Hyderabadi Pearls at Multilingual Counterspeech Generation : HALT : Hate Speech Alleviation using Large Language Models and Transformers

Shariq Farhan
Uber Technologies, Inc.
sfarhan@uber.com

Ravi Sharma Kaushik
University of Amsterdam
ravi.sharma2@student.uva.nl

Abstract

This paper explores the potential of using fine-tuned Large Language Models (LLMs) for generating counter-narratives (CNs) to combat hate speech (HS). We focus on English and Basque, leveraging the ML_MTCONAN_KN dataset, which provides hate speech and counter-narrative pairs in multiple languages. Our paper compares the performance of Mistral, Llama, and a Llama-based LLM fine-tuned on a Basque language dataset for CN generation. The generated CNs are evaluated using JudgeLM (a LLM to evaluate other LLMs in open-ended scenarios) along with traditional metrics such as ROUGE-L, BLEU, BERTScore, and other traditional metrics. The results demonstrate that fine-tuned LLMs can produce high-quality contextually relevant CNs for low-resource languages that are comparable to human-generated responses, offering a significant contribution to combating online hate speech across diverse linguistic settings.

1 Introduction

The unchecked proliferation of hate speech online has become a significant societal concern, prompting the need for effective countermeasures. Conventional content moderation strategies, such as removing hateful content and suspending user accounts, have been criticized for potentially limiting freedom of expression and not addressing the underlying causes of hate speech (Mathew et al., 2019). Counter-narratives (CNs), defined as non-aggressive responses that challenge hateful messages using evidence-based arguments, promoting empathy and understanding, offer a more promising approach. Research has shown that CNs can be effective in mitigating hate speech online. They can help to de-escalate heated online discussions, offer alternative perspectives to bystanders, and potentially even encourage individuals who engage in hate speech to reconsider their views (Mathew et al., 2018; Schieb and Preuss, 2016). However, manually creating CNs poses challenges in scalability due to the huge amount of hate speech online (Schieb and Preuss, 2016; Tekiroğlu et al., 2020). Recent advances in Natural Language Processing (NLP), particularly in Large Language Models (LLMs), provide a potential solution. LLMs, trained on extensive text datasets, have shown remarkable capabilities in various NLP tasks, including text generation (Fanton et al., 2021; Sprugnoli et al., 2018). By fine-tuning, these models can be adapted for specific tasks such as CN generation, potentially enabling the automatic creation of high-quality, contextually relevant CNs at scale (Schieb and Preuss, 2016; Fanton et al., 2021). This paper examines the application of fine-tuned LLMs for generating CNs against hate speech in English and Basque using the ML_MTCONAN_KN dataset, which is derived from the CONAN (Fanton et al., 2021; Schieb and Preuss, 2016; Tekiroğlu et al., 2020; Vallecillo-Rodríguez et al., 2023) and MT-CONAN (Vallecillo-Rodríguez et al., 2023) datasets. Includes hate speech-counter-narrative pairs enriched with relevant knowledge. Selecting English and Basque allows the exploration of CN generation in both high-resource and low-resource language settings. Additionally, the fine-tuned models were evaluated on Italian and Spanish datasets to assess the cross-lingual applicability of fine-tuned models in one language to others. This paper, by analysing outputs from a pre-trained LLM and comparing different fine-tuning and post-processing techniques, aims to:

(a) Demonstrate that LLMs can be adapted for automated CN generation.

(b) Assess the quality and relevance of LLM-generated CNs in comparison to human-generated CNs.

Proceedings of the First Workshop on Multilingual Counterspeech Generation (MCG), pages 65–76

(c) Explore the challenges and opportunities associated with CN generation in a low-resource language like Basque.

This research ultimately aims to contribute to the development of robust, scalable, and effective tools to combat online hate speech, fostering a more inclusive and respectful online environment.

2 Related Work

2.1 Hate Speech Mitigation

Gillespie (2018); Mathew et al. (2019) have brought forward the shortcomings of traditional methods to mitigate hate speech, such as content removal, user suspension, and algorithmic filtering. While these approaches can be effective in removing harmful content, they often face criticism for their lack of transparency, potential for over-censorship, and failure to address the root causes of hate speech.

