TrenTeam at Multilingual Counterspeech Generation: Multilingual Passage Re-Ranking Approaches for Knowledge-Driven Counterspeech Generation Against Hate

Daniel Russo
University of Trento, Italy
Fondazione Bruno Kessler, Italy
drusso@fbk.eu

Abstract

Hate speech (HS) in online spaces poses severe risks, including real-world violence and psychological harm to victims, necessitating effective countermeasures. Counterspeech (CS), which responds to hateful messages with opposing yet non-hostile narratives, offer a promising solution by mitigating HS while upholding free expression. However, the growing volume of HS demands automation, making Natural Language Processing a viable solution for the automatic generation of CS. Recent works have explored knowledge-driven approaches, leveraging external sources to improve the relevance and informativeness of responses. These methods typically involve multi-step pipelines combining retrieval and *passage re-ranking* modules. While effective, most studies have focused on English, with limited exploration of multilingual contexts. This paper addresses these gaps by proposing a multilingual, knowledge-driven approach to CS generation. We integrate state-of-the-art re-ranking mechanisms into the CS generation pipeline and evaluate them using the MT-CONAN-KN dataset, which includes hate speech, relevant knowledge sentences, and counterspeech in four languages: English, Italian, Spanish, and Basque. Our approach compares reranker-based systems employing multilingual cross-encoders and LLMs to a simpler end-to-end system where the language model directly handles both knowledge selection and CS generation. Results demonstrate that reranker-based systems outperformed end-to-end systems in syntactic and semantic similarity metrics, with LLM-based re-rankers delivering the strongest performance overall.[1]

Content warning: this paper contains unobfuscated examples some readers may find offensive

1 Introduction

Online spaces have become fertile ground for the proliferation of hateful content, which poses significant threats not only in digital environments but also in the offline world. Research highlights a direct connection between online hate speech and real-world violence (Awan and Zempi, 2016). Exposure to such content can severely impact the mental health of victims, fostering feelings of insecurity and exclusion (Saha et al., 2019; Persily et al., 2020; Dreißigacker et al., 2024).

Counterspeech (CS) – a strategy of responding to hateful messages with opposing, non-hostile narratives – emerges as a promising solution. Studies suggest that counterspeech can be more impactful than traditional moderation techniques like content removal or user bans, while also aligning with free speech principles (Schieb and Preuss; Fraser et al., 2021). Given the sheer volume of hateful content generated daily, researchers in Natural Language Processing (NLP) have increasingly focused on automating CS-related tasks, including classification (Chung et al., 2021a; Mathew et al., 2019), data curation (Chung et al., 2019; Fanton et al., 2021), and generation (Tekiroğlu et al., 2020; Chung et al., 2021b; Zhu and Bhat, 2021; Tekiroğlu et al., 2022).

Although the majority of the NLP work on counterspeech has centred on English, recent studies have expanded this scope to other languages. For instance, datasets and generation systems now exist for Italian (Chung et al., 2019; Fanton et al., 2021), French (Chung et al., 2019), Spanish (Vallecillo Rodríguez et al., 2024; Bengoetxea et al., 2024), and Basque (Bengoetxea et al., 2024). Despite these advancements, multilingual research remains underexplored, particularly in terms of cross-lingual adaptability and scalability.

Another promising frontier in CS generation is knowledge-grounded approaches, which can help improve the model's accuracy and lead to CS more

[1] This work is the result of our participation in the *Shared Task on Multilingual Counterspeech Generation* held at COLING 2025.

Proceedings of the First Workshop on Multilingual Counterspeech Generation (MCG), pages 77–91
January 19th, 2025. ©2025 Association for Computational Linguistics

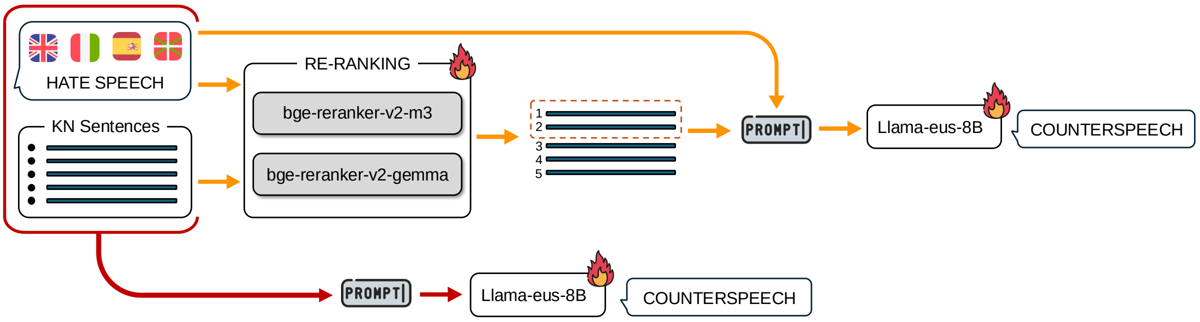

Figure 1: Graphical representation of the experimental design: orange lines indicate the `Rerank-CS` approach, red lines the `E2E Prompt-CS` approach. The fire emoji symbolizes model fine-tuning.

aligned with those produced by experts. By incorporating external knowledge (KN) sources, such as Wikipedia or discussion forums, these methods improve the relevance and informativeness of generated responses (Chung et al., 2021b; Jiang et al., 2023). For example, Chung et al. (2021b) leverage keyphrase extraction for KN retrieval, while Jiang et al. (2023) utilize metrics such as stance consistency to construct KN repositories. Both studies integrate the retrieval phase with a *passage re-ranking* module (Nogueira and Cho, 2019), enabling the fine-grained selection of retrieved KN sentences to be passed to the language model. Specifically, Chung et al. (2021b) propose using the `ROUGE-L` metric (Lin, 2004) to identify the most relevant sentences for countering hate speech, whereas Jiang et al. (2023) employ a fitness function for sentence selection. However, these techniques have primarily been developed and evaluated in English, leaving a significant gap in multilingual contexts.

In this paper, we aim to bridge this gap by proposing a multilingual, KN-driven approach to CS generation. Specifically, we focus on enhancing the *passage re-ranking* module by incorporating state-of-the-art re-ranking mechanisms into the KN-driven CS generation pipeline. To evaluate our approach, we tested the performance of multilingual cross-encoders and LLM-based re-rankers on the `MT-CONAN-KN` dataset.[2] We compared reranker-based systems to a simpler end-to-end approach, where all available information – hate speech and retrieved KN – was directly passed to an LLM tasked with selecting the appropriate KN and generating a CS grounded in it. Figure 1 graphically summarizes the proposed systems.

