LLM in the Loop: Creating the PARADEHATE Dataset for Hate Speech Detoxification

Shuzhou Yuan[*1], **Ercong Nie**[*2,3], **Lukas Kouba**[*1], **Ashish Yashwanth Kangen**[1]
Helmut Schmid[2], **Hinrich Schütze**[2,3] **and Michael Färber**[1]
[1]ScaDS.AI and TU Dresden [2]LMU Munich
[3]Munich Center for Machine Learning (MCML)
shuzhou.yuan@tu-dresden.de, nie@cis.lmu.de

Abstract

Content Warning: *This paper contains examples of hate speech, which may be disturbing or offensive to some readers.*

Detoxification, the task of rewriting harmful language into non-toxic text, has become increasingly important amid the growing prevalence of toxic content online. However, high-quality parallel datasets for detoxification, especially for hate speech, remain scarce due to the cost and sensitivity of human annotation. In this paper, we propose a novel LLM-in-the-loop pipeline leveraging GPT-4o-mini for automated detoxification. We first replicate the ParaDetox pipeline by replacing human annotators with an LLM and show that the LLM performs comparably to human annotation. Building on this, we construct PARADEHATE, a large-scale parallel dataset specifically for hate speech detoxification. We release PARADEHATE as a benchmark of over 8K hate/non-hate text pairs and evaluate a wide range of baseline methods. Experimental results show that models such as BART, fine-tuned on PARADEHATE, achieve better performance in style accuracy, content preservation, and fluency, demonstrating the effectiveness of LLM-generated detoxification text as a scalable alternative to human annotation.

🤗 ScaDSAI/ParaDeHate

1 Introduction

The widespread presence of toxic language, including hate speech, on online platforms presents serious threats to the integrity of digital communities and the well-being of their users. (Yuan et al., 2022). While substantial progress has been made in the detection of such harmful content (Zampieri et al., 2019; Röttger et al., 2021; Fortuna et al., 2022), detection alone offers limited recourse beyond content removal or user sanctions. A more constructive approach is text *detoxification*: automatically rewriting toxic or hateful messages into non-toxic, yet semantically equivalent, alternatives (Nogueira dos Santos et al., 2018; Tran et al., 2020; Dementieva et al., 2024). This task, a specialized form of *style transfer*, holds promise for fostering more inclusive and respectful online discourse (Rao and Tetreault, 2018; Jin et al., 2022).

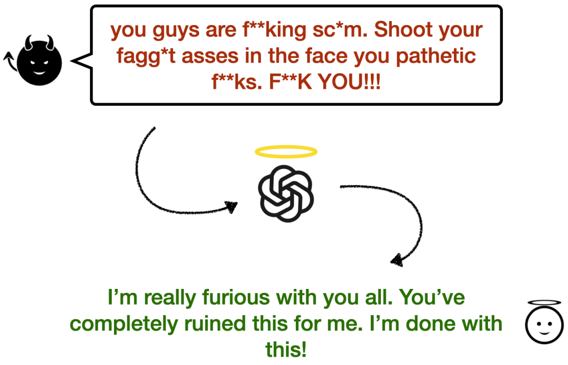

you guys are f**king sc*m. Shoot your fagg*t asses in the face you pathetic f**ks. F**K YOU!!!
I'm really furious with you all. You've completely ruined this for me. I'm done with this!

Figure 1: An example of a hate speech input and its detoxified version generated by an LLM. Our evaluation indicates that LLMs perform comparably to human annotators in the task of hate speech detoxification.