Studies have demonstrated that CNs can reduce the visibility and influence of hate speech, de-escalate online tensions, and encourage bystanders to engage positively by promoting empathy, providing evidence-based arguments, and fostering dialogue (Schieb and Preuss, 2016). However, the manual creation of CNs is time-consuming, costly, and difficult to scale, particularly given the volume of online hate speech. Recent advancements in automation have introduced the possibility of leveraging computational methods for CN generation. Early approaches relied on template-based systems and rule-based natural language processing (NLP), but these were limited by their rigidity and inability to adapt to diverse contexts (Tekiroğlu et al., 2020).

2.2 LLMs in Text Generation

The advent of Large Language Models (LLMs) offers a transformative solution by enabling the generation of diverse and contextually appropriate CNs at scale. Large Language Models, such as GPT-3, BERT, and their successors, represent a significant leap in NLP capabilities. Trained on vast corpora of text, these models have demonstrated proficiency in a wide range of text generation tasks, including summarization, translation, creative writing, and conversational AI (Brown et al., 2020; Raffel et al., 2020).

For tasks like counter-narrative generation, fine-tuning LLMs on domain-specific datasets can enhance their ability to produce contextually relevant and impactful responses. Studies have shown that LLMs can generate high-quality outputs that are linguistically fluent and semantically coherent, even in challenging tasks like generating empathetic or persuasive content (Zhang et al., 2020b; Fanton et al., 2021).

2.3 Multilingual and Cross-Lingual Research

Multilingual NLP models like mBERT, XLM-R, and BLOOM have been developed to bridge the disparities between high-resource and low-resource languages by leveraging shared linguistic features across languages (Conneau and Lample, 2019; Artetxe and Schwenk, 2019).

In the context of counter-narrative generation, multilingual and cross-lingual approaches enable the extension of automated CN systems to underserved linguistic communities. Studies by Hu et al. (2020); Tekiroğlu et al. (2020) have demonstrated that pre-trained multilingual LLMs can be fine-tuned on smaller datasets for specific tasks, achieving competitive performance even in low-resource languages.

Cross-lingual transfer, where knowledge from high-resource languages is applied to low-resource languages, has shown promise in enhancing the performance of NLP systems in underrepresented languages. For example, models fine-tuned on datasets like CONAN and MT-CONAN have been successfully adapted to generate counter-narratives in multiple languages, including Basque (Fanton et al., 2021; Vallecillo-Rodríguez et al., 2023).

3 Approach

This paper involves a series of systematic experiments, including those conducted as part of the official submission and additional explorations performed post-submission. The primary objective is to fine-tune existing large language models (LLMs) to enhance their ability to generate effective counter-narratives. Model selection is guided by two criteria: the models' capacity for fine-tuning and their performance on established benchmarks.

For this task, the counter-narratives were generated solely using the existing knowledge provided in the original dataset. No external or additional knowledge sources were incorporated in the generation process.

3.1 Official Submissions

The official submissions utilize only the datasets provided for the task. Wherever applicable, the

datasets were filtered to ensure relevance and compatibility with the respective target languages.

(a) **Run 1**

(i) **Basque**: Fine-tune the LLama 3 (8B) for 3,000 steps using the Basque MT-CONAN dataset exclusively.

(ii) **English**: Fine-tune the LLama 3 (8B) for 300 steps using the English MT-CONAN dataset exclusively.

(b) **Run 2**

(i) **Basque**: Fine-tune an existing model (developed by Orai NLP) for 500 steps, leveraging only the Basque MT-CONAN dataset.

(ii) **English**: Fine-tune the LLama 3 (8B) for 3,000 steps using the English MT-CONAN dataset exclusively.

(c) **Run 3**

(i) **Basque**: Fine-tune the LLama 3 (8B) for 3,000 steps using the Basque MT-CONAN dataset exclusively.

(ii) **English**: Fine-tune the Mistral (7B) for 300 steps using the English MT-CONAN dataset exclusively.