This work represents the outcome of our participation[3] in the *Multilingual Counterspeech Generation Shared Task*, organized as part of the *First Workshop on Multilingual Counterspeech Generation* (`MCG@COLING 2025`).[4] Results demonstrate that reranker-based systems achieved outstanding performance in terms of syntactic and semantic similarity with the `MT-CONAN-KN` test set, outperforming other systems in the competition. Additionally, LLM-based re-rankers produced better results on average according to these metrics. However, when evaluated using LLM-based metrics, the systems' performance was comparable to those tested on the `MT-CONAN-KN`, indicating strong alignment with the competition dataset but relative weakness in generating generally high-quality CS.

Although preliminary, these findings underscore the importance of passage re-ranking for KN-driven CS generation, particularly in multilingual contexts. Nonetheless, further research is necessary to develop high-quality, domain-specific KN bases and to refine retrieval strategies to enhance CS generation.[5]

2 Related Work

Although interest in CS generation is growing, most existing approaches rely on fine-tuning language models on ad-hoc datasets (Qian et al., 2019; Tekiroğlu et al., 2022; Halim et al., 2023) or, in more recent research, employing *in-context learning* techniques (Doğanç and Markov, 2023; Mun et al., 2023; Zheng et al., 2023). However, very

[2] https://huggingface.co/datasets/LanD-FBK/ML_MTCONAN_KN

[3] We participated as the *TrenTeam*.

[4] sites.google.com/view/multilang-counterspeech-gen/

[5] The code and data are publicly available in the following GitHub repository: https://github.com/drusso98/TrenTeam-MCG2025/

few steps have been taken toward a KN-driven generation of CS.

Efforts towards KN-driven CS generation remain limited due to two primary challenges: (i) the common lack of explicit, well-structured facts in hate speech (HS) and (ii) the scarcity of training data (Chung et al., 2021b). To address these challenges, Chung et al. (2021b) proposed to prepend to the generative step a KN retrieval one. To address the limitation of the lack of explicit facts in the HS, the authors developed a query generation module to extract keywords from HS instances in the CONAN dataset (Chung et al., 2019). These keywords were then used in a two-step KN retrieval procedure: first, a retrieval step of the top 25 relevant articles from a KN base comprising the Newsroom (Grusky et al., 2018) and WikiText-103 (Merity et al., 2016) datasets using BM25 (Robertson et al., 2009); second, a selection step of the top 5 most relevant sentences from these articles using the ROUGE-L metric (Lin, 2004). The retrieved sentences were combined with the HS instance to form a single input, which was then passed to generative models such as GPT-2 (Radford et al., 2019) and XNLG (Chi et al., 2020), fine-tuned for this purpose.

More recently, Jiang et al. (2023) introduced the RAUCG framework for unsupervised retrieval-augmented CS generation. Like Chung et al. (2021b), the RAUCG framework comprises two components: a KN retriever and a CS generator. Using data from the ChangeMyView subreddit[6], the retrieval module employed a multi-step process. This included stance consistency and semantic overlap rate to select counter-comments relevant to the HS post, ensuring these contained effective counter-arguments. The framework further refined the retrieved comments using a custom-designed *fitness function*, computed in terms of perplexity, to identify the most suitable sentences. Finally, the HS and the selected sentences were utilized to generate the CS through energy-based decoding, which was constrained to preserve the retrieved KN and counter the corresponding HS, all while ensuring fluency.

Both approaches emphasize the importance of fine-grained selection of effective sentences or counter-arguments to ensure that the retrieved KN provided as input to generative models is both appropriate and effective. Chung et al. (2021b) assessed sentence relevance based on textual overlap using the ROUGE-L metric, whereas Jiang et al. (2023) ranked sentences based on the model's confidence in next-word prediction (perplexity). The importance of assessing sentence relevance for KN-driven generation is also reflected by the growing emphasis on *passage re-ranking* within Retrieval-Augmented Generation (RAG; Lewis et al., 2020) systems. Indeed, recent advancements in RAG demonstrate that passage re-ranking is a critical step for improving retrieval performance (Nogueira and Cho, 2019), which ultimately enhances generation quality. State-of-the-art approaches increasingly utilize cross-encoders for passage re-ranking, which process query and passage information jointly to generate a relevance score. Although more computationally intensive than traditional bi-encoders (Reimers and Gurevych, 2019; Lin et al., 2023), cross-encoders provide superior performance by capturing the semantic relationship between query and passage more effectively. With the advent of LLMs, recent methods have also employed generative models for passage re-ranking by prompting the model to reason over query-passage pairs and output entailment labels (e.g., *true/false* or *yes/no*). The ranking score is then derived from the logits associated with the positive label (Zhuang et al., 2023; Li et al., 2024).

This work seeks to advance KN-driven CS generation by leveraging the latest developments and technologies in passage re-ranking and applying them to hate-speech countering. Specifically, we evaluate two re-ranking-based CS generation approaches and compare them with an end-to-end prompt-based generation approach. Additionally, we explore these methodologies in a multilingual setting using the MT-CONAN-KN dataset, which contains triplets of HS, a list of related KN sentences, and a CS written using one or more of the KN sentences across four languages: English, Italian, Spanish, and Basque.

