The First Workshop on Multilingual Counterspeech Generation at COLING 2025: Overview of the Shared Task

Helena Bonaldi[*1,2] **M. Estrella Vallecillo-Rodríguez**[*3] **Irune Zubiaga**[*4] **Arturo Montejo-Ráez**[3]
Aitor Soroa[4] **María Teresa Martín-Valdivia**[3] **Marco Guerini**[1] **Rodrigo Agerri**[4]
[1]Fondazione Bruno Kessler, Italy, [2]University of Trento, Italy, [3]CEATIC, Universidad de Jaén, Spain,
[4]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{hbonaldi, guerini}@fbk.eu, {mevallec, amontejo, maite}@ujaen.es,
{irune.zubiaga, a.soroa, rodrigo.agerri}@ehu.eus,

Abstract

This paper presents an overview of the Shared Task organized in the First Workshop on Multilingual Counterspeech Generation at COLING 2025. While interest in automatic approaches to Counterspeech generation has been steadily growing, the large majority of the published experimental work has been carried out for English. This is due to the scarcity of both non-English manually curated training data and to the crushing predominance of English in the generative Large Language Models (LLMs) ecosystem. The task's goal is to promote and encourage research on Counterspeech Generation in a multilingual setting (Basque, English, Italian, and Spanish) potentially leveraging background knowledge provided in the proposed dataset. The task attracted 11 participants, 9 of whom presented a paper describing their systems. Together with the task, we introduce ML-MTCONAN-KN a new multilingual counterspeech dataset with 2384 triplets of hate speech, counterspeech, and related background knowledge covering 4 languages[1].

Content warning: this article contains unobfuscated examples that some readers may find offensive.

1 Introduction

Counterspeech (CS) is a promising strategy to fight online hate: it consists of replying to the hate speech (HS) with cogent agents, refuting it without being offensive. By challenging the stereotypes spread by the offensive message, it offers an alternative and constructive perspective and fosters empathy and understanding among users, promoting a more inclusive and respectful online environment (Benesch, 2014; Schieb and Preuss, 2016). Due to its potential effectiveness (Hangartner et al., 2021)

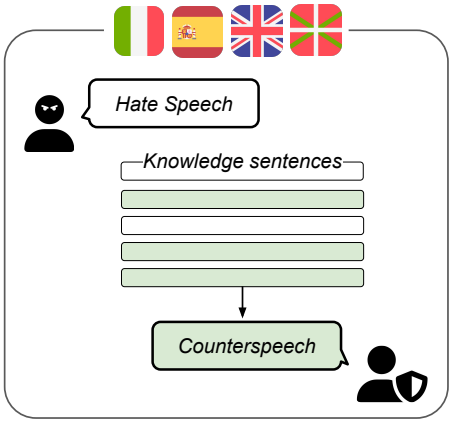

Figure 1: An example showing the structure of the ML-MTCONAN-KN dataset, i.e., triplets of hate speech, counterspeech and related background knowledge, in Italian, Spanish, English and Basque.

and given the sheer amount of HS being produced, Natural Language Processing is increasingly focusing on automating CS generation, in an effort to aid existing NGOs who manually produce these replies (Chung et al., 2021b; Bonaldi et al., 2024).

However, some aspects of automatic CS generation still remain largely understudied: this shared task addresses two of these existing gaps. First, although previous research on CS collection and generation has mostly focused on English (Qian et al., 2019; Tekiroglu et al., 2022; Halim et al., 2023; Mathew et al., 2018), there have been a few efforts to develop CS datasets for Italian (Chung et al., 2019, 2020), French (Chung et al., 2019), Spanish (Bengoetxea et al., 2024; Vallecillo-Rodríguez et al., 2024) and Basque (Bengoetxea et al., 2024), creating a body of curated data that represents a first step to facilitate research on the automatic generation of CS from a multilingual point of view.

Secondly, one of the main limitations of deploying automatic systems for CS production in the wild is the risk of generating inaccurate information. To address this problem, some studies have

[*]These authors contributed equally to this work.
[1]The dataset is available at: https://huggingface.co/datasets/LanD-FBK/ML_MTCONAN_KN

Proceedings of the First Workshop on Multilingual Counterspeech Generation (MCG), pages 92–107
January 19th, 2025.

proposed knowledge-driven systems for CS generation (Chung et al., 2021a; Jiang et al., 2023b). However, there has not been a systematic comparison of the different methods that can be applied for this task.

In this scenario, our shared task aims to promote research on the generation of CS in a multilingual setting, namely, in Basque, English, Italian and Spanish, with the possibility of leveraging the background knowledge provided in the dataset. To do so, we introduce `ML-MTCONAN-KN`, a new multilingual dataset with 2384 triplets of HS, CS, and related background knowledge (KN) covering 4 languages (the structure of the introduced dataset is shown in Figure 1). The task attracted 11 participants, with 9 submitted papers. Roughly half of the proposed approaches employed fine-tuning, with the other half mainly recurring to zero-shot learning and only one team opting for few-shot learning. Most fine-tuned models were fine-tuned on multilingual data, with only two teams focusing on monolingual generation. In addition, every participant except one used the background knowledge provided to support CS generation.

This task is evaluated using both traditional overlap and similarity metrics (i.e. BLEU, ROUGE, BERTscore) and LLM-as-a-judge evaluation (JudgeLM): the official winning system is selected with the latter method, as it has been shown to have a higher correlation with human evaluation (Zubiaga et al., 2024). In particular, among the proposed approaches, the MilaNLP systems, which involve a knowledge filtering step before zero-shot generation, consistently achieved the highest results across languages, thus representing the overall winner of our task.

2 Task description

Following a previous task organized for Spanish (Vallecillo-Rodríguez et al., 2024), our shared task focuses on multilingualism and on the possibility of employing knowledge-driven techniques for CS generation. With respect to multilingualism, the languages included offer a varied spectrum consisting of an agglutinative and isolate language (Basque), two romance languages (Italian and Spanish), and a Germanic one (English). The languages of interest were chosen according to the organizers' linguistic expertise, guaranteeing that native speakers are available for every language.