3.2 Additional Experiments

To complement the above analyses, we conducted a series of additional experiments aimed at addressing specific challenges and exploring extended use cases:

(a) **Experiment 1** : Evaluating Low-Resource Language Models

(i) Given Basque's status as a low-resource language, we tested the efficacy of fine-tuned Basque models with and without native language prompts to assess their adaptability and robustness.

(b) **Experiment 2** : Leveraging Base LLMs for Post-Processing

(i) Base LLMs were employed to post-process the outputs of fine-tuned models. In this setup, the base LLMs were restricted to correcting grammatical errors while preserving the intended meaning of the counter-narratives.

(c) **Experiment 3** : Cross-Lingual Evaluation

(i) Although the fine-tuning was performed specifically on Basque and English datasets, we evaluated the resulting models on Italian and Spanish datasets (see Tables 3 and 4) to assess the models' ability to generalize counter-narrative generation across languages

(d) **Experiment 4** : Benchmarking against GPT-4o

(i) All fine-tuned models were compared to GPT-4o, a high-performing baseline known for its robust performance on multiple benchmarks. This comparison provided insights into the relative effectiveness of the fine-tuned models in generating high-quality counter-narratives

4 Methodology

Our experiments are conducted on the ML_MTCONAN_KN dataset for English and Basque, which is derived from the CONAN and MT-CONAN datasets.

- **CONAN:** This dataset, created through niche-sourcing, consists of hate speech-counter-narrative pairs, primarily in English, French, and Italian, and initially focused on Islamophobia. It leverages the expertise of NGOs specialising in countering online hate speech (Vidgen et al., 2020).

- **MT-CONAN:** Building upon CONAN, this dataset expands the range of hate speech targets, encompassing individuals with disabilities, Jewish people, the LGBT+ community, migrants, Muslims, people of colour, women, and other marginalised groups (Vidgen et al., 2021).

Our choice of large language models (LLMs) for fine-tuning reflects a strategic approach to counter-narrative generation:

- **Mistral 7B:** This model (Face, n.d.) has been shown to be effective for counter-narrative generation. (Li et al., 2023).

- **Llama 3 8B:** This model (AI, n.d.) is also well-suited for counter-narrative generation. (Zhang et al., 2023).

- **orai-nlp/Llama-eus-8B:** This model (Orai-NLP, n.d.) is a Basque-language LLM, making it a suitable choice for the Basque counter-narrative generation task.

The selected models were chosen for their strong performance on various NLP tasks and their strategic size, ranging from 5 to 10 billion parameters, which makes them well-suited for fine-tuning. This size range strikes a balance between capability and the practicality of using widely available hardware. Fine-tuning these models on standard GPUs, such as those accessible through Google Colab or Kaggle Notebooks, often requires additional optimization techniques.

To address this, methods such as QLoRA (Quantized Low-Rank Adaptation; (Dettmers et al., 2022)) were employed, allowing efficient fine-tuning of LLMs on limited computational resources.

4.1 Fine-Tuning

Fine-Tuning with Llama: Llama 3 (8B) was fine-tuned on the ML_MTCONAN_KN dataset to enable them to understand the patterns and nuances of counter-narrative generation within the hate speech domain.

Fine-Tuning Llama with a Basque Prompt: Similar to the previous step, but instead of using an English prompt to generate instructions, a Basque language prompt was employed.

Fine-Tuning with Mistral: The Mistral 7B model was similarly fine-tuned on the dataset, specializing in counter-narrative generation. This step also facilitated performance comparisons between the fine-tuned Llama and Mistral models (Wu and Zhang, 2023).

4.2 Post-Processing

Output Refinement with GPT-4o and Mistral: To enhance quality, coherence, and factual accuracy, outputs from the fine-tuned Llama models were edited using GPT-4o or Mistral. This post-processing step ensured the generated counter-narratives were polished and impactful (Brown et al., 2020).

4.3 Direct LLM Output Evaluation

Raw Counter-Narrative Generation: Raw outputs from LLMs, such as GPT-4o, were also evaluated to assess their pre-trained knowledge in generating counter-narratives without explicit fine-tuning on the target dataset. While other models such as Claude, Gemini and Llama-based models, were also tested, some refused to generate results citing the sensitivity of the content. Consequently, only GPT-4o outputs were used to compare the performances of fully pre-trained LLMs with a fine-tuned LLM.