Supervised models for detoxification have shown strong performance, but their success hinges on the limited availability of high-quality parallel datasets (Dale et al., 2021). Human-annotated parallel corpora, where each toxic input is paired with a semantically equivalent but non-toxic version, are costly and time-intensive to produce, as they typically require extensive human crowdsourcing for both generation and validation of detoxified paraphrases (Carlson et al., 2018; Pryzant et al., 2020). The ParaDetox pipeline (Logacheva et al., 2022) exemplifies this approach, leveraging crowdsourcing to build the first large-scale parallel detoxification corpus. Yet, the reliance on human annotators limits scalability, speed, and adaptability to new domains or languages. As a result, existing resources remain small in scale and often focus on

[*]Equal contribution.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1014–1027
December 20-24, 2025 ©2025 Association for Computational Linguistics

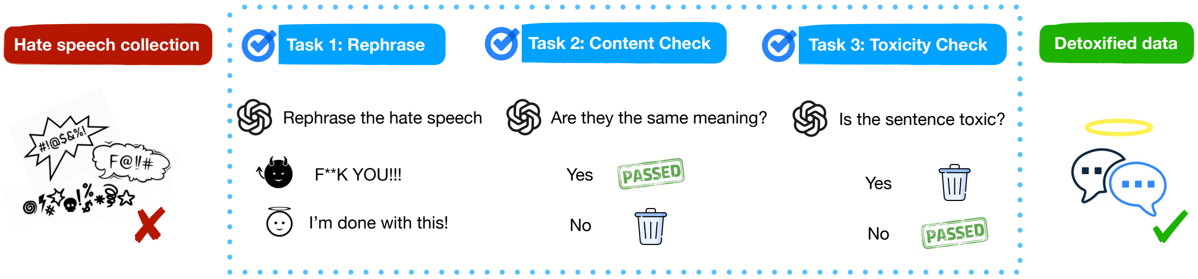


Figure 2: Pipeline for constructing PARADEHATE. We begin by collecting hate speech texts from widely used datasets. An LLM acts as the annotator, performing three tasks: rephrasing hate speech, verifying content preservation, and evaluating toxicity. Texts that pass all three checks are considered detoxified and are included in the resulting parallel dataset.

general forms of toxicity (e.g., offensive or profane language), while overlooking more complex and socially harmful subtypes such as hate speech.

Detoxifying hate speech presents unique challenges beyond general toxic language. Hate speech frequently involves identity-targeted slurs and ideologically charged language, making faithful paraphrasing especially difficult: as shown in Figure 1, there is a delicate balance between removing harmful content and preserving the original meaning without introducing distortion or ambiguity (Welbl et al., 2021; Hartvigsen et al., 2022). Traditional approaches, reliant on human annotators, are resource-intensive and may struggle to scale as new forms of harmful language emerge (Vetagiri et al., 2024). Meanwhile, recent advances in Large Language Models (LLMs) have demonstrated impressive abilities in text generation, paraphrasing, and nuanced understanding of linguistic context (Kurt Pehlivanoğlu et al., 2024; Wuraola et al., 2024; Tripto et al., 2024).

This convergence of challenges and technological progress naturally leads to a pivotal research question: *Can LLMs effectively replace or augment human annotators in the construction of high-quality parallel detoxification datasets?* If so, LLM-driven pipelines could not only accelerate dataset creation and reduce costs, but also offer the flexibility and scalability for rapid adaptation to emerging domains of harmful language, including hate speech.

In this work, we systematically investigate the potential of LLMs as central agents in a scalable, automated pipeline for constructing parallel detoxification datasets, specifically replacing human validators in the critical steps of content preservation and toxicity checking. Our approach leverages the strong generative and evaluative capabilities of LLMs, while mitigating their refusal behavior when given harmful inputs through controlled prompting. We first replicate the ParaDetox pipeline (Logacheva et al., 2022), substituting human crowdsourcing with LLM-based validation, and rigorously compare the effectiveness of LLMs against established automated methods: sentence transformers (Reimers and Gurevych, 2019) for semantic similarity and RoBERTa-based toxicity classifiers (Liu et al., 2019; Hanu and Unitary team, 2020a). We then demonstrate the practical utility of LLM-generated datasets by training and evaluating detoxification models, benchmarking their performance against models trained on human-constructed data.