3 Dataset

All systems developed in this study are based on the *Multilingual Multi-Target Knowledge-based CONAN dataset* (**ML_MTCONAN_KN**)[2], provided by the organizers of the *Multilingual Counterspeech Generation Shared Task* (MCG@COLING 2025[4]). The ML_MTCONAN_KN dataset is built upon the *Multi-Target CONAN dataset* (MT-CONAN; Fanton et al., 2021), which contains 5003 English HS-CS pairs addressing multiple hate targets, including *dis-*

[6] https://www.reddit.com/r/changemyview/

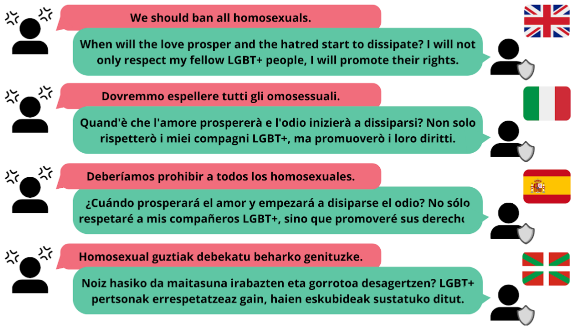
We should ban all homosexuals.
When will the love prosper and the hatred start to dissipate? I will not only respect my fellow LGBT+ people, I will promote their rights.
Dovremmo espellere tutti gli omosessuali.
Quand'è che l'amore prospererà e l'odio inizierà a dissiparsi? Non solo rispetterò i miei compagni LGBT+, ma promuoverò i loro diritti.
Deberíamos prohibir a todos los homosexuales.
¿Cuándo prosperará el amor y empezará a disiparse el odio? No sólo respetaré a mis compañeros LGBT+, sino que promoveré sus derech
Homosexual guztiak debekatu beharko genituzke.
Noiz hasiko da maitasuna irabazten eta gorrotoa desagertzen? LGBT+ pertsonak errespetatzeaz gain, haien eskubideak sustatuko ditut.

Figure 2: Example of an HS-CS pair from the ML_MTCONAN_KN dataset in English, Italian, Spanish, and Basque. Image sourced from the official website of the MCG Shared Task at COLING 2025.

abled, *Jews*, *LGBT+*, *migrants*, *Muslims*, *people of color*, and *women*. From this dataset, a subset of 596 HS instances was sampled to construct the ML_MTCONAN_KN dataset, focusing on five hate targets: *women*, *migrants*, *Jews*, and *people of color*. For each HS instance, five KN sentences were collected, and a novel CS was written using one or more of these KN sentences.

The resulting English dataset was automatically translated into Italian, Spanish, and Basque. To ensure high-quality translations, native speakers of each target language manually post-edited the CS. The final dataset comprises 2384 entries, divided into the following subsets: a training set with 396 HS-CS pairs per language, and development and test sets with 100 pairs per language each. Figure 2 illustrates an example of an HS-CS pair translated into the four languages included in the ML_MTCONAN_KN dataset.

4 Experimental Design

In this work, we compare two CS generation approaches. In the *first* approach we tested KN-driven CS generation leveraging multilingual re-rankers to identify the most relevant KN sentences for a given HS. The selected sentences were eventually passed to the LLM to guide its generation of the CS (Rerank-CS approach).

The *second* approach employs a prompt-based method where a multilingual LLM is directly fine-tuned to 'reason' over the entire set of KN sentences, identify the most relevant, and produce the CS in a single, end-to-end process. (E2E Prompt-CS approach). Figure 1 provides a graphical overview of the experimental design. In the following sections, we provide further details of the two approaches proposed.

4.1 Rerank-CS Approach

For the Rerank-CS approach, we tested two multilingual re-rankers: the lightweight bge-reranker-v2-m3[7] and the LLM-based bge-reranker-v2-gemma[8] (Chen et al., 2024). Both the re-rankers are part of the *BGE (BAAI General Embeddings)* family of embedding models and were chosen for two main reasons: (i) while carrying out the experiments they were the only re-rankers that *officially* supported all four languages, i.e. English, Italian, Spanish, and Basque; (ii) they were ranked high in the Massive Text Embedding Benchmark (MTEB) Leaderboard (Muennighoff et al., 2022).

The bge-reranker-v2-m3 model (**M3_RRank** hereafter) is a lightweight, multilingual cross-encoder based on the BGE-M3 model (Chen et al., 2024). It was built upon the XLM-RoBERTa pretrained model (Conneau et al., 2019) and fine-tuned on extensive unlabeled, labelled, and synthetic corpora. The bge-reranker-v2-gemma (**Gemma_RRank** hereafter), on the other hand, is a multilingual LLM-based re-ranking model with the Gemma-2B model (Team et al., 2024) as its backbone. This generative model is utilized for a binary classification task, employing the logits of the positive response (e.g., *'true'* or *'yes'*) to represent the final ranking score.

To evaluate re-rankers on their ability to rank KN sentences by relevance to hate speech, an annotated version of the MT-CONAN-KN dataset is needed. In this annotated version, for each entry, the KN sentence(s) used to write the CS are identified and labelled. This annotated dataset will also serve to fine-tune the re-rankers. The remainder of this section outlines the automatic annotation procedure, as well as the strategies for fine-tuning and evaluating the re-rankers.

KN Sentences Annotation A qualitative analysis of the MT-CONAN-KN dataset revealed substantial overlap between the CS and one or more KN sentences. To identify which KN sentences were used to compose each CS, we employed ROUGE-L (Lin, 2004) as a metric, which measures the similarity between texts based on their common longest common subsequences (LCS; Lin, 2004). Specifically, we calculated the ROUGE-L score between each CS sentence and the corresponding KN sentences. For cases where a single KN sentence was used to

[7] https://huggingface.co/BAAI/bge-reranker-v2-m3
[8] https://huggingface.co/BAAI/bge-reranker-v2-gemma

write multiple CS sentences, we also computed the ROUGE-L score for the entire CS against each KN sentence. Finally, we kept the highest ROUGE-L scores for each KN sentence; the KN sentences whose ROUGE score exceeded a specific threshold were labelled as those used to write the CS. In particular, a threshold of 35% was chosen to ensure at least one positive sentence per hate speech in the training set. [9] Further details and a graphical representation of the KN sentence annotation process can be found in Appendix A.

Re-Ranker Fine-Tuning To fine-tune a re-ranker, each hate speech instance requires a list of positive and negative passages (KN sentences in our case). Using the annotated KN sentences, we considered sentences with ROUGE scores above the threshold as positive examples. Sentences with ROUGE-L scores below this threshold were treated as negative examples. These triplets of hate speech, positive KN sentences, and negative KN sentences (along with their ROUGE scores) were then used to fine-tune the two re-rankers, i.e. M3_RRank and Gemma_RRank. The Gemma_RRank necessitates a prompt in the input that specifies the classification task to be performed for the extraction of the ranking score. The prompt used is shown below.

> Given an hateful content A and a possible argument B against it, determine whether the argument is an effective reply providing a prediction of either 'Yes' or 'No'.[10]

In Appendix B.1 we provide further re-rankers fine-tuning details.