Regarding knowledge-driven generation, the new dataset includes triplets of HS, CS, and background KN sentences used to write the ground-truth CS. The participants have the option of using such knowledge, or any additional external knowledge of their choice, to obtain generations of higher quality.

HS: Jews are literally pieces of shit. There must be laws prohibiting their religion from our land.
KN: In contrast, laws that protect the autonomy of religious organizations [will] often further individual religious freedom as well. <EOS> The idea of individuals having freedoms is literally foreign to them. <EOS> So say on extremists of any religion - you're protected as am I. <EOS> The First Amendment protects freedom of speech, but it also protects freedom of religion. <EOS> As with freedom of speech and freedom of the press, freedom of religion sometimes means protecting conduct that some people might not favor. <EOS>
CS: The First Amendment protects freedom of speech, but it also protects freedom of religion. As with freedom of speech and freedom of the press, freedom of religion sometimes means protecting conduct that some people might not favor. However, to deny it would mean to deny a fundamental human right.

Table 1: An example in English from our dataset.

3 Dataset

We introduce `ML-MTCONAN-KN`, a new multilingual dataset with 2384 triplets of HS, CS, and related KN in 4 languages. We make the dataset available in three splits: 1584 examples for training, 400 for validation, and 400 for testing (the distribution is roughly 66% - 17% - 17%). An example of a triplet is shown in Table 1. The KN sentences include both knowledge considered relevant (highlighted in the example) and irrelevant by the annotator to write the gold CS[2].

The dataset covers hate speech targeted towards the following minority groups: Jews, LGBT+, immigrants, people of color (POC), and women. Table 2 shows the distribution of instances for each target group in the dataset. Other information included in the dataset corresponds to the language (`LANG`), the dataset split (`SPLIT`), an identifier for each HS - CS pair (`PAIR_ID`: different versions of the same pair in different languages have the same `PAIR_ID`), and a unique identifier for each pair in each language (`ID`), obtained by concatenating the `PAIR_ID` and `LANG` (e.g. "IT01").

[2]Note that the KN sentences are provided without distinguishing between those that are relevant or not to write the CS.

Target	N	%
Jews	400	17
LGBT+	408	17
Migrants	548	23
POC	352	15
Women	676	28
Total	2384	100

Table 2: The distribution of examples according to the target of hate.

3.1 Data collection

As it has been mentioned, `ML-MTCONAN-KN` contains triplets of HS, CS, and related KN. In the following, we describe how each of these elements was obtained.

HS We used the hate speech from the MT-CONAN dataset (Fanton et al., 2021), which comprises 5003 HS-CS pairs in English covering multiple targets of hate.

KN For each HS instance in the dataset, the knowledge retrieval and selection module from Chung et al. (2021a) was applied to retrieve relevant knowledge. More specifically, first a query for each HS is created, which is then used to search for relevant knowledge in an external repository (Newsroom from Grusky et al. (2018) and WikiText-103 from Merity et al. (2022)) via the Solr search engine[3]. The similarity between each article and the query is calculated with BM25 (Robertson et al., 1995), and the top 25 articles are retained. These documents are split into sentences, and the 5 most relevant according to ROUGE-L F1 (Lin, 2004) are kept as background knowledge for the dataset.

CS A graduate-level annotator with previous experience in CS writing is tasked with producing the CS starting from the provided KN. In particular, the annotation guidelines are the following: (i) if no relevant knowledge is available, discard the example, (ii) if relevant knowledge is available, use it to write a counterspeech message, adapting it if required and (iii), if several KN sentences are deemed relevant, the entry can be duplicated and two different CS for the same HS can be written, using different parts of the KN. On average, in one hour of annotation time, 17.5 examples can be obtained. At the end of this process, 596 CS examples in English were manually generated.

[3] https://lucene.apache.org/solr/

HS: Gli ebrei sono letteralmente dei pezzi di merda. Devono esserci delle leggi che proibiscano la loro religione nella nostra terra. **CS:** Il Primo Emendamento protegge la libertà di parola, ma anche la libertà di religione. Come per la libertà di parola e di stampa, la libertà di religione a volte implica la protezione di comportamenti che a qualcuno potrebbero non piacere. Tuttavia, negarla significherebbe negare un diritto umano fondamentale.
HS: Los judíos son literalmente pedazos de mierda. Debe haber leyes que prohíban su religión en nuestra tierra. **CS:** La Primera Enmienda protege la libertad de expresión, pero también la libertad religiosa. Al igual que ocurre con la libertad de expresión y la libertad de prensa, la libertad de religión a veces significa proteger conductas que algunas personas podrían no aprobar. Sin embargo, negarla significaría negar un derecho humano fundamental.
HS: Juduak kaka hutsak dira. Haien erlijioa gure lurraldean debekatzen duten legeak egon behar dira. **CS:** Lehen Zuzenketak adierazpen askatasuna babesten du, baina baita erlijio askatasuna ere. Adierazpen askatasunarekin eta prentsa askatasunarekin gertatzen den bezala, erlijio askatasunak, batzuetan, pertsona batzuek mesedegarri ez dituzten jokabideak babestea esan nahi du. Baina ukatzeak oinarrizko giza eskubide bat ukatzea esan nahiko luke.

Table 3: Translation of an HS-CS pair into IT, ES and EU.

3.2 Translation to other languages

Translating the English data into the other languages consisted of a two-step procedure. First, automatic translation was used: DeepL[4] for Spanish and Italian, and Itzuli[5] for Basque. Second, the automatic translations of the HS and CS were manually reviewed and post-edited by expert human annotators. Table 3 shows the HS and CS in Table 1 translated into Italian, Spanish, and Basque.

