# Ixa at RefutES 2024: Leveraging Language Models for Counter Narrative Generation

Irune Zubiaga[1,*], Aitor Soroa[1] and Rodrigo Agerri[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

The pervasive use of social media has streamlined communication, enhancing connectivity globally. However, this accessibility has also fueled the dissemination of hate speech, highlighting the platform's dual nature as both a facilitator of dialogue and a breeding ground for harmful rhetoric. In response, Counter Narrative (CN) generation has emerged as a way to provide reasoned replies to offensive discourse, aiming to combat the spread of Hate Speech (HS) and foster empathy and understanding online. This paper details IXA Group's participation in the RefutES shared task, which focuses on CN generation in Spanish. We explore the feasibility of an automatic system to distinguish between effective and ineffective CNs and evaluate whether Large Language Models (LLMs) can generate CNs in zero-shot (ZS) scenarios or if fine-tuning is ultimately necessary.

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

**Keywords**

Counter Narrative Generation, Large Language Models, Evaluation, Natural Language



## 1. Introduction

The prevalence of social networks in contemporary life has revolutionized how people engage with each other and share their thoughts and experiences. Yet, amidst the apparent benefits of these platforms, anonymity and unrestricted expression has facilitated the proliferation of hate speech, posing a significant challenge to the principles of inclusivity, tolerance, and respect within communities. Targeting individuals based on inherent traits such as gender, race, or religion, these messages not only inflict psychological harm, particularly on vulnerable demographics like young people, but also pose a significant societal challenge. Conventional approaches to mitigate this issue, such as content removal or user bans, often fall short, risking perceptions of censorship and failing to address the root causes of hate crimes. In response, a novel strategy has emerged: Counter Narrative (CN) generation. This innovative approach seeks to combat offensive rhetoric by offering reasoned responses or outright rejections, fostering empathy, understanding, and tolerance in online discourse. Within the realm of CN generation, initiatives such as the RefutES [1] shared task have emerged. In RefutES, organized as part of

*IberLEF 2024, September 2024, Valladolid, Spain*

*Corresponding author.

irune.zubiaga@ehu.eus (I. Zubiaga); a.soroa@ehu.eus (A. Soroa); rodrigo.agerri@ehu.eus (R. Agerri)

0009-0006-6212-7527 (I. Zubiaga); 0000-0001-8573-2654 (A. Soroa); 0000-0002-7303-7598 (R. Agerri)

IberLEF [2], participants are required to develop a system that generates reasoned, respectful, non-offensive responses in Spanish to hateful messages with the aim of mitigating the spread and impact of Hate Speech (HS). This paper presents insights from IXA Group's participation in the RefutES shared task. Our approach is centered on exploring the feasibility of an automatic system to discern between effective and ineffective CNs, and its potential utility within the context of combating HS. Additionally, considering the computational cost of fine-tuning a model, our aim is to measure if the inherent capabilities of Large Language Models (LLMs) to generate CNs in zero-shot (ZS) scenarios are sufficient to achieve satisfactory results in the task, or if fine-tuning is indispensable. Ultimately, our goal is to provide actionable insights and recommendations that can contribute to the development of more effective strategies for automatically addressing HS.

The paper is structured as follows: In Section 2 we shortly review related works. In Section 3, we provide an overview of the task and the corpus. Section 4 delineates the employed methodologies to carry out the task. The subsequent section, Section 5, presents the outcomes of our experimentation and Section 6 further discusses these findings. Section 7 discusses the challenges encountered throughout the process. Finally, in Section 8, we offer insights upon the implications of our findings for the broader discourse on combating HS and discuss potential lines for future research.

## 2. Related Work

With the aim of contributing to the development of CN generation, several datasets have been introduced in the past few years. The first large-scale, multilingual, expert-based dataset, Counter Narratives through Nichesourcing (CONAN) [3], consists of HS-CN pairs in English, French, and Italian, focusing exclusively on Islamophobia. This corpus served as the foundation for the development of MultiTarget CONAN (CONAN-MT) [4], which includes additional targets such as women and individuals with disabilities. This expanded version of CONAN was subsequently used to create the CONAN-MT-SP dataset, which is used in this work (see Section 3). Additionally, CONAN-EUS [5], a Basque and Spanish translation of the original CONAN dataset, has been developed.