Re-Ranker Evaluation We evaluated the performance of the M3_RRank and Gemma_RRank re-rankers on the MT-CONAN-KN dev set on the task of scoring KN sentences based on their relevance to the corresponding HS. To measure the effectiveness of the re-rankers, we employed *Mean Average Precision* (MAP)[11], a metric that computes the average precision at each relevant position in the ranked list, offering a comprehensive evaluation of ranking quality. MAP is particularly useful in ranking tasks like this, as it rewards systems that place relevant items (in this case, gold KN sentences) higher in the ranking. Gold KN sentences were identified using ROUGE scores between KN sentences and CS sentences, as detailed earlier. The MAP results for both re-rankers, in their off-the-shelf and fine-tuned versions, are presented in Table 1.

Model	All	EN	IT	ES	EU
M3_RRank	0.637	0.625	0.659	0.648	0.616
M3_RRank FT	**0.753**	**0.772**	**0.753**	**0.753**	**0.732**
Gemma_RRank	0.670	0.660	0.687	0.685	0.647
Gemma_RRank FT	**0.764**	**0.782**	**0.792**	**0.780**	**0.702**

Table 1: Mean Average Precision results for M3_RRank and Gemma_RRank re-rankers, with and without fine-tuning (FT). We present results on the entire development set, as well as partial results on the language-specific subsets.

Our analysis shows that fine-tuning significantly enhances the performance of both re-rankers across all languages. Additionally, the Gemma_RRank model consistently outperforms the M3_RRank model, both with and without fine-tuning, indicating superior ability in ranking the most relevant KN sentences higher. Interestingly, the fine-tuned M3_RRank shows better MAP scores in Basque when compared to Gemma_RRank.

Counterspeech Generation The re-ranker module was followed by a KN-driven generation step, where the input consisted of the HS and the relevant KN sentences selected by the re-ranker. Following the automatic annotation of the KN sentences, we noticed that in the training set, on average, two KN sentences were used to write the CS. Therefore, in the generation phase, we provided the LLM with the top two previously ranked KN sentences. In particular, we employed the Llama-eus-8B model (Corral et al., 2024), the only open LLM that officially claims to be trained in all four languages present in the dataset. The Llama-eus-8B model is a multilingual adaptation of Meta's Llama3.1-8B, specifically tailored for the Basque language while retaining its multilingual capabilities.

We fine-tuned this model using the newly annotated version of the dataset. The input for fine-tuning was structured as follows:

> You will be provided with a hateful comment (hate speech) and 2 sentences comprising arguments against the comment (knowledge). Generate a reply to the hateful content using only the informa-

[9] A subset of the annotated data has been manually checked to ensure the effectiveness of this annotation strategy.

[10] This prompt was a slight modification of the default originally used for developing the Gemma_RRank.

[11] MAP was computed using the ranx library (Bassani, 2022).

Lang.	System	JudgeLM Score	rougeL (%)	bleu (%)	bertscore (%)	novelty (%)	gen_len
EN	`Rerank-CS M3_RRank`	1.056,0	49,6	45,3	82,0	78,0	34,4
	`Rerank-CS Gemma_RRank`	**1.145,5** ↓	**53,9**	**48,3**	**83,4**	78,1	36,3
	`E2E Prompt-CS`	999,5	52,5	43,3	82,2	**79,0** ↑	35,4
	Gold	1.175,5	100,0	100,0	100,0	77,7	32,7
IT	`Rerank-CS M3_RRank`	880,0	46,4	38,6	81,2	77,9	37,8
	`Rerank-CS Gemma_RRank`	**965,5** ↑	**48,6**	**41,2**	**81,7**	77,8	37,0
	`E2E Prompt-CS`	791,0	47,4	37,9	80,9	**78,8** ↑	35,5
	Gold	929,5	100,0	100,0	100,0	77,9	35,3
ES	`Rerank-CS M3_RRank`	879,0	48,2	39,3	81,7	**75,8** ↑	41,2
	`Rerank-CS Gemma_RRank`	**987,5** ↓	**51,6**	**42,9**	**82,8**	75,6	40,9
	`E2E Prompt-CS`	769,0	50,2	40,3	82,0	75,4	37,9
	Gold	899,0	100,0	100,0	100,0	75,1	36,9
EU	`Rerank-CS M3_RRank`	1.364,5	**33,8**	**22,4**	**77,6**	85,2	28,2
	`Rerank-CS Gemma_RRank`	**1.394,5** ↓	32,8	20,9	77,1	85,7	27,5
	`E2E Prompt-CS`	1.246,0	31,7	18,2	76,6	**85,9** ↑	24,0
	Gold	1.534,5	100,0	100,0	100,0	85,3	26,5
All	`Rerank-CS M3_RRank`	1044,9	44,5	36,4	80,6	79,2	35,4
	`Rerank-CS Gemma_RRank`	**1123,3** ↓	**46,7**	**38,3**	**81,3**	79,3	35,4
	`E2E Prompt-CS`	951,4	45,5	34,9	80,4	**79,8** ↑	33,2
	Gold	1134,6	100,0	100,0	100,0	79,0	32,9

Table 2: Generation results with the three systems: `Rerank-CS` with `M3_RRank` and `Gemma_RRank` and the `E2E Prompt-CS` approaches. We highlighted the best system among the three in bold. The arrows denote improved or diminished performance relative to test set data (*Gold* line). We have highlighted the results that achieved first, second, and third place in the general leaderboard of the shared task with gold, silver, and bronze respectively (Complete Results in Appendix C).

> tion present in the knowledge. Reply in the following language: {language}
>
> Hate speech: {hate speech}
>
> Knowledge: {KN sentences}
>
> Reply: {counterspeech}

For further details on the fine-tuning process, refer to Appendix B.2.

4.2 E2E Prompt-CS Approach

We evaluate the reranking-based generation methods against a more direct approach where the 'reasoning' abilities of a multilingual LLM have been examined. In particular, we tested an end-to-end prompt-based approach for selecting relevant sentences from the five KN sentences and eventually generating CS based on these sentences. For this approach, we employed `Llama-eus-8B` LM. Initial zero-shot experiments with `Llama-eus-8B` revealed that the model's output quality was suboptimal, as it tended to over-generate, including irrelevant information in the CS. Consequently, we decided to fine-tune the model to improve its performance and to adapt the CS style to the training examples. To this end, we provide the model with the HS input and five related KN sentences during both training and inference. The model is then prompted to identify the most relevant sentences and generate a CS that effectively leverages the provided information. Details on the fine-tuning procedure can be found in Appendix B.2.