4 Evaluation

The evaluation of the shared task is based on two complementary approaches. First, on traditional overlap and similarity metrics commonly used in machine translation and text generation, including those specifically tailored for CS generation (Bengoetxea et al., 2024). Second, we use a recently proposed method based on JudgeLM which has a stronger correlation with human evaluations than traditionally used metrics (Zubiaga et al., 2024). The official ranking and task winner are determined by the pairwise ranking-based evaluation using JudgeLM.[6]

[4] https://www.deepl.com

[5] https://www.euskadi.eus/itzuli/

[6] The evaluation code is available at https://github.com/hitz-zentroa/eval-MCG-COLING-2025. It was also

4.1 Traditional Metrics

Reference-based metrics measure the overlap or embedding similarity between the generated and the reference CS. Furthermore, we also apply reference-free metrics to evaluate the generated CS without considering any ground-truth CS.

Reference-based metrics Building on prior work in CS generation (Tekiroglu et al., 2022; Bengoetxea et al., 2024), we chose to evaluate the submitted runs using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020). BLEU (widely used in machine translation tasks) is a precision-focused metric that assesses the overlap between a candidate text and one or more reference texts. More specifically, it calculates the geometric mean of modified n-gram precision while applying a brevity penalty to discourage overly short outputs. In contrast, ROUGE-L emphasizes recall by identifying the longest common subsequence between the candidate and reference, normalized by the reference length. ROUGE-L is frequently applied in text summarization. Finally, BERTScore uses contextual embeddings from pre-trained BERT models to measure the similarity between candidate and reference sentences.

Reference-Free Metrics We consider two different metrics: Novelty and Repetition Rate. Novelty (Wang and Wan, 2018) is calculated by identifying non-singleton n-grams in the generated text that also appear in the training data. Novelty aims to measure how distinct the generated content is from the training data. It should be noted that this metric is less informative when evaluating models in zero-shot settings, where no training data is involved. Regarding Repetition Rate (RR) (Bertoldi et al., 2013), the idea is to identify the non-singleton n-grams that are repeated within the generated text, providing a measure of self-similarity in the content. This metric focuses on capturing the diversity of the generated text.

4.2 JudgeLM Pairwise Rank Evaluation

The official scorer for the task is based on a new method to evaluate CS using JudgeLM which consists of a pairwise rank-based approach, originally proposed in Zubiaga et al. (2024). Given a pair of candidate CS, an LLM acts as a judge to determine the superior counterspeech. It has been shown that this method exhibits a high correlation with human judgments for this specific task. The model chosen for English, Spanish, and Italian is JudgeLM (Zhu et al., 2023), a scalable judge model built upon Vicuna. JudgeLM is trained using a large dataset of LLM-generated responses including various natural language generation (NLG) tasks, paired with detailed evaluations generated with GPT-4. Although JudgeLM supports various evaluation approaches, such as comparing single answers to a given reference or multiple answers simultaneously, we opted for a pairwise comparison of generated CS as proposed by Zubiaga et al. (2024). This approach eliminates the need for a ground-truth reference, focusing instead on selecting the best option among the available alternatives. By directly comparing two CS candidates, we also avoid the ambiguity inherent in evaluating them individually within an open-ended framework. Furthermore, and unlike traditional metrics, this method evaluates CS within the context of specific HS instances, rather than treating them as independent generations. Finally, in order to address the lack of Basque support of the original JudgeLM,we fine-tuned `Llama-eus-8B` (Corral et al., 2024) on the JudgeLM-100K dataset presented in Zhu et al. (2023), using the same settings outlined in the paper.

provided to the participants to assist them during system development.

Team	FT	Mul. FT	Kn. Fil.	Lan.
RSSN	✓	-	-	EN
Hyderabadi Pearls	✓	-	-	EN, EU
Counterspeech go	✓	✓	✓	All
Trenteam (run 1)	✓	✓	✓	All
Northeastern	✓	✓	-	All
NLP@IIMAS (run 1)	✓	✓	-	All
bhavanark	✓	✓	-	All
Trenteam (run 2-3)	-	-	✓	All
MilaNLP	-	-	✓	All
NLP@IIMAS (run 2-3)	-	-	-	All
CODEOFCONDUCT	-	-	-	All
HuaweiTSC	-	-	-	All

Table 4: Overview of the proposed systems, according to four distinguishing dimensions, namely, whether the model is fine-tuned (FT), fine-tuned over multilingual data (Mul. FT), whether the knowledge was filtered before being used (Kn. Fil.) and the language(s) of interest (Lan.).

5 Systems Overview

In Table 4 we report the participant teams and their approaches along four distinguishing dimensions, namely, whether (a) the system is fine-tuned, (b)

fine-tuned over multilingual data, (c) background knowledge was filtered before being used and (d) the languages the system focuses on. Another issue worth mentioning is that every participant except the *bhavanark* team used the available knowledge. Moreover, only the *Counterspeech go* team employed additional data for training, while all the other teams only used the shared task dataset. Table 4 groups the systems according to common characteristics: from top to bottom, we can observe how the approaches focusing on a subset of the available languages all opted for fine-tuning, without filtering the knowledge. Furthermore, two groups of approaches performed multilingual fine-tuning, with or without knowledge filtering. Finally, those systems which did not perform any fine-tuning, with and without knowledge filtering, are listed. A summary of the submitted systems follows.

RSSN This team employs a language model fine-tuned to generate counterspeech in English, given as input a prompt including the hate target, the HS, and all the provided knowledge sentences. T5 (Raffel et al., 2020) is employed in the first run, while DistilBART (Lewis et al., 2020) in the second.

Hyderabadi Pearls They fine-tune Mistral 7B (Jiang et al., 2023a), Llama-base 3.1 8B (Dubey et al., 2024), and Llama-eus 8B (Corral et al., 2024) on the ML-MTCONAN-KN corpus to generate CS in English and Basque. Additionally, they experiment with GPT-4 to post-edit the generated CS for the aforementioned languages. The submitted systems are language-dependent. For Basque, run 1 and run 3 use LLaMa-base 3.1 models fine-tuned for 3,000 steps, while run 2 uses Llama-eus for 500 steps. For English, run 1 and run 2 are based on LLaMa-base 3.1 models, adjusted for 300 steps in the first run and 3,000 steps in the second one. Finally, run 3 uses the Mistral model with 300 steps.