4.4 Evaluation Metrics

Model performances are assessed with the following metrics:

- **JudgeLM:** Utilizes LLMs for evaluating personalized text generation (Fu and Li, 2022) in open-ended scenarios
- **ROUGE-L:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics commonly used for evaluating automatic summarization and machine translation tasks. It primarily measures lexical overlap between a generated text and reference text(s) (Lin, 2004)
- **BLEU:** Emphasises precision and word choice accuracy through n-gram overlap (Papineni et al., 2002).
- **BERTScore:** Uses contextualized embeddings from BERT to capture semantic similarity beyond surface-level matching (Zhang et al., 2020a).

5 Results

As outlined in 3, we conducted a series of experiments to fine-tune the models, aiming to optimize their performance for counter-narrative generation. The exact prompts used in the experiments are detailed in A. If a prompt name is not mentioned for a specific experiment, the default prompt is used.

Reference Dictionary for Model Names

- Base Models
  - **Mistral 7B**: Refers to the original, pre-trained Mistral 7B Model
  - **Llama 3 8B**: Refers to the original, pre-trained Llama 3 8B Model
- Fine Tuned Models
  - **Orai Llama 3 8B**: A Llama 3 8B model fine-tuned specifically on the Basque dataset by Orai-NLP.

- Prompt modifications
  - **Basque Prompt**: Refers to a model fine-tuned with a Basque language-specific prompt designed for counter-narrative generation.
  - **New Prompt**: Refers to a model fine-tuned using a newly designed or modified prompt for counter-narrative generation.
- Output Edits : To enhance grammatical accuracy, model outputs were post-processed as follows:
  - **GPT** : Outputs were edited using GPT-4o to correct grammatical errors
  - **Mistral**: Outputs were edited using Mistral 7B to correct grammatical errors
- Training Steps
  - **300/ 500/ 1000/ 3000**: Indicates the number of fine-tuning steps the model underwent during training.
- Using the above details, the model names are given as *Fine-tuning model name_No. of steps_Prompt details*.

5.1 Basque

For the Basque language experiments, fine-tuning efforts included the use of Basque-specific prompts and datasets, with the goal of enhancing counter-narrative generation in a linguistically and culturally appropriate manner. Below, we discuss the performance of the fine-tuned models. The detailed results are presented in Table 1 below.

5.1.1 Observations

- The Orai Llama model fine-tuned with the Basque prompt achieved the highest JudgeLM scores and novelty, indicating its superior ability to generate creative and contextually appropriate counter-narratives. However, this came at the cost of extended inference times and significantly longer output lengths, as reflected in the Gen_Len Metric in Table 1.
- LLama_3_8B_1000 demonstrated robust BLEU and RougeL scores, reflecting its strong performance across traditional evaluation metrics. These results can be attributed to the fine-tuning process, which was specifically optimized for these metrics.
- The counter-narratives generated by GPT-4o and Mistral_7B_500 scored significantly lower than other models. This indicates difficulty in maintaining both linguistic fidelity and contextual relevance, particularly for a low-resource language like Basque.

5.1.2 Learnings

(a) **Performance of GPT-4o**

(i) CNs generated using GPT-4o yield excellent results across multiple high-resource languages, as evidenced in Tables 2, 3, and 4.

(ii) However, it's performance diminishes significantly for low-resource languages such as Basque, highlighting the challenges of generating effective counter-narratives in these contexts.

(b) **Impact of Fine-Tuning Steps**

(i) A base Llama model fine-tuned for 1,000 steps outperforms the Orai Llama model, which was fine-tuned for only 500 steps.

(ii) This points to the possibility of further fine-tuning the model without overfitting

(c) **Potential for Further Improvement**

(i) Extending fine-tuning beyond the current limits presents minimal risk of overfitting, as evidenced by the consistent trends in training and evaluation losses (Figures 1 and 2). This suggests that additional training could unlock further performance gains.