5 Results and Discussion

The `Rerank-CS` and `E2E Prompt-CS` generation systems were evaluated using both automatic metrics and LLM-based metrics. The evaluation of the generated CS was conducted against the gold CS across several dimensions: *recall* and *precision* of the overlapping LCS and unigrams, calculated using `ROUGE-L` (Lin, 2004) and BLEU (Papineni et al., 2002), respectively; *novelty*, quantified via the Jaccard distance between the set of words from the generated CS and the words present in the train set (Wang and Wan, 2018; Jaccard, 1901); *semantic similarity*, measured using `BERTScore` (Zhang et al., 2020), by leveraging the pre-trained contextual embeddings of the `bert-base-multilingual-cased` model (Devlin et al., 2019). To further assess the quality of the generated CS, an LLM-based metric was applied. `JudgeLM`, a `Vicuna`-based model (Chiang

HATE SPEECH

Women should stay at home and raise the children, this should be their only job.

KN SENTENCES

1) Fifty years ago, women working full-time earned just 59 cents for every dollar their male counterparts made.
2) Close to 60% of women also said that working online allowed them to spend more time at home with their families.
3) Women now working, in the absence of necessity, said they would prefer part-time (41 percent) over full-time (17 percent) employment.
4) Who would have thought that only 17% of the US Congress would be women?
5) Only 80 different women were cited, with Lady Gaga and Angela Merkel each mentioned four times.

COUNTERSPEECH

Fifty years ago, women working full-time earned just 59 cents for every dollar their male counterparts made. Women now working, in the absence of necessity, said they would prefer part-time (41 percent) over full-time (17 percent) employment. Women should be free to choose their own job. [KN 1-3]

Cinquant'anni fa, le donne che lavoravano a tempo pieno guadagnavano solo 59 centesimi per ogni dollaro guadagnato dai loro colleghi maschi. Questo non è un buon motivo per dire alle donne che dovrebbero stare a casa e crescere i figli. Inoltre, quasi il 60% delle donne ha dichiarato che lavorare online permette loro di trascorrere più tempo a casa con la famiglia. [KN 1-2]

Hace cincuenta años, las mujeres que trabajaban a tiempo completo ganaban sólo 59 céntimos por cada dólar que ganaban sus homólogos masculinos. Cercan del 60% de las mujeres afirmaron también que trabajar en línea les permitía pasar más tiempo en casa con sus familias. ¿Qué es lo que no entiendes? [KN 1-2]

Emakumeek etxean geratu eta umeak hazi beharko lituzkete, hau izan beharko litzateke euren lan bakarra? Emakumeen % 60k esan zuen Internet bidez lan egiteak aukera ematen ziela denbora gehiago pasatzeko etxean beren familiekin. Duela 50 urte, emakumeek 59 zentimo irabazten zituzten gizonek egindako dolar bakoitzeko. [KN 1-2]

Table 3: Examples of generations using the Rerank-CS approach combined with Gemma_RRank. Due to space constraints, only the English version of the hate speech and its corresponding KN sentences are reported. The generations in the four languages (English, Italian, Spanish, and Basque) originated from hate speech and KN sentences in those respective languages. The KN sentences chosen by the re-ranker and ultimately utilized by the LLM to guide the generation are indicated in square brackets at the end of each CS.

et al., 2023) fine-tuned on the JudgeLM-100K dataset, was used for English, Italian, and Spanish. For Basque, an ad-hoc fine-tuned version of Llama-eus-8B was employed. These models were adapted for the specific task of CS generation following the approach from Zubiaga et al. (2024).

Results are reported in Table 2. On average, the Rerank-CS system using Gemma_RRank demonstrated superior performance compared to other approaches. Interestingly, both Rerank-CS approaches achieved higher scores in terms of text overlap and semantic similarity with the gold CS (ROUGE-L and BLEU), while the E2e Prompt-CS approach outperformed the other systems in terms of *novelty*. A closer examination of individual language performance reveals that the Rerank-CS Gemma-RRank system outperformed other systems across all languages except Basque. For Basque, the lightweight M3-RRank yielded the best results in generation for overlap metrics (ROUGE-L and BLEU) and semantic similarity (BERTScore).

Additionally, the Rerank-CS Gemma-RRank system consistently received the highest scores from the *JudgeLM* model across all languages. Interestingly, the LLM-based evaluation recorded the highest scores for the Basque language. This phenomenon may be due to the fact that the generation model and the evaluation model were the same, namely, Llama-eus-8B. All systems enhanced the novelty in their outputs when compared to the gold CS. Nevertheless, the E2E Prompt-CS method consistently yielded the most novel results, with the exception of Spanish.

When considering overall results (see Appendix C), the Rerank-CS systems performed exceptionally well in overlap-based metrics (ROUGE-L, BLEU) and semantic similarity (BERTScore) across the four languages. This suggests that: (i) the fine-tuned re-rankers were generally able to assign higher scores to the proper KN sentences; (ii) the fine-tuned generative model successfully learned the task of generating according to the KN provided in input and properly adapted its output to align with the MT-CONAN-KN style, i.e, in generating CS that adhere to the KN sentences. However, these systems received lower rankings from the LLM-based judge as the generated CS adhered strictly to the MT-CONAN-KN style, which, when evaluated against CS generated by a less constrained model, may appear less flexible or creative.

Table 3 presents an example of generated CS in the four languages for the given hate speech input, utilizing the KN sentences previously selected by Gemma-RRank. A qualitative analysis of the outputs indicates that the fine-tuned Llama-eu-8B model effectively incorporates the KN sentences into its responses. In most cases, the model adds relevant text to directly address the HS, as demonstrated by examples such as *"Women should be free to choose their own job"* or *"Questo non è un buon motivo per dire alle donne che dovrebbero stare a casa e crescere i figli"* in the provided example.