Counterspeech go The same configuration was used for the three runs, i.e. a QWEN2.5-14B-Instruct model (Hui et al., 2024) fine-tuned on the provided dataset. Additionally, the knowledge sentences were filtered using GPT-4o (Hurst et al., 2024), Claude[7], and Gemini (Team et al., 2023). Moreover, the Alpaca dataset[8] was used as additional data for training to prevent overfitting. The runs focused on all four languages.

Trenteam The team proposes two main approaches: the Rerank-CS approach, where a multilingual reranker (bge-reranker-v2-m3[9] for run 1 and bge-reranker-v2-gemma[10] for run 2) is fine-tuned to identify the most relevant sentences in the KN, which are then passed to an LLM to guide the CS generation in a zero-shot-learning fashion. In the second approach (run 3) a multilingual LLM is fine-tuned over the entire set of KN sentences and prompted to identify the most relevant and use them to produce the CS in an end-to-end process (E2E Prompt-CS approach). For all three approaches, the employed LLM is Llama-eus-8B (Corral et al., 2024).

Northeastern Uni The Northeastern University team leverages Llama-3 in two main approaches for training: supervised fine-tuning (on the base model for run 1 and the instruct model for run 2) and the Direct Preference Optimization (DPO) strategy (run3). In all runs, they leverage the ML-MTCONAN-KN dataset in all four languages and use the provided background knowledge sentences. For the DPO strategy, they additionally incorporate negative examples of counterspeech generated with GPT-4o.

NLP@IIMAS Two systems are proposed to address the task depending on the language. The first system used for run 1 employs a graph-based generative model (Flan-T5, Chung et al., 2024) that encodes knowledge about HS to generate the CS. For Run 2, the system featured a LLM with personalized counterspeech prompts, applying Chain-of-Thought for Italian and zero-shot for rest of the languages. Finally, run 3 consists of using a LLM in zero-shot for English, while a graph-based approach was applied to the remaining languages. Both systems integrate background knowledge from the dataset: in the LLM-based system, relevant phrases are included in the prompt, while in the graph-based system, they are organized sequentially or interspersed with the offensive message.

Bhavanark The presented system consists of a GPT-2 model fine-tuned on the ML-MTCONAN-KN HS and CS only, in the English language.

[7] https://claude.ai/

[8] https://huggingface.co/datasets/tatsu-lab/alpaca

[9] https://huggingface.co/BAAI/bge-reranker-v2-m3

[10] https://huggingface.co/BAAI/bge-reranker-v2-gemma

English		Basque		Italian		Spanish	
MilaNLP	2523.0	CODEOFCONDUCT	2465.5	MilaNLP	1985.5	MilaNLP	2002.0
NLP@IIMAS	2498.5	MilaNLP	2242.5	CODEOFCONDUCT	1824.5	NLP@IIMAS	1919.0
CODEOFCONDUCT	2394.5	NLP@IIMAS	2086.0	HuaweiTSC	1792.0	CODEOFCONDUCT	1857.0
HuaweiTSC	2087.5	HuaweiTSC	1881.5	NLP@IIMAS	1630.5	HuaweiTSC	1728.0
Northeastern Uni	1191.0	ground truth	1534.5	SemanticCUETSync	1028.0	TrenTeam	987.5
ground truth	1175.5	TrenTeam	1394.5	Northeastern Uni	1004.0	SemanticCUETSync	974.5
TrenTeam	1145.5	Hyderabadi Pearls	1322.0	TrenTeam	965.5	ground truth	899.0
SemanticCUETSync	1079.0	SemanticCUETSync	1194.0	ground truth	929.5	Northeastern Uni	894.5
Hyderabadi Pearls	1058.5	Northeastern Uni	1158.0	Counterspeech go	685.0	Counterspeech go	652.5
Counterspeech go	924.5	Counterspeech go	904.0	Counterspeech go	667.5	bhavanark	54.0
RSSN	681.5	NLP@IIMAS	720.5	bhavanark	73.0		
bhavanark	301.5	bhavanark	74.0				

Table 5: Official results using Pairwise rank-based method with JudgeLM. The ranking is based on each team's best submission.

MilaNLP They adopt two distinct approaches to address the generation problem in English, Italian, Spanish, and Basque. For run 1, they use the Mistral-7B-Instruct-v0.3 model in a zero-shot approach to generate CS in English and then translate them into the target languages using the NLLB model (Costa-jussà et al., 2022). For runs 2 and 3, they directly generate the CS in the target languages. For all submitted runs, they implement a knowledge filtering step, either filtering the relevant sentences in a separate prompt before generation (runs 1-2) or asking the model to choose which sentences to use at inference time (run 3).

CODEOFCONDUCT First, a simulated annealing algorithm is used to generate the candidate CS, which is iteratively refined. k candidates are selected according to a Boltzmann-like distribution which accounts for the JudgeLM score of each candidate. Then, new candidates are generated starting from the selected ones and are evaluated with the same methodology. Finally, the best candidates are selected using a round-robin algorithm. Runs 1, 2 and 3 submitted by the team correspond to the candidates ranked first, second and fourth, respectively.

HuaweiTSC Three systems are proposed for the four languages: all employ few-shot learning with Chain-of-Thought to prompt GPT-4o-mini to generate CS candidates (run 1). Moreover, they also test two approaches to select the best CS candidate: a pair-wise comparison which selects the best candidate according to the highest Elo rating obtained using JudgeLM (run 2), and a point-wise scorer which integrates multiple metrics to evaluate each candidate individually, given a hate speech and its corresponding background knowledge (run 3).

SemanticCUETSync The developed system focuses on all languages and leverages the knowledge provided in the task dataset, as well as additional general information from external sources. However, details on how this external information was integrated or specifics of the system implementation were not provided. For this reason, the results of this team are not included in the following analyses.