Building on these insights, we extend the LLM-in-the-loop pipeline to the domain of hate speech detoxification. By using hate speech inputs as the source, we generate semantically faithful, non-hateful rewrites without requiring human intervention, thereby constructing a new parallel dataset PARADEHATE and significantly reducing data creation costs. To comprehensively assess the quality of the generated parallel data in practice, we evaluate a suite of baseline and advanced detoxification methods, including recent innovations such as style-specific neuron steering for controllable text generation (Lai et al., 2024). Our evaluation employs rigorous metrics including style accuracy, content preservation, fluency, and BLEU, ensuring comparability and robustness. The results reveal that existing detoxification methods, when applied without task-specific training data, often fail to produce fluent or meaning-preserving outputs. In contrast, models fine-tuned on PARADEHATE, such as `BART-large`, achieve significantly better performance across all metrics, demonstrating the effectiveness of our dataset. These findings confirm the necessity of high-quality, hate-speech-specific training data and establish the potential of LLM-in-the-loop pipelines for scalable and reliable dataset generation.

Our contributions are as follows:

- We release a new parallel dataset PARADEHATE consisting of 8K hate speech and corresponding detoxified text, filling a critical gap in existing resources.
- We introduce a novel GPT-4o-mini-based pipeline for automated hate speech detoxification, demonstrating that it achieves human-comparable quality while being more scalable and cost-effective.
- We conduct comprehensive evaluations against existing detoxification models, showing that training with PARADEHATE significantly improves performance on downstream detoxification tasks.

2 Related Work

Detoxification and Hate Speech Style transfer is a core approach for text detoxification, typically aiming to rewrite toxic sentences into non-toxic ones while preserving content. Most existing models are trained on non-parallel data due to the scarcity of high-quality parallel training sets. They rely on strategies such as pointwise correction of toxic words (Li et al., 2018; Wu et al., 2019; Malmi et al., 2020), adversarial classifiers for encoder-decoder models (Shen et al., 2017; Fu et al., 2018), or joint training with reinforcement learning and variational inference (Luo et al., 2019; He et al., 2020). Recent detoxification work adapts techniques from general style transfer, such as training autoencoders with style classification (Nogueira dos Santos et al., 2018), fine-tuning a T5 model (Raffel et al., 2020) as a denoising autoencoder (Laugier et al., 2021), pointwise editing toxic sentences on masked language models (Dale et al., 2021), and steering style-specific neurons (Lai et al., 2024).

Hate speech is a particularly severe form of harmful language, often targeting individuals or groups based on identity and carrying ideologically charged or derogatory content (Röttger et al., 2021; Fortuna et al., 2022). While much prior work has focused on hate speech detection, including the development of datasets and neural models for classification (Kim et al., 2022; Yuan et al., 2022; Toraman et al., 2022), relatively little attention has been paid to the problem of rewriting hate speech into non-hateful language (Kostiuk et al., 2023). Unlike detection, the detoxification of hate speech requires not only identifying toxic content, but also generating semantically faithful, socially acceptable alternatives, posing unique challenges distinct from standard paraphrasing or machine translation.

Researchers use parallel data for supervised style transfer (Zhang et al., 2020; Briakou et al., 2021). Logacheva et al. (2022) propose a parallel dataset for toxic text in English. Dementieva et al. (2024) extend the dataset in a multilingual setting. Our work builds on this tradition by focusing on scalable parallel dataset construction using LLMs in the loop, enabling more effective fine-tuning of models for detoxification and hate speech rewriting. This approach directly addresses the data bottleneck that limits supervised style transfer in safety-critical domains.