The model's tendency to reproduce the KN sentences verbatim (or with minimal alterations) can be attributed to its training on the MT-CONAN-KN dataset. In this dataset, CS often included extended portions of the KN sentences, as evidenced by the high ROUGE scores observed during the annotation of the KN sentences (see Section 4.1). This strong alignment with the MT-CONAN-KN dataset further explains the relatively low JudgeLM scores. Indeed, the CS generated by our systems remain closely tied to the re-ranked KN sentences, limiting the stylistic and argumentative diversity of the output.

6 Conclusion

In this work, we addressed the challenges of multilingual, KN-driven CS generation, proposing an approach that integrates advanced *passage re-ranking* mechanisms into the generation pipeline. By leveraging multilingual cross-encoders and LLM-based re-rankers, we demonstrated the effectiveness of fine-grained KN selection in enhancing the quality and relevance of generated CS. Our results, evaluated on the MT-CONAN-KN dataset, show that reranker-based generation systems consistently outperform end-to-end approaches in both syntactic and semantic similarity metrics, underscoring the importance of re-ranking in this domain.

Despite these promising outcomes, our findings also reveal limitations in generating high-quality, unconstrained CS, particularly when evaluated using LLM-based metrics. These insights emphasize the need for further advancements, including the development of high-quality, domain-specific KN bases and more sophisticated retrieval and re-ranking strategies, and ad-hoc fine-grained metrics.

Overall, this study highlights the potential of KN-driven CS generation, particularly in multilingual contexts, as a critical tool in combating hate speech. Future work should focus on improving adaptability across languages and optimizing CS quality to better address the complex challenges posed by online hate speech.

Limitation

Despite the promising results of our approach, several limitations remain. The performance of multilingual re-rankers and models varied across languages, indicating challenges in achieving consistent cross-lingual adaptability. Moreover, in this work, we employed Llama-eus-8b, the only open-source LLM officially trained on all four target languages. However, as a base model, it lacks instruction-based fine-tuning, which we believe could significantly enhance counterspeech quality, particularly by leveraging conversational nuances. Additionally, the input data were automatically preprocessed, which may have introduced alignment issues or errors in pairing hate speech with KN sentences, eventually affecting the generated counterspeech quality. Manually curated annotations could help refine the training data and further improve performance. Finally, the KN sentences used for grounding the generation were often short and lacked sufficient contextual depth. Expanding the context available to both the re-ranker and the LLM could improve retrieval precision and lead to the generation of more coherent and impactful CS.

Ethical Statement

This study addresses the challenge of generating CS and constraining it on selected KN sentences. While the outcomes are encouraging, it's crucial to highlight that the success of these systems depends heavily on two factors: the quality of the input data and the capabilities of the LLM employed. A robust LLM may produce subpar CS if the ground KN is inaccurate or insufficient. On the other hand, weaker generative models may struggle to utilize the provided information effectively, leading to factual inaccuracies (Zellers et al., 2019; Solaiman et al., 2019) and ineffective CS, which hinders the goal of automating this task. Hence, in the context of KN-driven generation, particularly when addressing sensitive issues such as hate speech countering, it is crucial to maintain a standard quality of the resources employed. Nonetheless, it is important to note that automatic systems for CS generation are not deployed as autonomous systems. Instead, they should be considered as suggestion tools that serve as an aid for humans.

References

Imran Awan and Irene Zempi. 2016. The affinity between online and offline anti-muslim hate crime: dynamics and impacts. *Aggression and Violent Behavior*, 27:1–8.

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia. ELRA and ICCL.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7570–7577.

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, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. 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. 2021b. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Ander Corral, Ixak Sarasua, and Xabier Saralegi. 2024. Llama-eus-8b, a foundational sub-10 billion parameter llm for basque.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mekselina Doğanç and Ilia Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 1–12, Prague, Czechia. Association for Computational Linguistics.

Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science*, 13(1):4.

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, Online. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*.

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.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):369–380.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

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

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

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.

Nathaniel Persily, Joshua A Tucker, and Joshua Aaron Tucker. 2020. Social media and democracy: The state of the field, prospects for reform.

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, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 255–264, New York, NY, USA. Association for Computing Machinery.

Carla Schieb and Mike Preuss. Governing hate speech by means of counterspeech on facebook.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

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

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

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

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

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

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against neural fake news*. Curran Associates Inc., Red Hook, NY, USA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.

A KN sentences selection

Figure 3 illustrates the process of automatic sentence selection. The ROUGE-L score was used to evaluate the overlap between the CS and all KN sentences. This overlap was calculated for the entire CS (central column in the ROUGE scores matrix in the Figure) as well as for each of its sentences. Subsequently, the highest ROUGE-L value for each KN sentence was retained. Eventually, sentences whose ROUGE-L value was higher than a given threshold were labelled as those used for creating the CS (the green squares in the Figure).

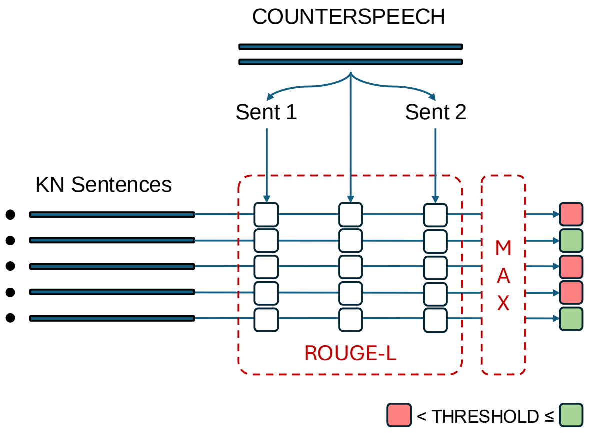

Figure 3: Graphical representation of the automatic procedure employed for selecting the KN sentences employed for writing the CS.

B Fine-Tuning Details

B.1 Re-Ranker Fine-Tuning

Starting from the annotated dataset, as detailed in Section A, we proceeded to fine-tune the M3_RRank and Gemma_RRank re-rankers. The following sections provide specifics for each re-ranker.

M3_RRank We fed the cross-encoder with the hate speech, the list of KN sentences used to create the CS, the list of the discarded KN sentences, and their ROUGE-L scores computed as explained in Section A. The information was formatted in JSON, and structured as follows.

```
{
    "query": hate speech,
    "pos": selected KN sentences,
    "neg": discarded KN sentences,
    "pos_scores": ROUGE-L scores selected KN
    ↪  sentences,
    "neg_scores": ROUGE-L scores discarded KN
    ↪  sentences
}
```

The re-ranker was trained on an NVIDIA Ampere A40 GPU with 48GB of memory for 5 epochs, using a learning rate of 6×10^{-5}, a training batch size of 8, and a weight decay of 0.01.

