does not have any implicit filtering. This pollutes the generated output, which is affected by irrelevant and wrong/hateful information. An example can be seen in Table 4, where the piece of BK selected produces more hate speech instead of countering it. In addition, we compare same-prompt with separate-prompt filtering. We discard the translated version since it did not yield better results. By observing the same subsample across run 2 and run 3, we observe that run 2 produces shorter responses, by selecting only 2 or 3 BK pieces every time. However, it tends to select irrelevant and hateful BK, like the one in Table 4 or *è stato dimostrato che molte comunità, incluse quelle religiose, possono essere dominanti o abusare del loro potere* (en: *it's been proved that many communities, including religion ones, can be dominant or abuse of their power*, which are discarded by filtering the BK in the same prompt. Therefore, addressing the two tasks together is better than separately. This tendency has been observed in other hate-related tasks (Muti et al., 2022, 2024b). However, even in instances with the highest scores, generated responses tend to rely exclusively on the BK, without providing a logical link between the BK and a final statement to counter hate, which occurs in the ground truth CN.

6 Conclusion

We presented our approach to the knowledge-grounded generation of counternarratives by investigating two aspects: *(i)* generating in English and then translating to the target language vs generating in the original language and *(ii)* filtering (either within the same prompt or in a separate prompt as a preliminary step) vs feeding the model with all the knowledge pieces. The human evaluation performed on the development set shows a contrast in the results. Run 5, which is the simplest setting - no translations nor BK filtering - results in the best-performing run based on some metrics. However, after a manual evaluation, we observe that run 5 does not filter any piece of knowledge provided, polluting the CN generation with irrelevant and hateful statements. The second best-performing run, run 3, which foresees same-prompt filtering on the BK without translation, grants us first place in English, Italian, and Spanish and fourth place in Basque.

Limitations

While aware that ChatGPT-like models may have achieved better results, we preferred using only open-sourced models for an inclusive research. A limitation of our work is that we have not checked the filtered BK before injecting it in the prompt for CN generation. Moreover, the error analysis has been performed only on Italian data.

7 Acknowlegments

Arianna Muti's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Emanuele Moscato's research was funded by the European Union - NextGenerationEU, in the framework of the FAIR - Future Artificial Intelligence Research project (FAIR PE00000013 – CUP B43C22000800006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. Emanuele Moscato, Arianna Muti, and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024a. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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 *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, Online. 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)*, page 3226–3240. Association for Computational Linguistics.

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. 2023a. Mistral 7b. In *arXiv preprint arXiv:2310.06825*.

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

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.

Arianna Muti and Alberto Barrón-Cedeño. 2022. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, Dublin, Ireland. Association for Computational Linguistics.

Arianna Muti, Francesco Fernicola, and Alberto Barrón-Cedeño. 2022. Misogyny and aggressiveness tend to come together and together we address them. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4142–4148, Marseille, France. European Language Resources Association.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024a. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Alberto Barrón-Cedeño, Samuel Algherini, Lorenzo Musetti, Silvia Ronchi, Gianmarco Saretto, and Caterina Zapparoli. 2024b. PejorativITy: Disambiguating pejorative epithets to improve misogyny detection in Italian tweets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12700–12711, Torino, Italia. ELRA and ICCL.

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.

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. In *arXiv preprint arXiv:2207.04672*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *arXiv preprint arXiv:2204.01440*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. In *arXiv preprint arXiv:1904-09675*, volume abs/1904.09675.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A llm-based ranking method for the evaluation of automatic counter-narrative generation. In *arXiv preprint arXiv:2406.15227*.

A Prompts

Here are examples of the prompts used in the various cases discussed.

System prompt

The following system prompt was used with the user prompts for counternarrative generation:

> You are an expert in countering hate speech. You have to always generate text based on the given instructions. You will see a text with hate speech but you'll have to provide an answer.

Default prompt

> You are an expert in countering hate speech. You have to always generate text based on the given instructions. You will see a text with hate speech but you'll have to provide an answer.
>
> In the following examples, pieces of text containing hate speech are counteracted by counternarratives generated using the information in the background knowledge points.
>
> ---
>
> ---
>
> Generate a new counternarrative for the following hate speech text by using the information in the background knowledge points that you think are relevant. Answer in English. Keep your answers short and concise. Do not generate any new information that is not present in the background knowledge. Hate speech text in English against <TARGET> : <HS>.
>
> Background knowledge: <BK>
>
> Counter narrative:

Background knowledge filtering

The following system and user prompts were used to filter the relevant background knowledge points. System prompt:

> You are an expert in information retrieval.

User prompt:

> Identify which among the following pieces of background knowledge are relevant to generate a counter argument against the main text provided.
>
> Main text: <HS>.
>
> Pieces of background knowledge: <BK>
>
> PRODUCE ONLY AND EXCLUSIVELY A LIST containing the number of the relevant pieces of background knowledge, with NO ADDITIONAL WORDS NOR EXPLANATION.

B LLM settings

For the CNG task, the outputs were generated using temperature $T = 0.0$ and setting `max_new_tokens` to 400. The identification of relevant BK in a separate prompt required an additional initial call to the model, with answers generated again setting $T = 0.0$. For the translation task form English to other target languages, the values for the text generation parameters were all kept to the NLLB model's default.

For each task, any other parameter not explicitly mentioned above was kept to default value.