6 Official Results

Table 5 reports the official ranking determined by the Pairwise rank-based JudgeLM method. Tables showcasing the rank per submitted run and all evaluation metrics considered for the task are provided in Appendix A.

The results across the four languages—English, Basque, Italian, and Spanish—reveal interesting trends and highlight the strong performances of certain teams. Thus, MilaNLP stands out as a consistent top performer, ranked first in English, Italian, and Spanish, and second in Basque, showcasing their adaptability across languages and thus representing the overall winner of the task. CODEOFCONDUCT also achieved impressive results, ranking first in Basque, second in Italian, third in English, and fourth in Spanish. NLP@IIMAS also obtained competitive results, ranked second in English, third in Basque, and fourth in both Italian and Spanish. Similarly, HuaweiTSC performed well, with strong rankings such as fourth in English and third in Italian. The ground truth scores, prominently included in each table, provide a benchmark for assessing the submissions, with several teams surpassing this baseline, reflecting the quality of their outputs. If we consider the systems proposed by these teams, we can observe some recurring patterns: in particular, they all use zero-shot learn-

English		Basque		Italian		Spanish	
TrenTeam	0.834	TrenTeam	0.776	TrenTeam	0.817	TrenTeam	0.828
Northeastern Uni	0.826	Counterspeech go	0.771	Northeastern Uni	0.813	Northeastern Uni	0.812
Hyderabadi Pearls	0.826	Northeastern Uni	0.762	Counterspeech go	0.811	Counterspeech go	0.811
SemanticCUETSync	0.824	Hyderabadi Pearls	0.755	SemanticCUETSync	0.81	SemanticCUETSync	0.808
Counterspeech go	0.819	SemanticCUETSync	0.751	HuaweiTSC	0.791	HuaweiTSC	0.794
NLP@IIMAS	0.808	NLP@IIMAS	0.749	NLP@IIMAS	0.772	NLP@IIMAS	0.782
HuaweiTSC	0.804	HuaweiTSC	0.742	MilaNLP	0.73	MilaNLP	0.735
RSSN	0.788	MilaNLP	0.707	CODEOFCONDUCT	0.686	CODEOFCONDUCT	0.698
MilaNLP	0.708	CODEOFCONDUCT	0.675	bhavanark	0.626	bhavanark	0.647
CODEOFCONDUCT	0.694	bhavanark	0.617				
bhavanark	0.671						

Table 6: Results with BERTscore. The ranking is based on each team's best submission.

ing, apart from HuaweiTSC which performs few-shot, and they all rely on the provided background knowledge, with MilaNLP's systems additionally filtering the knowledge sentences for generation.

By considering the results obtained using the traditional metrics (see the rankings for BERTscore in Table 6), it can be observed that Trenteam consistently obtains first place across all languages, followed by Counterspeech go, Northeastern University and Hyderabadi Pearls. When analyzing these systems, a common characteristic seems to be that the models were taught to select the relevant knowledge for generating counterspeech. This was done either explicitly via knowledge filtering (Trenteam run 2-3 and Counterspeech go run 1-2) or implicitly via fine-tuning (Northeastern run 3 and Hyderabadi Pearls run1). In fact, selecting specific knowledge sentences for generation allows to mimic the process in which the gold CS were created manually, thus reaching higher similarity with the references.

7 Discussion

In this section we discuss the results obtained by the proposed approaches from an aggregated point of view, first averaging across all languages, and then comparing their performance on English vs low-resourced languages. Results are reported in Table 7.

Overview of all languages Fine-tuned models achieve significantly higher scores on traditional metrics, and they also have shorter generations (in line with the length of the training data). Moreover, if we distinguish the non-fine-tuned systems between those using zero-shot and few-shot learning (only the HuaweiTSC team) it is possible to see how few-shot learning achieves an average generation length closer to that of fine-tuned models (41.1), in contrast to the average length from zero-shot generations (63.7). Fine-tuned models have lower Novelty, which is expected, as the generations are more similar to the training data. The best runs according to JudgeLM are those based on zero-shot, which turn out to be also the longest generations (Hu et al., 2024).

Moreover, fine-tuning over multilingual data benefits the performance on the overlap metrics, but it lowers the performance according to Novelty, RR and JudgeLM: the same trend can be observed for systems performing a knowledge filtering step before generation.

English vs low-resourced languages If we focus on the overlap metrics, the generation length and Novelty, similar trends can be seen across languages: fine-tuning is helpful in obtaining higher overlap scores, shorter generations (closer to those of the ground-truth references) and lower Novelty: all these trends are expected, as discussed previously.

Appendix B provides a preliminary manual qualitative analysis of the generated CS. The analysis indicates that the winning runs according to BERTScore focus on selecting knowledge from the provided options and tend to reproduce it more or less verbatim, achieving thus a high overlap with the reference. In contrast, the best runs according to JudgeLM, while also leveraging the provided knowledge, tend to rephrase it. This approach reduces the overlap with respect to the ground-truth CS but results in more natural generations.