Gemma_RRank The fine-tuning of this LLM-based re-ranker utilized the same input as the M3_RRank, with the addition of a prompt instruction. The prompt used is detailed in Section 4.1 (paragraph *'Re-Ranker Fine-Tuning'*). The LLM underwent training on an NVIDIA Ampere A40 GPU with 48GB of memory, employing 'Low-Rank Adaptation' (LoRA; Hu et al., 2021) with a rank of 32 and an α value of 64. We trained the model for 5 epochs, with a learning rate of 5×10^{-5}, a weight decay set at 0.01, and a warm-up ratio of 0.1.

B.2 LLM Fine-Tuning for Generation

We utilized Llama-eus-8B for CS generation. Two versions of the LMM were fine-tuned, one corresponding to each CS generation approach, namely the Rerank-CS and E2e Prompt-CS approaches. The same hyperparameters were used across both fine-tuning, with the only variation being the training data. The training was performed on an NVIDIA Ampere A40 GPU with 48GB of memory, and no quantization was applied. Low-Rank Adaptation (LoRA) was utilized with a rank of 16, an α value of 16, and a dropout rate of 0. Training parameters included a learning rate of 5×10^{-5}, a training batch size of 2, an evaluation batch size of 4, and gradient accumulation steps of 4. The model was trained for 3 epochs, with a weight decay of 0.01, and a warm-up ratio of 0.03.

For the Rerank-CS approach we employed the prompt reported in Section 4.1 (paragraph *'Counterspeech Generation'*) filling it with the hate speech, the top 2 sentences selected by the retriever, and the gold CS from the train and dev sets of the MT-CONAN-KN. The dev set has been used as an evaluation set during training. For the E2E Prompt-CS both the hate speech and all the KN sentences were passed as input to the language model, formatted into a unique prompt, as shown below:

> You will be provided with a hateful comment (hate speech) and *{nof_sent}* sentences comprising arguments against the comment (knowledge).
>
> Select the most effective sentences and use them to generate a reply to the hateful content. Reply in the following language: language

Hate speech: *{hate speech}*

Knowledge: *{knowledge}*

Reply: *{counterspeech}*

C Complete Results

In Tables 4, 5 , 7 , 6 we report the general results of the shared tasks. Teams are reported in alphabetical order, and for each metric we highlighted in gold, silver, and bronze the first, second and third best results accordingly. We took part in the shared task under the name **TrenTeam**. The `Rerank-CS` systems utilizing `M3_RRank` and `Gemma_RRank` were submitted as *run1* and *run2* respectively; results for the `E2E Prompt-CS` system are designated with *run3*.

Team	JudgeLM Score	ROUGE-L (%)	BLEU (%)	BERTScore (%)	Novelty (%)	Gen_len
bhavanark run1	301.5	14.0	1.7	67.1	81.3	54.2
CODEOFCONDUCT run1	2374.5	16.2	2.8	69.4	83.4	84.8
CODEOFCONDUCT run2	2344.0	16.4	3.2	69.4	83.7	85.6
CODEOFCONDUCT run3	2394.5	16.2	2.9	69.1	83.4	88.3
counterspeech go run1	924.5	49.6	34.0	81.9	76.5	24.4
counterspeech go run2	854.0	49.7	34.0	81.8	77.2	24.0
counterspeech go run3	840.0	49.8	33.9	81.9	77.4	23.6
HuaweiTSC run1	1635.0	40.4	27.2	78.2	80.7	38.2
HuaweiTSC run2	2087.5	33.6	18.8	76.1	80.8	48.3
HuaweiTSC run3	1682.0	46.6	34.6	80.4	79.0	39.2
Hyderabadi Pearls run1	861.0	53.1	40.9	82.6	78.2	28.7
Hyderabadi Pearls run2	1058.5	44.3	34.8	79.5	77.0	32.1
Hyderabadi Pearls run3	996.5	45.2	35.2	79.5	77.0	30.9
MilaNLP run1	2326.5	18.1	3.2	70.7	82.3	64.5
MilaNLP run2	2357.5	18.5	3.8	70.8	82.5	66.7
MilaNLP run3	2523.0	19.0	4.9	70.8	83.0	84.7
NLP@IIMAS run1	704.0	48.8	41.2	80.8	78.2	29.8
NLP@IIMAS run2	2498.5	14.7	2.0	68.8	83.1	73.5
NLP@IIMAS run3	2494.5	14.7	2.0	68.8	83.1	73.5
Northeastern Uni run1	965.5	48.3	40.1	81.0	76.8	30.4
Northeastern Uni run2	990.0	51.6	42.1	82.3	76.6	30.9
Northeastern Uni run3	1191.0	51.8	40.3	82.6	78.1	43.0
RSSN run1	681.5	46.3	35.7	78.8	78.4	40.8
RSSN run2	59.0	24.5	13.2	69.2	80.8	31.0
SemanticCUETSync run1	1079.0	51.8	44.4	82.4	77.5	33.4
TrenTeam run1	1056.0	49.6	45.3	82.0	78.0	34.4
TrenTeam run2	1145.5	53.9	48.3	83.4	78.1	36.3
TrenTeam run3	999.5	52.5	43.3	82.2	79.0	35.4
ground truth	1175.5	100.0	100.0	100.0	77.7	32.7