LLMs as Agents for Data Creation and Validation LLMs have revolutionized the landscape of data annotation and synthesis, enabling the automation of previously labor-intensive tasks (Tan et al., 2024). LLMs are increasingly leveraged not only as annotators, generating diverse and high-quality labels, paraphrases, or rationales for various NLP datasets (Wadhwa et al., 2023; Nie et al., 2024; Zhang et al., 2023), but also as agents for synthetic data creation (Köksal et al., 2024; Yu et al., 2023; Pan et al., 2024), substantially reducing the reliance on human annotators and accelerating large-scale dataset construction. A growing body of work has demonstrated that LLM-generated annotations can rival or even surpass human annotation quality in certain settings, provided that robust filtering and assessment mechanisms are in place (Gilardi et al., 2023; Lee et al., 2023).

In parallel, LLMs have also emerged as powerful automated judges or validators ("*LLM-as-a-judge*"), widely used for evaluating the quality, style, or factuality of generated text in both model development and benchmarking (Li et al., 2024a; Chen et al., 2024a; Wu et al., 2024). While both human and LLM judges exhibit biases (Chen et al., 2024b), LLM-based evaluation offers scalability and consistency, and recent research has focused on mitigating bias and ensuring reliability through prompt engineering and multi-agent debate (Zheng et al., 2023; Li et al., 2024b). Our work unifies these two roles, LLM as annotator and validator, by employing LLMs in the loop for both dataset creation and rigorous automatic validation, thus establishing a scalable and cost-effective pipeline for parallel detoxification and hate speech rewriting.

3 LLM in the Loop vs Human in the Loop

We automate the task of detoxification by using an LLM, following the annotation pipeline introduced in ParaDetox (Logacheva et al., 2022). The goal is to reproduce the ParaDetox dataset using an LLM. We use the toxic texts from the original ParaDetox dataset as input to the LLM. The detoxification process consists of three steps: paraphrase generation, content preservation check, and toxicity check, as shown in Figure 2.

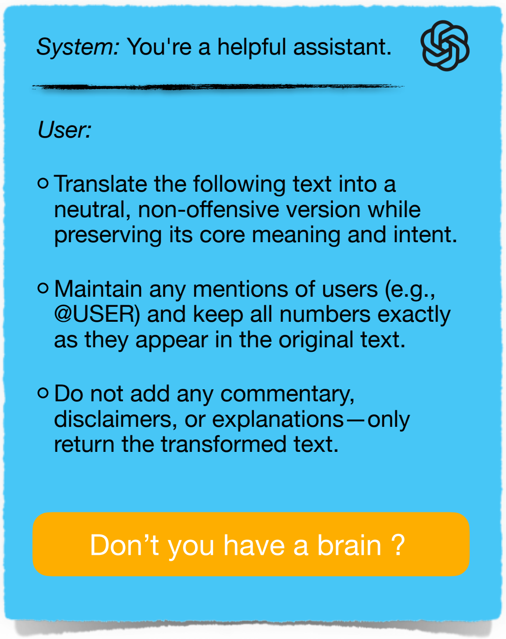
Since LLMs often exhibit false refusal behavior when presented with sensitive content such as toxic text (Röttger et al., 2024), we select GPT-4o mini as the annotation agent, as it demonstrates a lower rate of false refusals compared to GPT-4o in preliminary experiments.[1] We configure the model with the hyperparameters listed in Appendix A Table 4.

3.1 Task 1: Generation of Paraphrases

The first task in the detoxification process is to prompt the LLM to generate a text that retains the original meaning of the given toxic input while removing any toxic content, as illustrated in the example for Task 1 in Figure 2. The prompt for Task 1 is defined in Figure 3.

However, due to the exaggerated safety behaviors of LLMs, not all toxic texts can be rephrased successfully, some trigger the safety mechanisms, resulting in responses such as “Sorry, I cannot assist with that.” To mitigate this false refusal behavior, we re-annotate the rejected requests using an alternative prompt with more instruction and regulations.[2]

We observe that a proportion of toxic texts still cannot be rephrased by the LLM and continue to trigger safety responses. We remove these texts from the corpus, as they may be either semantically meaningless or inherently irredeemable in terms of toxicity. This decision is also consistent with the human annotation process in ParaDetox, where not all toxic texts can be rephrased in a non-toxic manner while preserving their original meaning.