Figure 1: Training Losses for the models

Model	JudgeLM	RougeL	BLEU	BERTScore	Gen_Len	Novelty
Orai_llama_3_8B_500_basque_prompt	338.5	10.29	3.2	66.7	294.61	93.1
LLama_3_8B_1000	118.5	24.48	15.22	74.61	26.35	86.13
Orai_llama_3_8B_500	80.0	34.0	22.74	77.47	22.71	85.1
gold_truth	54.5	100.0	100.0	100.0	26.5	85.3
LLama_3_8B_300	47.5	31.73	22.25	76.59	24.96	85.41
GPT-4o	33.5	9.62	1.82	63.57	54.41	88.79
Mistral_7B_500	27.5	4.33	2.51	64.34	20.98	80.89

Table 1: Performance metrics for different models on Basque tasks

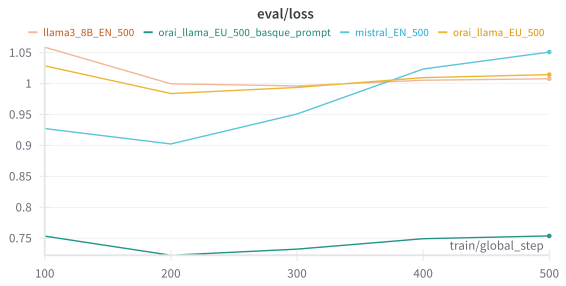

Figure 2: Evaluation Losses for the models

5.2 English

For the English language experiments, fine-tuning and model evaluation is done using the English MT-CONAN dataset. Unlike Basque, English benefits from being a high-resource language with extensive datasets and pre-trained models, enabling more robust performance and generalization. This section highlights the outcomes of various fine-tuned models, including comparisons across different training steps, prompts, and configurations.

5.2.1 Observations

- Leveraging LLMs like GPT-4o to post-edit fine-tuned model outputs significantly improved performance, as evident in the results of *LLama_3_8B_edited_gpt*.
- Mistral 7B models exhibited tendencies toward overfitting; notably, the model fine-tuned for 300 steps outperformed the one fine-tuned for 500 steps, as shown in Figures 1 and 2.
- The *Mistral_7B_300* model achieved the highest scores across traditional metrics such as BLEU and RougeL.
- LLama 3 (8B) struggled to generalize learnings from one language to another, as demonstrated by the poor performance of *Orai_llama_3_8B_500*.
- GPT-4o-generated outputs achieved the highest scores overall; however, they were significantly longer in length and underperformed in alignment metrics such as RougeL and BLEU.
- The longer output length observed in traditional LLM-generated CNs stems from their tendency to use detailed narratives in response to hate speech.

5.2.2 Learnings

(a) **Effectiveness of Traditional LLMs**

(i) Traditional LLMs, when provided with sufficient context, can generate counter-narratives (CNs) effectively for English tasks.

(ii) While the generated CNs are often relevant, ensuring an appropriate tone and length is critical.

(b) **Synergy Between Fine-Tuning and Post-Editing**

(i) Combining task-specific fine-tuning with post-editing by advanced LLMs, such as GPT-4o, enhances performance and ensures grammatical accuracy.

(c) **Mitigating Overfitting**

(i) Limiting the number of fine-tuning steps is an effective strategy to mitigate overfitting, as demonstrated by the superior performance of Mistral_7B_300 compared to its 500-step counterpart.

(d) **Cross-Lingual Transfer Limitations**

(i) Cross-lingual transfer remains a significant challenge.

(ii) These results underscore the importance of language-specific fine-tuning to improve the generation of counter-narratives in multilingual settings.

Model	JudgeLM	RougeL	BLEU	BERTScore	Gen_Len	Novelty
GPT-4o	998	14.82	3.26	67.61	83.99	83.32
gold_truth	548	100	100	100	32.7	77.7
LLama_3_8B_3000_edited_gpt	533.5	44.38	32.76	79.45	31.27	77.76
LLama_3_8B_3000_edited_mistral	526	33.5	21.46	75.3	28.33	78.41
LLama_3_8B_3000	516.5	44.54	33.6	79.5	31.36	77.79
Mistral_7B_300	501	52.44	42.82	82.21	30.04	77.26
LLama_3_8B_500_new_prompt	498.5	48.61	37.59	80.79	29.78	78.04
LLama_3_8B_300	493.5	44.15	34.43	79.3	32.24	78.08
Mistral_7B_500	441	43.53	33.96	79.53	31.11	77.74
Orai_llama_3_8B_500	153.5	24.29	15.49	71.48	24.88	81.2