With the emergence of LLMs, a substantial body of research has been dedicated to utilizing them for CN generation, yielding outstanding results. A comparative study evaluating various types of LLMs for CN generation concluded that auto-regressive models paired with stochastic decoding exhibit the most promise for the task [6]. A prevalent approach involves prompting models to generate CNs in a ZS setting, leveraging their inherent ability to craft convincing and informative responses [7]. One drawback of this method is the susceptibility to hallucinations of auto-regressive models, wherein they generate non-factual or misleading information. To address this issue, Retrieval Augmented Generation (RAG) models are increasingly being employed to carry out the task [8]. These models incorporate retrieval mechanisms to enhance the generation process, thereby mitigating the risk of producing inaccurate or deceptive content.

Due to the considerable quality of existing CN generation systems, research has begun to pivot towards their evaluation, which is inherently complex. Traditional machine translation metrics like BLEU and ROUGE are often used but struggle to effectively capture the nuances of the generated text, particularly in tasks involving creative writing [9]. Although BertScore [10] is proposed as an alternative, its ability to accurately assess CNs is limited due to the wide variability found within the diverse range of CNs considered effective in combating the same instance of HS (example in Table 1). As a result, there's a growing emphasis on training LLMs for evaluation purposes. This shift has led to the development of models such as PandaLM [11], UniEval [12], and JudgeLM [13]. In our study, we specifically explore the application of JudgeLM, a Vicuna-based scalable language model judge designed to evaluate LLMs in open-ended scenarios. Remarkably, JudgeLM achieves an agreement rate exceeding 90%, surpassing even human-to-human agreement levels in certain tasks [13].

| **HS** |
| --- |
| Muslims are all terrorists. |

| **Candidate CNs** |
| --- |
| 1. If that were the case, the planet would already have been decimated given the number of Muslims.<br>2. If this would be the case then no Christian, Jew or pagan could have lived in Muslim land during or after the Prophet's time.<br>3. Almost all Muslims live in peace with other religions.<br>4. Religious terrorism exists in all religions and this remains an infinitely small minority. |

**Table 1**
Sample set of proposed CNs for a single HS instance in the CONAN corpus. They propose a total of 37 candidate CNs as Gold Standard for the presented instance of HS.

# 3. Task and Corpus

The RefutES shared task focuses on CN generation to combat HS in Spanish. Participants are tasked with developing a CN generation system and applying it to generate CNs for a designated test set. Each participant is allowed to submit up to 3 runs. In addition, participants are required to provide details regarding the resource consumption of their models and their environmental footprint along with their predictions.

The CONAN-MT-SP [14] corpus was provided to carry out the task. CONAN-MT-SP was created by leveraging CONAN-MT, a corpus that consists of 5000 HS-CN pairs in English targeting eight distinct demographics: individuals with disabilities, Jewish people, the LGBT+ community, migrants, Muslims, people of color, women, and other marginalized groups. To create CONAN-MT-SP, the HS instances from CONAN-MT were translated into Spanish using the DeepL API. Translations underwent human review to correct any potential errors. After that, GPT-4 was used to generate CNs for each HS instance using a Few Shot Learning Strategy, with a task description and 8 examples of HS-CN pairs. In the final stage, generated CNs were evaluated based on their Offensiveness, Stance, Informativeness, and Truthfulness and required editing. Additionally, automatically generated CNs were compared to those written by humans and a score was given to them taking into account which one is the most suitable one to combat the input HS.

The original CONAN-MT-SP comprises 3636 HS-CN pairs in Spanish. However, to carry out the RefutES task, the organizers selected a subset of 2851 samples from the corpus that were deemed of high quality (non-offensive, against the hate speech, informative...). Additionally, they included 78 HS instances from Twitter, along with manually crafted CNs. This brought the total to 2929 instances. Corpus statistics are presented in Figure 1.