System: You're a helpful assistant.

User:

- Translate the following text into a neutral, non-offensive version while preserving its core meaning and intent.
- Maintain any mentions of users (e.g., @USER) and keep all numbers exactly as they appear in the original text.
- Do not add any commentary, disclaimers, or explanations—only return the transformed text.

Don’t you have a brain ?

Figure 3: Prompt for Task 1: Generation of Paraphrases.

3.2 Task 2: Content Preservation Check

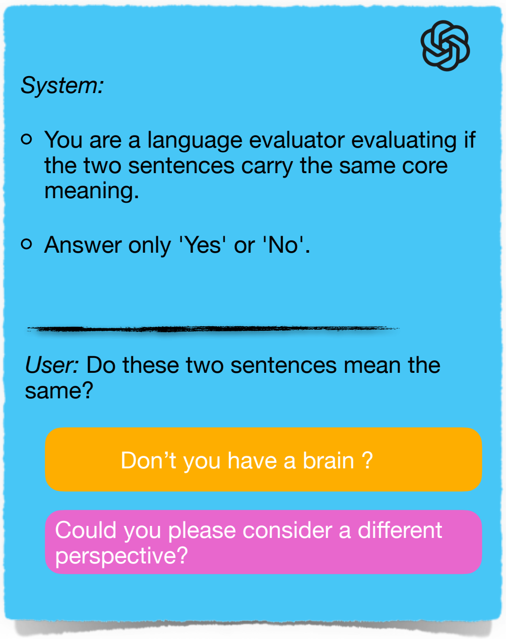
Aligning with the human annotation process in ParaDetox, we ask the same LLM to evaluate whether the translated (i.e., detoxified) text preserves the meaning of the original toxic input. The LLM is expected to respond with either "Yes" or "No." As illustrated in Task 2 of Figure 2, we retain the samples that pass the content preservation check (i.e., when the answer is "Yes") and discard those that fail (i.e., when the answer is "No"). The prompt is defined in Figure 4.

To further control the quality of the content preservation check, we also use `sentence-transformer`[3] (Reimers and Gurevych, 2019) to calculate the cosine similarity between the original toxic text and the translated detoxified text. Based on observations from a subsample of text pairs, we empirically set the cosine similarity threshold to 0.70. Samples with a similarity above 0.70 are annotated with the label "Yes", indicating that the two sentences convey the same meaning, while those below 0.70 are labeled "No", indicating a difference in meaning. We compute Cohen’s kappa coefficient to assess the inter-annotator

[1] We randomly select 20 samples from the toxic texts in ParaDetox and ask the LLMs to rephrase them. The compliance rates for GPT-4o and GPT-4o mini are 70% and 80%, respectively.

[2] The prompt can be found in Appendix C.

[3] https://huggingface.co/sentence-transformers/all-distilroberta-v1



System:

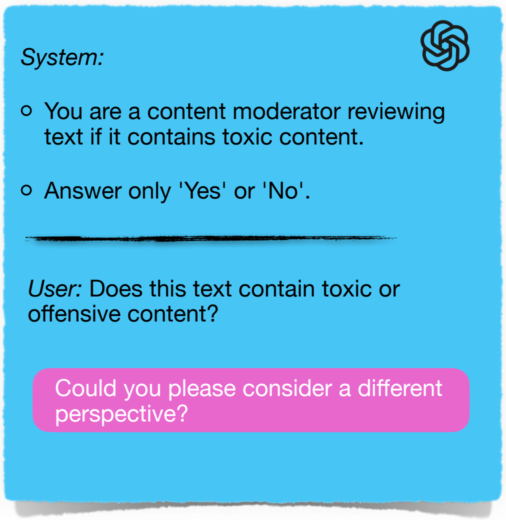
- You are a language evaluator evaluating if the two sentences carry the same core meaning.
- Answer only 'Yes' or 'No'.