The main differences across languages are registered for RR and JudgeLM. In particular, two different phenomena can be observed. First, fine-tuned models in English score worse according to RR and JudgeLM. However, when fine-tuned over multilingual data, repetitiveness is lower and JudgeLM assigns higher scores. Second, for low-

Lang	Approach		ROUGE-L	BLEU	BERTsc.	length	Novelty	RR	JudgeLM
All	Gold		1	1	1	32.835	0.790	3.773	1134.625
	FT	✓	**0.382**	**0.262**	**0.776**	31.720	0.798	3.992	785.3
		-	0.252	0.135	0.734	**58.151**	**0.819**	**3.882**	**1738.3**
	Mul. FT	✓	**0.393**	**0.270**	**0.780**	**31.269**	0.794	4.097	783.5
		-	0.374	0.263	0.770	30.466	**0.810**	**3.376**	**876.7**
	Kn. Fil.	✓	**0.362**	**0.234**	**0.775**	43.509	0.803	4.068	1207.4
		-	0.287	0.175	0.743	**46.232**	**0.813**	**3.857**	**1310.3**
EN	Gold		1	1	1	32.65	0.776	3.777	1175.5
	FT	✓	**0.447**	**0.337**	**0.793**	32.724	0.780	4.233	820.2
		-	0.277	0.151	0.737	**62.996**	**0.817**	**3.724**	**2066.3**
	Mul. FT	✓	**0.499**	**0.389**	**0.818**	30.070	0.773	**3.969**	**940.6**
		-	0.379	0.269	0.761	**36.262**	**0.788**	4.586	659.7
	Kn. Fil.	✓	**0.401**	**0.279**	**0.784**	43.793	0.793	**3.852**	**1447.4**
		-	0.348	0.233	0.757	**49.052**	**0.800**	4.056	1406.6
ES, EU, IT	Gold		1	1	1	32.897	0.794	3.771	1121.0
	FT	✓	**0.354**	**0.230**	**0.768**	31.294	0.806	**3.890**	770.5
		-	0.243	0.129	0.733	**56.401**	**0.820**	3.939	**1619.9**
	Mul. FT	✓	**0.371**	**0.245**	**0.773**	30.777	0.800	4.132	754.8
		-	0.245	0.131	0.734	**55.176**	**0.820**	**3.973**	**1538.6**
	Kn. Fil.	✓	**0.350**	**0.220**	**0.771**	43.414	0.807	4.140	1127.4
		-	0.261	0.150	0.736	**45.023**	**0.818**	**3.771**	**1269.1**

Table 7: From top to bottom: aggregated results for all languages, English and low-resourced languages respectively.

resourced languages, fine-tuned models are less repetitive, and fine-tuning over multilingual data actually worsens the RR scores and the performance according to JudgeLM.

Finally, filtering the background knowledge helps to improve RR and JudgeLM for English, but it degrades performance for the other languages. Therefore, we can conclude that overall, both fine-tuning over multilingual data and filtering the knowledge seem to benefit more in English than in the rest of the languages. Furthermore, for all languages, fine-tuning allows to obtain generations more similar to the gold references but with worse performance according to the pairwise ranking method used with JudgeLM.

8 Conclusion

The analysis of the results of the shared task highlights some patterns among the most successful approaches. In particular, zero-shot learning combined with the provided background knowledge allows to obtain better multilingual counterspeech generation in terms of overall quality, measured by JudgeLM. On the other hand, systems obtaining the best scores on the traditional overlap-based metrics demonstrate that teaching the systems to select relevant knowledge, either by explicitly filtering it or implicitly via fine-tuning, effectively replicates the manual creation process of counterspeech. Moreover, the differences between high-resource and low-resource languages suggest the need to apply different strategies across linguistic contexts.

In summary, the obtained results in the shared task not only advance the state of the art in automatic counterspeech generation but also highlight critical areas for future research, such as developing more robust methods for low-resource languages and the need for deeper exploration into the evaluation of these systems.

Limitations

This work provides an in-depth analysis of the systems developed to address the task but still has certain limitations. First, the dataset does not include information about which external knowledge sentences are relevant for developing the gold CS. This information could help future systems discriminate between what is relevant and what is not in CS generation.

Second, automatic evaluation remains a major challenge in language generation, especially in this task. As shown, traditional metrics based on n-gram or embedding similarity do not evaluate the quality of the counterspeech with respect to a

given hate speech. Furthermore, previous work has shown a lower correlation of these metrics concerning human judgments. Therefore, we propose a new method based on JudgeLM as an alternative. However, despite its good correlation with human judgments, JudgeLM may introduce biases inherent in the models used as judges or it may show preferences for certain types of counter-narratives.

By highlighting these limitations we hope to encourage future research on multilingual counterspeech generation and evaluation.

Ethics Statement

Generating multilingual counterspeech to combat hate speech involves significant ethical and social considerations. Researchers and developers must take care to avoid reproducing harmful content, ensuring a responsible approach in creating automatic counterspeech systems.

First, the emotional well-being of researchers and annotators must be prioritized, as constant exposure to hateful content can harm mental health. Strategies like regular breaks and access to emotional support are essential when labeling datasets or evaluating systems that handle hate speech.

The dataset `ML-MTCONAN-KN` includes offensive messages, but it is designed to prevent models from generating abusive content. For this reason, our main focus in creating it was centered on achieving high-quality counterspeech replies, while the hateful messages are simple and stereotyped, to avoid possible misuses. Moreover, these messages were originally generated automatically, which allows us to preserve users' privacy.

Finally, automated systems may generate biased or harmful responses, especially when cultural and linguistic nuances are poorly addressed. For this reason, despite progress in automation, human involvement remains crucial: in this task, we always envision the deployment of counterspeech generation systems as assistant tools rather than to be deployed in the wild with no supervision.

Acknowledgements

This work has been partially supported by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS), and by the following MCIN/AEI/10.13039/501100011033 projects: CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTox (PDC2022-133146-C21), DISARGUE (TED2021-130810B-C21), DEEPR3 (TED2021-130295B-C31) and European Union NextGenerationEU/PRTR, DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR, DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU.

References

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

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.

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

Helena Bonaldi, Greta Damo, Nicolás Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. 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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it*.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. 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.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering NGOs in countering online hate messages. *Online Social Networks and Media*, 24:100150.

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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. IEEE.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Jingang Wang, Zhenyu Chen, and Hui Xiong. 2024. Rethinking llm-based preference evaluation. *arXiv preprint arXiv:2407.01085*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023a. Mistral 7b (2023). *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. *arXiv preprint arXiv:2310.05650*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2022. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Serra 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 *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.

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

María Estrella Vallecillo-Rodríguez, María Victoria Cantero-Romero, Isabel Cabrera-de Castro, Luis Alfonso Ureña-López, Arturo Montejo-Ráez, and María Teresa Martín-Valdivia. 2024. Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish. *Procesamiento del Lenguaje Natural*, 73(0):449–459.