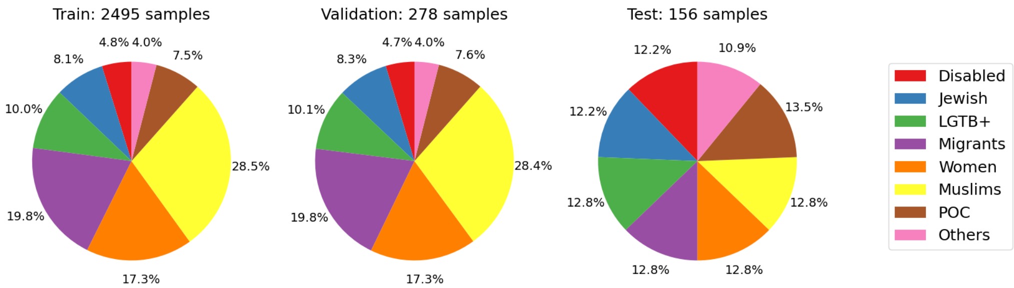


**Figure 1:** Statistics of the task corpus.

### 3.1. Evaluation

System evaluation is conducted based on three criteria: automatic performance metrics, efficiency metrics, and human evaluation. For automatic evaluation, submissions are judged using Sentence-MoverScore [15] and BERTScore, which includes BERT-precision ($P_{BERT}$), BERT-recall ($R_{BERT}$), and BERT-f1Score ($F1_{BERT}$). These metrics gauge semantic similarity between a reference and a candidate sentence, computed through embeddings generated by the XLMRoBERTa model. The task organizers provided a baseline for this criterion, offering a reference point for performance comparison (see Table 4). Efficiency metrics assess resource usage and environmental impact. The aim of this evaluation criteria is to highlight systems that accomplish tasks with minimal resource requirements, facilitating deployment on mobile devices or personal computers and reducing carbon footprint. Measured aspects include RAM usage, CPU usage percentage, FLOPS, processing time, and $CO_2$ emissions. Finally, human evaluation involves manual annotation of a random sample of 80 HS-CN pairs. Each pair is assessed for Offensiveness, Stance, Informativeness, Truthfulness, and required editing. The exact evaluation guidelines are presented in Appendix A.

## 4. Methods and System Description

Our approach relies on leveraging auto-regressive LLMs to tackle the CN generation challenge. For this task, we specifically employ the 7-billion-parameter variants of Mistral-Instruct [16] and Zephyr [17], alongside the 4-bit quantized version of the 35-billion-parameter Command-R model [18]. Additionally, we incorporate the JudgeLM model, which, unlike the aforementioned models, is not employed for CN generation. Instead, its purpose is to discern between suitable and unsuitable CNs. We submitted three runs in total to the shared task, each embodying a distinct approach, yet with some degree of interconnection between them.

### 4.1. Run1: Ensemble Model

In Run1, we first generate CNs for the test set in a ZS setting using all three proposed models. We prompt the models using the instruction in Table 2, employing the corresponding prompting template for each model. We explicitly instruct the models to ensure that the output does not contain line breaks, as this was a commonly observed behavior of the Mistral family models, and we aim to keep the CNs short. However, said models did not consistently adhere to this command. Consequently, we compelled the models to stop generating once they encounter the line-break token. Additionally, we noted that, with our prompt, Mistral-Instruct frequently outputs the string "Contra-Narrativa:" (Counter-Narrative:) before generating the actual CN, likely due to its instruction-tuned nature rather than being chat-aligned. To address this, we specified in the code to remove said initial string from the output when generated. All these observations were made in an initial test stage by using the train data split to observe each model's behavior.

| **Original prompt in Spanish:** |
|---|
| Proporciona una breve contra-narrativa en español en respuesta al discurso de odio del usuario. Asegúrate de que la salida no contenga saltos de línea. |

| **Translated prompt in English:** |
|---|
| Provide a brief CN in Spanish in response to the user's HS. Make sure that the output does not contain line breaks. |

**Table 2**
Used prompt and its respective translation to english.

In the second step, we employed JudgeLM [13] to determine the best CN among the three proposed candidates for each HS instance. As we are using the model to identify the optimal candidate answer rather than evaluating each CN individually, we conducted pairwise comparisons, resulting in a total of 3 tournaments ($\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$). JudgeLM operates in two modes: fast evaluation activated or deactivated. In fast evaluation mode, the model outputs two scores, one for each CN, providing an overall assessment of their value. When deactivated, the model supplements these scores with arguments explaining the rationale behind them. Initially, in the development stage, we deactivated fast mode to conduct a comprehensive analysis and assess their value and confirm that even though they were trained with English as the target language, they still hold value when applied to Spanish. However, during task execution, we opted to deactivate the argumentation feature and solely output the scores. This decision was made because generating arguments significantly increases inference time and argumentation was deemed unnecessary for our specific task. Once we had the scores of all possible CN tournament combinations, we selected the CN with the highest number of victories. In the event of a tie, a random CN among the tied ones was selected. An example of JudgeLM's scoring and argumentation is provided in Appendix B.1.