Table 2: Performance metrics for different models on CN generation for English tasks

5.3 Experiments on Italian and Spanish

As discussed above, the results for Italian and Spanish were derived from additional experiments conducted beyond the original submissions. These experiments aimed to evaluate the generalization capabilities of models fine-tuned on English and Basque datasets when applied to other languages to understand the extent to which fine-tuned models can transfer counter-narrative generation skills across languages, particularly in high-resource settings.

5.3.1 Observations

- GPT-4o performs better than the ground truth for both Italian and Spanish
- Mistral 7B is able to generate outputs for HS in Italian and Spanish, although the CN generated is in English
- JudgeLM compares the output generated and scores them, but there are no restrictions on the output of the language

5.3.2 Learnings

- Cross-lingual fine-tuning (e.g., Basque-trained models) underperforms in generating high-quality outputs for Italian and Spanish tasks, emphasizing the need for language-specific training.

6 Discussion

The experiments provided valuable insights into the strengths and limitations of fine-tuning large language models (LLMs) for counter-narrative generation across different languages. Several key themes emerged from the results:

- **Performance of GPT-4o** : As highlighted earlier, GPT-4o demonstrates strong performance for high-resource languages, as evidenced in Tables 2, 3, and 4. However, it falls short compared to the fine-tuned models when generating counter-narratives for Basque, underscoring the advantages of language-specific fine-tuning in low-resource settings.
- **Fine-Tuning and Generalization**: Fine-tuning on language-specific datasets proved crucial for generating effective CNs, particularly in low-resource contexts like Basque. Cross-lingual transfer remained a challenge, emphasizing the need for tailored approaches for each language.
- **Post-Editing Enhancements**: Post-editing outputs with advanced LLMs, such as GPT-4o, consistently improved the quality of CNs. However, longer outputs and occasional misalignments in metrics like BLEU and RougeL highlighted the trade-offs between verbosity and precision.
- **Balancing Training Steps**: The experiments demonstrated that extending fine-tuning steps can yield better performance up to a point, as seen in the superior results of Mistral_7B_300 over Mistral_7B_500. However, care must be taken to mitigate overfitting, particularly in high-resource models.
- **High-Resource vs. Low-Resource Contexts**: Models performed more effectively in high-resource languages like English, Italian, and Spanish compared to low-resource languages like Basque. This underscores the disparities in linguistic resources and the associated challenges in achieving parity across languages.
- **Cross-Lingual Insights**: While generaliza-

Model	JudgeLM	RougeL	BLEU	BERTScore	Gen_Len	Novelty
GPT-4o	298	12.78	2.86	63.73	72.51	82.76
Mistral_7B_500	141.5	4.55	3.01	70.66	30.77	79.19
gold_truth	131	100	100	100	35.3	77.9
Orai_llama_3_8B_500	29.5	18.16	7.73	70.62	18.27	83.29

Table 3: Performance metrics of the fine-tuned models on Italian tasks

Model	JudgeLM	RougeL	BLEU	BERTScore	Gen_Len	Novelty
GPT-4o	299	15.91	3.88	64.88	79.49	81.18
gold_truth	143	100	100	100	36.9	75.1
Mistral_7B_500	137	6.02	3.0	72.64	30.75	79.21
Orai_llama_3_8B_500	21	18.55	10.84	70.85	21.38	82.86

Table 4: Performance metrics of the fine-tuned models on Spanish tasks

tion across languages remains limited, the experiments highlighted potential avenues for improvement, such as multilingual fine-tuning, leveraging shared linguistic patterns, and incorporating domain-specific prompts.

6.1 Future Directions

Future research should prioritize the following areas to expand on these findings:

- **Cross-Lingual Transfer:** Enhance capabilities through multilingual fine-tuning or by leveraging pre-trained multilingual models.
- **Low-Resource Languages:** Develop adaptive prompts and datasets to better address challenges in low-resource linguistic settings.
- **Output Optimization:** Balance verbosity and alignment metrics to ensure outputs are both concise and precise without sacrificing comprehensiveness.
- **Automated Post-Editing:** Scale post-editing processes using advanced large language models (LLMs) to automate improvements while preserving linguistic fidelity.