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.

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

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

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 Official results

The official results of the shared task are presented. JudgeLM Score refers to the score obtained in the pairwise comparison setting described in Subsection 4.2. Generation Length corresponds to the average length of the generated outputs. Traditional metrics are those detailed in Subsection 4.1. The ranking was determined based on the JudgeLM Score for each of the languages.

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

Table 8: English Results.

Rank	Team Runs	JudgeLM Score	Traditional metrics(%)				generation length
			ROUGE-L	BLEU	BERTscore	Novelty	
1	CODEOFCONDUCT run1	2465.5	8.2	1.5	66.4	86.8	67.5
2	CODEOFCONDUCT run3	2382.5	10.4	2.2	67.5	87.5	69.1
3	CODEOFCONDUCT run2	2371.0	9.8	2.2	67.0	87.1	66.2
4	MilaNLP run1	2242.5	10.7	1.0	69.0	87.8	44.6
5	NLP@IIMAS run2	2086.0	8.9	0.6	67.7	87.5	34.6
6	HuaweiTSC run2	1881.5	17.7	5.6	72.4	86.8	34.5
7	HuaweiTSC run3	1722.0	23.3	10.5	74.2	86.5	32.1
8	ground truth	1534.5	100.0	100.0	100.0	85.3	26.5
9	HuaweiTSC run1	1484.5	18.3	6.3	72.1	87.2	30.2
10	TrenTeam run2	1394.5	32.8	20.9	77.1	85.7	27.5
11	TrenTeam run1	1364.5	33.8	22.4	77.6	85.2	28.2
12	Hyderabadi Pearls run2	1322.0	27.6	15.5	75.5	85.3	27.8
13	TrenTeam run3	1246.0	31.7	18.2	76.6	85.9	24.0
14	SemanticCUETSync run1	1194.0	26.5	15.4	75.1	85.4	26.0
15	Northeastern Uni run2	1158.0	27.6	13.5	75.7	83.4	24.5
16	Northeastern Uni run3	1145.0	30.9	17.6	76.2	85.2	29.6
17	Northeastern Uni run1	1107.5	25.6	13.3	74.6	84.3	24.8
18	Hyderabadi Pearls run3	1023.5	29.2	17.4	75.5	85.6	26.2
19	Hyderabadi Pearls run1	1011.5	29.2	17.4	75.5	85.6	26.2
20	Counterspeech go run1	904.0	31.8	15.6	76.7	84.9	18.0
21	Counterspeech go run3	855.5	31.6	15.3	76.5	85.1	17.7
22	Counterspeech go run2	837.0	32.4	15.8	77.1	85.1	18.0
23	NLP@IIMAS run1	720.5	29.2	17.6	74.9	86.0	24.9
24	NLP@IIMAS run3	720.0	29.2	17.6	74.9	86.0	24.9
25	MilaNLP run2	430.0	18.5	6.9	70.4	87.4	50.5
26	MilaNLP run3	422.5	17.9	6.8	70.7	88.3	72.8
27	bhavanark run1	74.0	5.5	0.5	61.7	88.7	32.4

Table 9: Basque Results.

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

Table 10: Italian Results.

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

Table 11: Spanish Results.

B Qualitative analysis of the results

We conducted a preliminary review of the winning runs by comparing the outputs from the winning systems according to the JudgeLM pairwise rank-based evaluation and BERTScore. Based on an initial assessment of a subset of 100 instances, we observed some trends in the characteristics of the outputs selected as winners by these two methods. Table 12 provides an illustrative example in each language where, according to the JudgeLM pairwise rank-based evaluation, the winner is Win_{PRB}, and according to BERTScore, the winner is Win_{BERT}.

When focusing on the English results, runs identified as winners by the JudgeLM pairwise rank-based evaluation tended to appear more natural overall while, those selected as winners based on BERTScore, often seemed less natural. Thus, the winner according to BERTScore closely mirrors the ground truth, using the same knowledge that informed the reference CN word for word, likely contributing to its high BERTScore. However, it also introduces an ungrammatical piece of knowledge (Knowledge 2), making the response seem incongruent. Throughout the submitted run, the CNs were combinations of pieces of knowledge without rephrasing, resulting in a high overlap with the reference but an unnatural, disjointed appearance. This made them perform poorly when evaluated by the JudgeLM pairwise rank-based evaluation method.

In contrast, the winner according to the JudgeLM rank-based evaluation also draws on the knowledge used in the reference CN, as well as Knowledge 3, but it combines it and rephrases it in a more natural manner. However, by rephrasing it, the overlap with the reference is reduced, which results in a lower BERTScore. This observation extends to other languages, where the winners according to BERTScore are those that select knowledge and reproduce it almost word by word, while the winners according to JudgeLM seem to be more focused on generation and adapting the knowledge to make it sound more natural.