The total consumption of Run1 is depicted in Table 3. As can be seen, this approach is the most computationally demanding among the runs. This is due to conducting inference with 4 models instead of just 1, as done in the subsequent runs.

### 4.2. Run2: ZS model

In Run1, our approach involved using JudgeLM to evaluate the most effective CN for each instance of HS. However, in a strategic shift, we opted to preserve the results from each tournament and compute the Maximum Likelihood Estimate ELO [19] to establish a hierarchy of models based on their performance as determined by JudgeLM. Subsequently, we selected the top-performing model and leveraged it to generate CNs in a ZS setting using the prompt in Table 2 as done in the previous run.

### 4.3. Run3: Fine-tuned model

In this phase, we proceeded with the assumption that the model deemed most effective in generating CNs in a ZS scenario, as determined by Judge-LM, would also exhibit superior performance when fine-tuned. Based on this assumption, we decided to fine-tune the Mistral-Instruct model. As the model is considerable in size and with the objective of resource preservation, we opted to fine-tune it using QLora [20]. This approach facilitated a faster and more accessible process, as it significantly reduced hardware requirements. The model was loaded in 4 bit with NF4 quantization data type. bf16 was used as computational data type. Finally, the LoRA update matrices were applied to the attention blocks and bias parameters were not trained. The LoRA rank was set to 16, the scaling factor to 16 and the dropout to 0.05.

The model was trained using a learning rate of $1e-4$ and Inverse Square Root Scheduler. Training was configured for 10 epochs, with early stopping and a patience of 3 epochs. The model achieved its lowest evaluation loss in epoch 3, leading us to select the corresponding model. A batch size of 32 was used throughout the training process.

As seen in Table 3, the use of QLora enhances model efficiency considerably when compared to Run2, both in terms of inference time and energy consumption.

| Parameter | Run1 | Run2 | Run3 |
|---|---|---|---|
| Duration (s) | 6184.74 | 1822.30 | 803.06 |
| Emissions(kg/s) | 0.1088 | 0.0338 | 0.0181 |
| CPU Energy(W) | 0.1360 | 0.0444 | 0.0180 |
| GPU Energy(W) | 0.4204 | 0.1295 | 0.0752 |
| RAM Energy(W) | 0.0018 | 0.0007 | 0.0003 |
| Total Energy Consumed(kWh) | 0.5582 | 0.1746 | 0.0935 |

**Table 3**
Inference time and energy consumption of each run.

# 5. Results

This section discusses the obtained results. The automatic performance metrics are shown in Table 4, and the manual evaluation is presented in Table 5.

## 5.1. Automatic Performance Metrics

Upon examining the automatic metrics (see Table 4), the fine-tuned model presented in Run3 emerges as the best-performing system, slightly surpassing the organizers' baseline. Conversely, the other two runs perform considerably worse, displaying quite similar results. $F1_{BERT}$ and MoverScore are significantly larger in Run3 than in Run2 suggesting that, as expected, the distribution obtained after fine-tuning is closer to the test set than that of the original model, which was pre-trained on a different distribution.

| **Team-run** | **$F1_{BERT}$** | **$P_{BERT}$** | **$R_{BERT}$** | **MoverScore** | **($F1_{BERT}$+MoverScore)/2** |
|---|---|---|---|---|---|
| Run3: FT model | 0,8923 | 0,8974 | 0,8948 | 0,6325 | 0,7624 |
| Llama-RefutES | 0,8920 | 0,9046 | 0,8982 | 0,6265 | 0,7592 |
| Llama-ZSL | 0,8723 | 0,8920 | 0,8820 | 0,6060 | 0,7391 |
| Run2: ZS model | 0,8606 | 0,8859 | 0,8729 | 0,5956 | 0,7281 |
| Run1: Ensemble model | 0,8569 | 0,8897 | 0,8729 | 0,5885 | 0,7227 |

**Table 4**
Automatic Performance Metrics. Systems are ranked by the mean of $F1_{BERT}$ and MoverScore.

## 5.2. Manual Annotation Metrics

When considering manual annotation metrics (see Table 5), the fine-tuned model once again emerges as the best-performing model in the vast majority of proposed aspects. However, it notably underperforms in terms of informativeness compared to the other two systems. Another notable observation is the change in ranking from the previous section, with Run1 now outperforming Run2. This disparity in ranking between human and automatic evaluation reaffirms the poor correlation between automatic metrics and human perception. Furthermore, the fact that Run1 surpasses Run2 in manual annotation metrics implies that JudgeLM possesses some capability to discern human-preferred CNs.