Table 4: Results for English

Team	JudgeLM Score	ROUGE-L (%)	BLEU (%)	BERTScore (%)	Novelty (%)	Gen_len
bhavanark run1	73.0	11.0	2.1	62.6	84.6	39.8
CODEOFCONDUCT run1	1824.5	10.7	2.7	68.6	81.6	78.0
CODEOFCONDUCT run2	1740.5	10.2	2.2	68.5	82.5	80.2
CODEOFCONDUCT run3	1803.5	10.1	2.4	68.3	81.6	75.2
counterspeech go run1	667.5	47.0	32.2	80.9	77.5	28.1
counterspeech go run2	663.0	47.1	31.7	81.0	77.8	27.6
counterspeech go run3	685.0	46.5	32.3	81.1	77.7	27.7
HuaweiTSC run1	1260.5	36.1	21.7	77.2	80.9	40.8
HuaweiTSC run2	1792.0	30.8	16.6	75.9	80.3	49.5
HuaweiTSC run3	1372.5	41.1	26.6	79.1	79.1	41.9
MilaNLP run1	1824.0	16.8	3.7	70.8	82.0	62.1
MilaNLP run2	1912.0	22.7	9.1	73.0	81.1	73.4
MilaNLP run3	1985.5	21.1	8.9	72.6	82.1	101.4
NLP@IIMAS run1	529.5	36.7	27.6	77.2	78.3	32.4
NLP@IIMAS run2	1630.5	13.6	1.9	68.4	81.9	50.1
NLP@IIMAS run3	503.0	36.5	25.8	77.1	79.4	31.6
Northeastern Uni run1	830.0	42.6	30.8	79.7	77.8	32.0
Northeastern Uni run2	905.5	45.4	33.7	80.8	76.9	33.5
Northeastern Uni run3	1004.0	47.5	36.2	81.3	77.8	40.7
SemanticCUETSync run1	1028.0	46.7	36.2	81.1	78.3	34.9
TrenTeam run1	880.0	46.4	38.6	81.2	77.9	37.8
TrenTeam run2	965.5	48.6	41.2	81.7	77.8	37.0
TrenTeam run3	791.0	47.4	37.9	80.9	78.8	35.5
ground truth	929.5	100.0	100.0	100.0	77.9	35.3

Table 5: Results for Italian

Team	JudgeLM Score	ROUGE-L (%)	BLEU (%)	BERTScore (%)	Novelty (%)	Gen_len
bhavanark run1	54.0	14.7	2.5	64.7	81.0	42.7
CODEOFCONDUCT run1	1857.0	12.0	2.8	69.8	81.3	86.4
CODEOFCONDUCT run2	1820.5	12.0	2.8	69.8	81.5	87.2
CODEOFCONDUCT run3	1839.0	11.5	3.0	69.5	81.8	87.8
counterspeech go run1	639.0	47.6	29.9	81.1	75.3	27.1
counterspeech go run2	646.5	46.7	29.8	80.9	75.6	27.1
counterspeech go run3	652.5	47.4	29.7	80.9	75.7	26.5
HuaweiTSC run1	1228.5	36.8	21.7	77.6	77.5	43.1
HuaweiTSC run2	1728.0	33.5	17.7	76.7	77.4	52.3
HuaweiTSC run3	1339.5	41.9	27.2	79.4	75.8	43.2
MilaNLP run1	1852.5	19.6	4.8	71.5	79.2	67.7
MilaNLP run2	1942.0	23.7	8.6	73.5	78.0	72.7
MilaNLP run3	2002.0	24.2	8.9	73.5	79.6	99.3
NLP@IIMAS run1	492.5	39.7	30.7	78.2	77.3	36.3
NLP@IIMAS run2	1919.0	16.7	3.3	69.6	79.6	64.9
NLP@IIMAS run3	466.0	38.5	27.6	78.1	76.1	33.6
Northeastern Uni run1	894.5	45.6	34.5	80.6	74.0	35.1
Northeastern Uni run2	845.0	46.7	33.6	81.2	73.9	33.4
Northeastern Uni run3	873.0	45.3	33.4	80.5	76.6	43.8
SemanticCUETSync run1	974.5	46.5	35.6	80.8	75.3	36.5
TrenTeam run1	879.0	48.2	39.3	81.7	75.8	41.2
TrenTeam run2	987.5	51.6	42.9	82.8	75.6	40.9
TrenTeam run3	769.0	50.2	40.3	82.0	75.4	37.9
ground truth	899.0	100.0	100.0	100.0	75.1	36.9

Table 6: Results for Spanish

Team	JudgeLM Score	ROUGE-L (%)	BLEU (%)	BERTScore (%)	Novelty (%)	Gen_len
bhavanark run1	74.0	5,5	0,5	61,7	88,7	32,4
CODEOFCONDUCT run1	2465.5	8,2	1,5	66,4	86,8	67,5
CODEOFCONDUCT run2	2371.0	9,8	2,2	67,0	87,1	66,2
CODEOFCONDUCT run3	2382.5	10,4	2,2	67,5	87,5	69,1
counterspeech go run1	904.0	31,8	15,6	76,7	84,9	18,0
counterspeech go run2	837.0	32,4	15,8	77,1	85,1	18,0
counterspeech go run3	855.5	31,6	15,3	76,5	85,1	17,7
HuaweiTSC run1	1484.5	18,3	6,3	72,1	87,2	30,2
HuaweiTSC run2	1881.5	17,7	5,6	72,4	86,8	34,5
HuaweiTSC run3	1722.0	23,3	10,5	74,2	86,5	32,1
Hyderabadi Pearls run1	1011.5	29,2	17,4	75,5	85,6	26,2
Hyderabadi Pearls run2	1322.0	27,6	15,5	75,5	85,3	27,8
Hyderabadi Pearls run3	1023.5	29,2	17,4	75,5	85,6	26,2
MilaNLP run1	2242.5	10,7	1,0	69,0	87,8	44,6
MilaNLP run2	430.0	18,5	6,9	70,4	87,4	50,5
MilaNLP run3	422.5	17,9	6,8	70,7	88,3	72,8
NLP@IIMAS run1	720.5	29,2	17,6	74,9	86,0	24,9
NLP@IIMAS run2	2086.0	8,9	0,6	67,7	87,5	34,6
NLP@IIMAS run3	720.0	29,2	17,6	74,9	86,0	24,9
Northeastern Uni run1	1107.5	25,6	13,3	74,6	84,3	24,8
Northeastern Uni run2	1158.0	27,6	13,5	75,7	83,4	24,5
Northeastern Uni run3	1145.0	30,9	17,6	76,2	85,2	29,6
SemanticCUETSync run1	1194.0	26,5	15,4	75,1	85,4	26,0
TrenTeam run1	1364.5	33,8	22,4	77,6	85,2	28,2
TrenTeam run2	1394.5	32,8	20,9	77,1	85,7	27,5
TrenTeam run3	1246.0	31,7	18,2	76,6	85,9	24,0
ground truth	1534.5	100,0	100,0	100,0	85,3	26,5

Table 7: Results for Basque