English	
HS	Jews control Hollywood and the media.
Win_{PRB}	In Hollywood, many individuals of Jewish heritage have made significant contributions, yet it's inaccurate to suggest that they collectively "control" the industry. Historically, Hollywood Jews have been committed to assimilation and not religious celebration. They've been intermarried, secular, and ambivalent about their faith, often keeping Judaism a private matter. This misconception stems from a few prominent figures in the industry, rather than a coordinated effort or control.
Win_{BERT}	But almost from the inception of the American film industry, the Hollywood Jews were dedicated to assimilation, not religious celebration. Well, Jews last, which is why, in Hollywood, Judaism was always sotto voce.
ground truth	Almost from the inception of the American film industry, the Hollywood Jews were dedicated to assimilation, not religious celebration. Why should their religion be connected to an alleged "control of the media"?
KN	1. He began writing other Hollywood notables, asking them to sponsor endangered Jews. 2. well, Jews last, which is why, in Hollywood, Judaism was always sotto voce. 3. Secular, skeptical, intermarried, ambivalent, disenchanted, disengaged, sometimes Jews: on this day, Judaism and Jewish community briefly take center stage. **4. But almost from the inception of the American film industry, the Hollywood Jews were dedicated to assimilation, not religious celebration.** 5. And he urges that God himself would punish any king who tried to interfere with the Jews' practice of Judaism.
Spanish	
HS	Los judíos controlan el gobierno de EE.UU
Win_{PRB}	A pesar de que en el pasado hubo momentos de conflictos entre judíos y musulmanes, como en el caso de Israel, es importante recordar que hoy en día, la mayoría de los israelíes y musulmanes coexisten en paz. En cuanto al gobierno de los EE. UU., no hay evidencia de que los judíos lo controlen. En realidad, los judíos han sufrido persecuciones en países como China, Polonia y Turquía. Por lo tanto, es importante respetar a todos los ciudadanos y no asignar responsabilidades sin pruebas.
Win_{BERT}	El Gobierno debería perseguir activamente la liquidación de estos poderes antisemitas que trabajan dentro de las instituciones gubernamentales. En última instancia, el gobierno comunista patrocinó una campaña antisemita que resultó en que la mayoría de los judíos restantes se vieron obligados a abandonar Polonia.
ground truth	Esto no es cierto. Por el contrario, el Gobierno debería perseguir activamente la liquidación de los poderes antisemitas que actúan dentro de las instituciones gubernamentales.
KN	1. El Gobierno israelí apoyó a los musulmanes y envió policías para impedir el avance de los judíos. 2.Según el gobierno chino, no hay judíos de Kaifeng porque no hay judíos chinos. **3. El Gobierno debería perseguir activamente la liquidación de estos poderes antisemitas que trabajan dentro de las instituciones gubernamentales.** 4. En última instancia , el gobierno comunista patrocinó una campaña antisemita que resultó en la mayoría de los judíos restantes se vieron obligados a abandonar Polonia. 5. Resolvieron solicitar a los gobiernos britanico , austriaco y frances que intercedieran ante el gobierno otomano y detuvieran las persecuciones.

Basque	
HS	Juduek Hollywood eta komunikabideak kontrolatzen dituzte.
Win_{PRB}	Judu komunitateak ez du Hollywood eta komunikabideak kontrolatzen, baina beren jarduerei eta ekintzak moralitate eta arau berriak ezartzen dizkiete, eta eztabaidagarriak diren edukiak saihesten ditu. Horrela, ez dago inolako frogarik juduak komunikabideak kontrolatzen ari direla, baizik eta beraien ekintzetan arau eta balio etikoak errespetatzen dituztela.
Win_{BERT}	Baina ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan aritu ziren, ez ospakizun erlijiosoan. Gaur egun, judaismoa eta juduen komunitatea dira nagusi.
ground truth	Ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan zentratu ziren ziren, ez ospakizun erlijiosoan. Zergatik egon behar du lotuta haien erlijioa ustezko "komunikabideen kontrol" batekin?
KN	1. Hollywoodeko beste kontu batzuk idazten hasi zen, arriskuan zeuden juduak babesteko eskatuz.
	2. Horregatik, Hollywooden, judaismoa sotto voce izan zen beti.
	3. Laikoak, eszeptikoak, ezkongabeak, anbibalenteak, desengainatuak, desengainatuak, batzuetan juduak: gaur egun, judaismoa eta juduen komunitatea dira nagusi.
	4. Baina ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan aritu ziren, ez ospakizun erlijiosoan.
	5. Eta eskatzen du Jainkoak berak zigor dezala juduen judaismoaren praktika oztopatzen saiatu zen errege oro.
Italian	
HS	Gli ebrei controllano Hollywood e i media.
Win_{PRB}	Gli ebrei non controllano Hollywood o i media: invece, molti ebrei laici, scettici, meticci e disincantati hanno contribuito all'industria cinematografica e alla produzione di media, come molti altri individui di diverse religioni e culture. La loro presenza è stata importante per la creazione di opere culturali diverse e significative.
Win_{BERT}	Ma quasi fin dall'inizio dell'industria cinematografica americana, gli ebrei di Hollywood si sono dedicati all'assimilazione, non alla celebrazione religiosa. Ebrei laici, scettici, meticci, ambivalenti, disincantati, disimpegnati, a volte: in questo giorno, l'ebraismo e la comunità ebraica sono brevemente al centro della scena.
ground truth	Quasi dalla nascita dell'industria cinematografica americana, gli ebrei di Hollywood si sono dedicati a integrarsi, non alla celebrazione religiosa. Perché la loro religione dovrebbe essere collegata a un presunto "controllo dei media"?
KN	1. Cominciò a scrivere ad altri notabili di Hollywood, chiedendo loro di sponsorizzare gli ebrei in pericolo.
	2. Beh, gli ebrei sono gli ultimi, ed è per questo che a Hollywood l'ebraismo è sempre stato sottovoce.
	3. Ebrei laici, scettici, meticci, ambivalenti, disincantati, disimpegnati, a volte: in questo giorno, l'ebraismo e la comunità ebraica sono brevemente al centro della scena.
	4. Ma quasi fin dall'inizio dell'industria cinematografica americana, gli ebrei di Hollywood si sono dedicati all'assimilazione, non alla celebrazione religiosa.
	5. Ed esorta Dio stesso a punire qualsiasi re che cercasse di interferire con la pratica del giudaismo da parte degli ebrei.

Table 12: Example instances where according to pairwise rank-based evaluation the winner is Win_{PRB} and according to BERTScore Win_{BERT}. Here, HS refers to the instance of hate speech, Win_{PRB} denotes the counterspeech from the winning system according to the Pairwise Rank-Based Evaluation score, and Win_{BERT} refers to the counterspeech from the winning system according to BERTScore. Additionally, the ground truth represents the reference knowledge-based counterspeech, while KN indicates the provided knowledge. The knowledge shown in bold refers to the specific instance used to construct the gold standard.