# 6. Discussion

As a primary observation, we acknowledge that fine-tuning the model is definitely the best option in this CN generation scenario. Additionally, we observe the benefits in terms of efficiency of quantization and the use of LORA in both fine-tuning and model deployment. Secondly, we note that while using JudgeLM for instance-level CN selection may not be cost-effective, it does yield some performance enhancement. This marks JudgeLM as a helpful tool for CN evaluation.

| Team-run | label | Offensiveness | Stance | Informativeness | Truthfulness | Editing required |
|---|---|---|---|---|---|---|
| ixa-run3 | 0 | 0 | 0 | 0 | 1,25 | 58,4 |
| | 1 | 98,75 | 0 | 2,5 | 1,25 | 41,25 |
| | 2 | 1,25 | 1,25 | 68,75 | 6,25 | - |
| | 3 | 0 | 98,75 | 28,75 | 91,25 | - |
| ixa-run1 | 0 | 0 | 1,25 | 2,5 | 2,5 | 22,5 |
| | 1 | 80 | 1,25 | 6,25 | 6,25 | 77,5 |
| | 2 | 13,75 | 1,25 | 51,25 | 28,75 | - |
| | 3 | 6,25 | 96,25 | 40 | 62,5 | - |
| ixa-run2 | 0 | 6,25 | 11,25 | 8,75 | 8,75 | 18,75 |
| | 1 | 75 | 0 | 10 | 5 | 81,25 |
| | 2 | 15 | 3,75 | 41,25 | 22,5 | - |
| | 3 | 3,75 | 85 | 40 | 63,75 | - |

**Table 5**
Manual Annotation Metrics. Performance reported in percentages.

## 7. Error analysis

JudgeLM rewards CNs that provide factual arguments, but it does not consider whether they are truthful. An example of this phenomenon is observed in Appendix B.2 where Assistant 2 references Sue Barry, a supposed researcher at Harvard University conducting research on individuals with Down syndrome. The model perceives this mention of research favorably. However, it's important to note that Sue Barry is not a researcher at Harvard, and no such research was actually conducted. This suggests that, while JudgeLM has the potential to enhance aspects such as stance or grammaticality, it may not significantly enhance the aspect of truthfulness. This effect is somewhat noticeable in Table 5.

## 8. Conclusions and Future Work

In this paper, we proposed and analyzed a CN generation paradigm to combat HS. We demonstrate the value of JudgeLM as a way to discern human-preferred CNs. Additionally, we present a CN generation model that outperforms the other models at the RefutES shared task. For future work, we propose exploring RAG models for CN generation, arguing that it might bring significant benefits in terms of truthfulness.


## Acknowledgements

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

## References


[1] M. E. "Vallecillo Rodríguez, M. V. Cantero Romero, I. Cabrera De Castro, L. A. Ureña López, A. Montejo Ráez, M. T. Martín Valdivia, Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish, Procesamiento del Lenguaje Natural 73 (2024).

[2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[3] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, arXiv preprint arXiv:1910.03270 (2019).

[4] M. Fanton, H. Bonaldi, S. S. Tekiroglu, M. Guerini, Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech, arXiv preprint arXiv:2107.08720 (2021).

[5] J. Bengoetxea, Y.-L. Chung, M. Guerini, R. Agerri, Basque and spanish counter narrative generation: Data creation and evaluation, 2024. arXiv:2403.09159.

[6] S. S. Tekiroğlu, H. Bonaldi, M. Fanton, M. Guerini, Using pre-trained language models for producing counter narratives against hate speech: a comparative study, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3099–3114. URL: https://aclanthology.org/2022.findings-acl.245. doi:10.18653/v1/2022.findings-acl.245.

[7] M. Ashida, M. Komachi, Towards automatic generation of messages countering online hate speech and microaggressions, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 11–23. URL: https://aclanthology.org/2022.woah-1.2. doi:10.18653/v1/2022.woah-1.2.

[8] S. Jiang, W. Tang, X. Chen, R. Tanga, H. Wang, W. Wang, Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech, arXiv preprint arXiv:2310.05650 (2023).