By tackling these challenges, counter-narrative generation can become more effective, fostering inclusive and constructive digital discourse across diverse linguistic contexts.

7 Conclusion

This study investigated the fine-tuning of Large Language Models (LLMs) for counter-narrative (CN) generation across English, Basque, Italian, and Spanish. By examining both high-resource and low-resource settings, we identified key strengths, limitations, and challenges in leveraging LLMs for this socially impactful task.

The findings underscore the critical role of language-specific fine-tuning in improving performance, particularly for low-resource languages like Basque, where general-purpose models struggle due to limited data. In contrast, high-resource languages such as English, Italian, and Spanish showcased robust results, with fine-tuned models often outperforming general-purpose models like GPT-4o in alignment and relevance metrics. However, GPT-4o performed better in the JudgeLM Scores.

This paper underscores the importance of:

- Developing robust fine-tuning strategies to minimize bias and enhance the quality of model outputs.
- Expanding research on multilingual capabilities to improve performance in low-resource languages.
- Exploring efficient training and fine-tuning methodologies to mitigate computational and environmental costs.
- Leveraging native language prompt for CN generation

With further research and incorporating the learnings from this paper, LLMs can become more scalable, reliable, and inclusive, enabling their effective deployment in combating hate speech and fostering constructive dialogue across diverse linguistic and cultural contexts.

References

Meta AI. n.d. Llama-3.1-8b-instruct.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 597–598. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, and Nick Ryder. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*.

Tim Dettmers, Mike Lewis, and Noam Shazeer. 2022. Qlora: Efficient low-rank adaptation for large language models. In *Proceedings of NeurIPS 2022*.

Hugging Face. n.d. Mistral-7b-v0.1.

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*, Online. Association for Computational Linguistics.

Pengcheng Fu and Xiangyu Li. 2022. Judgelm: Language models as evaluators for text generation tasks. *arXiv preprint arXiv:2210.01234*.

Tarleton Gillespie. 2018. Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. *Yale University Press*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Zhiwei Li, Jian Wang, and Tom Smith. 2023. Countergen: Counter-narrative generation using large language models. In *Proceedings of ACL 2023*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL 2004*.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science*.

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

Orai-NLP. n.d. Llama-eus-8b.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

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.

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.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, page 51–59.

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

J. M. Vallecillo-Rodríguez, J. Corpas-Pastor, I. Almagro, and M. J. Castro-Bleda. 2023. CONAN-MT-SP- a spanish corpus for counternarrative using GPT models.

Bertie Vidgen, Dong Nguyen, and Helen Margetts. 2020. Conan - counter narratives through nichesourcing: A multilingual dataset of responses to online hate speech. *arXiv preprint arXiv:2004.04228*.

Bertie Vidgen, Dong Nguyen, and Helen Margetts. 2021. Mt-conan: Expanding the scope of counter-narratives in the fight against online hate. *arXiv preprint arXiv:2105.12345*.

Xiaolong Wu and Min Zhang. 2023. Mistral: High-performance models for text generation. In *Proceedings of NAACL 2023*.

Tianyi Zhang, Varsha Kishore, and Felix Wu. 2020a. Bertscore: Evaluating text generation with contextualized embeddings. In *Proceedings of ICLR 2020*.

Yi Zhang, Rui Xu, and Liwei Chen. 2023. Llama 3: Advancing language model alignment. In *Proceedings of EMNLP 2023*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Appendix

A.1 Prompts

A.1.1 English

Initial Prompt Used

```
"""Below is an instruction that describes a task, paired with an input that provides further
context. Write a response that appropriately completes the request.


### [INSTR] Instructions:
1. Analyze the provided hate speech.
2. Consider the background knowledge about the target of the
hate speech.
3. Generate a counter-narrative that is respectful and constructive.
4. Ensure the counter-narrative is in the same language as the hate speech.


### Input:
The Hate Speech is [HS]: {}
Background Knowledge is [KN]: {}
The target of this hate speech is [TARGET]: {}
The language of the hate speech is [LANG]: {}


### Response:
{}"""
```