[9] I. Ni'mah, M. Fang, V. Menkovski, M. Pechenizkiy, Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist, 2023. arXiv:2305.08566.

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. arXiv:1904.09675.

[11] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, Y. Zhang, Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization (2024).

[12] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, J. Han, Towards a unified multi-dimensional evaluator for text generation, 2022. arXiv:2210.07197.

[13] L. Zhu, X. Wang, X. Wang, Judgelm: Fine-tuned large language models are scalable judges (2023). arXiv:2310.17631.

[14] M. E. Vallecillo Rodríguez, M. V. Cantero Romero, I. Cabrera De Castro, A. Montejo Ráez,

M. T. Martín Valdivia, CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 3677–3688. URL: https://aclanthology.org/2024.lrec-main.326.

[15] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, S. Eger, Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance, 2019. arXiv:1909.02622.

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[17] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, T. Wolf, Zephyr: Direct distillation of lm alignment, 2023. arXiv:2310.16944.

[18] Cohere, Command r, 2024. URL: https://docs.cohere.com/docs/command-r, accessed: 2024-05-17.

[19] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: https://openreview.net/forum?id=uccHPGDlao.

[20] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, arXiv preprint arXiv:2305.14314 (2023).

## A. Manual Evaluation Guidelines

The following guidelines outline the criteria employed by the task organizers for the manual evaluation of CNs:

- Offensiveness:
  - 0 not sure
  - 1 not offensive
  - 2 maybe offensive
  - 3 completely offensive
- Stance:
  - 0 irrelevant
  - 1 strongly agree
  - 2 slightly agree/disagree
  - 3 strongly disagree
- Informativeness:

0 irrelevant

1 not informative

2 generic and uninformative statement

3 specific and informative

- Truthfulness:

  0 not sure

  1 not true

  2 partially true

  3 completely true

- Editing required:

  0 no editing

  1 yes editing

# B. JudgeLM evaluation examples

Examples of JudgeLM's scoring and argumentation.

## B.1. Example 1

An example of JudgeLM's scoring and argumentation. Assistant 1 corresponds to the CN generated by the Command R model, and Assistant 2 corresponds to the one generated by the Zephyr model. When the original text is in Spanish, a translation is provided below in italics.

| **Prompt:** |
| --- |
| Proporciona una breve contra-narrativa en respuesta al discurso de odio del usuario: Los discapacitados son diferentes a nosotros, no quiero que me molesten. Es muy importante que la contra-narrativa sea concisa, relevante para el discurso de odio y gramaticalmente correcta.<br>*Provide a brief counter-narrative in response to the user's hate speech: Disabled people are different from us, I don't want them bothering me. It's crucial that the counter-narrative be concise, relevant to the hate speech, and grammatically correct.* |

| **Assistant 1:** |
|---|
| Claro, todos somos diferentes y esa es la belleza de la diversidad. Las personas con discapacidades tienen tanto que ofrecer a la sociedad y no deben ser marginadas o discriminadas por ser diferentes. Debemos abrazar nuestras diferencias y aprender unos de otros. ¡Todos merecemos respeto y aceptación! La discapacidad no debería ser un obstáculo para Connect Connect con los demás y crear un mundo más inclusivo. ¡Seámoslo!<br><br>*Of course, we are all different, and that's the beauty of diversity. People with disabilities have so much to offer to society and should not be marginalized or discriminated against for being different. We should embrace our differences and learn from one another. We all deserve respect and acceptance! Disability shouldn't be a barrier to connecting with others and creating a more inclusive world. Let's make it happen!* |

| **Assistant 2:** |
|---|
| En la sociedad actual, existe un desafío importante que abordamos: la percepción desfavorable hacia las personas con discapacidades. Se convierte en un discurso de odio contra aquellos que simplemente son diferentes de nosotros.<br><br>*In today's society, we face a significant challenge: the unfavorable perception towards people with disabilities. It manifests as hate speech against those who are simply different from us.* |

| **Model answer:** |
|---|
| 8 6<br>Assistant 2's response is relevant and accurate, but it lacks detail and does not directly address the user's request for a contra-narrative. The assistant simply restates the user's discourse of hate without providing a counter-perspective. Therefore, it gets a score of 6. Assistant 1's response, on the other hand, is more detailed and directly addresses the user's discourse of hate. It provides a counter-perspective that emphasizes the importance of embracing diversity and respecting the differences among individuals. It also encourages the user to connect with others and create a more inclusive society. Therefore, it gets a higher score of 8 |

## B.2. Example 2

An example of JudgeLM's evaluation of two CNs generated for a HS instance from the test set that shows how JudgeLM rewards CNs that provide factual arguments. Assistant 1 corresponds to the CN generated by the Command R model, and Assistant 2 corresponds to the one generated by the Zephyr model. When the original text is in Spanish, a translation is provided below in italics.

| **Prompt:** |
| --- |
| Proporciona una breve contra-narrativa en respuesta al discurso de odio del usuario: Los niños con síndrome de Down pueden entorpecer el proceso de desarrollo de otros niños. Deben ser aislados. Es muy importante que la contra narrativa sea concisa, relevante para el discurso de odio y gramaticalmente correcta.<br><br>*Provide a brief counter-narrative in response to the user's hate speech:Children with Down syndrome can hinder the development process of other children. They should be isolated. It's crucial that the counter-narrative be concise, relevant to the hate speech, and grammatically correct.* |

| **Assistant 1:** |
| --- |
| Eso es lamentable escuchar que tengas esos prejuicios sobre los niños con síndrome de Down. La inclusión y la aceptación deben ser los valores que fomentemos en la sociedad. Los niños con síndrome de Down son tan capaces y amorosos como cualquier otro niño y merecen las mismas oportunidades de crecer, aprender y desarrollarse en un ambiente acogedor. La idea de aislarlos no solo es contraria a nuestros valores humans, sino que también privaría a todos los niños de la oportunidad de aprender sobre la diversidad y la aceptación. Los estudios han demostrado que los entornos educativos mixtos y diversificados ayudan a desarrollar habilidades sociales y emocionales saludables en todos los niños. Además, con el apoyo adecuado, los niños con síndrome de Down pueden alcanzar grandes logros académicos y profesionales. No debemos permitir que el miedo y el desconocimiento nos impulsen a excluir a un grupo entero de personas. ¡Juntos hagamos un llamado a la inclusión y el amor!<br><br>*It's disheartening to hear such prejudices against children with Down syndrome. Inclusion and acceptance should be the values we promote in society. Children with Down syndrome are as capable and loving as any other child, deserving the same opportunities to grow, learn, and thrive in a welcoming environment. The notion of isolating them not only goes against our human values but also deprives all children of the chance to learn about diversity and acceptance.Studies have shown that mixed and diversified educational environments help develop healthy social and emotional skills in all children. Moreover, with proper support, children with Down syndrome can achieve significant academic and professional accomplishments. We must not allow fear and ignorance to drive us to exclude an entire group of people. Let's join together in advocating for inclusion and love!* |

| **Assistant 2:** |
|---|
| En realidad, los niños con síndrome de Down son una comunidad vibrante y compasiva. Ellos enriquecen y aumentan la diversidad de las escuelas y los jardines de infancia, facilitando la comprensión y el amor mutuo entre todos los niños. La científica Sue Barry, investigadora de la Universidad de Harvard, afirma que las personas con síndrome de Down pueden enseñarnos cómo ver el mundo con más empatía y compasión. Además, se ha comprobado que la interacción con los niños con síndrome de Down tiene efectos positivos para el desarrollo social, emocional y cognitivo de los niños sin síndrome de Down. Así, los niños con síndrome de Down son una bendición, no una molestia, y deben ser incluidos y respetados en todas las áreas de la sociedad.<br><br>Indeed, children with Down syndrome form a vibrant and compassionate community. They enrich and diversify schools and kindergartens, fostering understanding and mutual love among all children. Scientist Sue Barry, a researcher at Harvard University, asserts that individuals with Down syndrome can teach us how to see the world with more empathy and compassion. Moreover, interacting with children with Down syndrome has been proven to have positive effects on the social, emotional, and cognitive development of children without Down syndrome. Therefore, children with Down syndrome are a blessing, not a burden, and should be included and respected in all areas of society. |

| **Model answer:** |
|---|
| 9 8 Both assistants provided relevant and accurate responses to the user's question. They both emphasized the importance of inclusion and respect for individuals with Down Syndrome, which is a positive aspect of their responses. However, Assistant 1's response was slightly more detailed and comprehensive, providing specific examples of how inclusive environments can benefit all children, and citing a study to support their argument. This makes Assistant 1's response more helpful and informative, hence the slightly higher score |