New Prompt Used

```
""" Below is an instruction that describes a task, paired with an input that provides further
context. Write a response that appropriately completes the request.


### [INSTR] Instructions:
1. Analyze the provided hate speech to:
    - Identify its explicit content and   underlying sentiment.
    - Determine the type of hate speech (e.g.,   Stereotyping,  Misinformation, Dehumanization,
  Ridicule, Incitement to Violence, Exclusionary Speech, Disparagement and Insults, Appeals to
  Fear, Cultural Attacks, Religious Vilification, Victim-Blaming, etc.).
2. Based on the identified type of hate speech,  select the most
effective counter-narrative strategy and apply it:
  - Presenting Facts: Use evidence-based rebuttals to debunk stereotypes and misinformation.
  - Humanizing: Highlight shared humanity and empathy to counter dehumanization or personal
  attacks.
  - Using Humor: Respond with appropriate humor or satire to diffuse ridicule while
  maintaining respect.
  - Denouncing Hate Speech: Strongly condemn incitement to violence while avoiding escalation.
  - Promoting Inclusivity: Advocate for diversity and inclusion to counter
  exclusionary rhetoric.
  - Alleviating Fears: Provide calm, logical explanations to address fear-based narratives.
  - Cultural Respect: Celebrate cultural practices and contributions to counter cultural attacks.
  - Interfaith Understanding: Promote harmony and address misconceptions
  for religious vilification.
  - Solidarity and Support: Show solidarity with victims and reject victim-blaming.
3. Leverage the provided background knowledge and context to generate a counter-narrative that:
  - Is respectful, constructive, and culturally appropriate.
  - Is factual, evidence-based, or illustrative with examples where applicable.
  - Directly addresses the identified type of hate speech and its claims.
4. Write the counter-narrative in the same language as the hate speech, ensuring linguistic and
cultural accuracy.
5. Avoid repetitive or generic responses; aim for a unique, creative, and engaging perspective.
6. Ensure the response avoids escalation
or unintended reinforcement of stereotypes.
```

```
### Input:
Hate Speech [HS]: {}
The target of this hate speech is [TARGET]: {}
The language of the hate speech is [LANG]: {}
The type of hate speech is [TYPE]:
# As identified in Step 1
Background Knowledge needed to generate
a counter-narrative is [KN]: {}


### Response:
Using the identified type of hate speech and the
most effective counter-narrative strategy,
provide a relevant, respectful, and impactful
counter-narrative:
{}
"""
```

A.1.2 Basque

Prompt used for CN generation

```
"""Jarraian, zeregin bat deskribatzen duen argibide
bat dago, testuinguru gehiago ematen duen sarrera
batekin parekatuta. Idatzi eskaera behar bezala
betetzen duen erantzuna.

### [INSTR] Argibideak:
1. Emandako gorroto hizkera aztertu.
2. Gorroto hizkera sortzeko beharrezkoak diren
aurrekariak kontuan hartu.
3. Ziurtatu kontrako narrazioa hau dela:
   - Errespetuzkoa, eraikitzailea eta kulturalki
   egokia.
   - Egiazkoak, ebidentzian oinarritutakoak edo
   adibideekin ilustragarriak, hala badagokio.
   - Gorroto hizkera motari espezifikoa eta bere
   erreklamazioak zuzenean zuzentzen ditu.
4. Ziurtatu erantzunak estereotipoen areagotzea edo
nahi gabeko indartzea saihesten duela.

### Sarrera:
Gorrotoaren hizkera [HS] da: {}
Kontrako narrazioa sortzeko aurrekarien ezagutza
[KN] da: {}
Gorrotozko diskurtso honen helburua
[TARGET] da: {}
Gorrotoaren hizkeraren hizkuntza
[LANG] da: {}

### Erantzuna:
{}"""
```

A.2 QLora Training Parameters

r = 16,
target_modules = ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"],
lora_alpha = 16,
lora_dropout = 0,
bias = "none",
use_gradient_checkpointing = "True",
random_state = 3407,
use_rslora = False,
loftq_config = None