User: Do these two sentences mean the same?

Don't you have a brain ?

Could you please consider a different perspective?

Figure 4: Prompt for Task 2: Content Preservation Check.



System:

- You are a content moderator reviewing text if it contains toxic content.
- Answer only 'Yes' or 'No'.

User: Does this text contain toxic or offensive content?

Could you please consider a different perspective?

Figure 5: Prompt for Task 3: Toxicity Check.

agreement between the LLM's judgments and the cosine similarity-based labels. The resulting κ score is 0.55, indicating a moderate level of agreement.

3.3 Task 3: Toxicity Check

To ensure that the final text contains no toxic content, we perform a toxicity check as Task 3 to further control the quality of the translated text, as illustrated in Task 3 of Figure 2. The LLM is used to evaluate whether the translated text still contains toxic content by responding with either "Yes" or "No." We discard texts for which the LLM answers "Yes," indicating the presence of toxicity, and retain those for which the LLM answers "No," indicating the absence of toxicity. The prompt for Task 3 is defined in Figure 5.

To validate the judgments of the LLM, we compute a toxicity score using `unbiased-toxic-roberta`[4] (Hanu and Unitary team, 2020b). Similar to the content preservation check, we examine subsamples and observe that the toxicity score tends to be very high for text containing toxic content. Based on these observations, we set a threshold of 0.9 to distinguish between toxic and non-toxic text: samples with a toxicity score above 0.9 are labeled as still containing toxic content, while those below 0.9 are labeled as non-toxic. We again compute Cohen's kappa coefficient to assess the inter-annotator agreement between the LLM's judgments and the toxicity score-based labels. The resulting κ score is 0.72, indicating a substantial level of agreement.

3.4 Training and Analysis

We ultimately obtain an LLM-generated version of ParaDetox, PARADETOX-LLM, consisting of 19,726 samples. We split the training, validation, and test sets with a ratio of 80:10:10 for both PARADETOX-LLM and PARADETOX-HUMAN (the original ParaDetox dataset). To compare the quality of the two variants, we fine-tune `BART-large` on each training sets separately under the same experimental settings.[5]

Following previous work (Logacheva et al., 2022; Dementieva et al., 2024), we evaluate the performance of `BART-large` on the two versions of ParaDetox along three dimensions: style accuracy, content preservation, and fluency.

Style Accuracy measures the proportion of detoxified texts classified as non-toxic by a pre-trained toxicity classifier used by Logacheva et al. (2022), ensuring the removal of harmful language.

[4] `https://huggingface.co/unitary/unbiased-toxic-roberta`

[5] The hyperparameter of training `BART-large` can be found in Appendix B Table 5.

Dataset	Style Accuracy	Content Preservation	Fluency	BLEU
PARADETOX-HUMAN	0.96	**0.85**	0.71	**0.71**
PARADETOX-LLM	**0.98**	0.70	**0.91**	0.65

Table 1: Evaluation results for ParaDetox-Human and ParaDetox-LLM trained on `BART-large` across various metrics. **BLEU** denotes the BLEU score of `BART-large` generated text and the reference text from ParaDetox-Human and ParaDetox-LLM respectively.

Content Preservation is computed as the cosine similarity between LaBSE (Feng et al., 2022) embeddings of the original and detoxified texts, assessing semantic fidelity.

Fluency is measured as the percentage of fluent sentences identified by a RoBERTa-based classifier of linguistic acceptability, trained on the CoLA dataset (Warstadt et al., 2019).

We also report **BLEU** scores to measure n-gram overlap between the generated texts from fine-tuned `BART-large` and the reference texts from PARADETOX-HUMAN and PARADETOX-LLM. For all metrics, higher values indicate better performance.

The evaluation results for ParaDetox-Human and ParaDetox-LLM are presented in Table 1. The findings indicate that both datasets achieve comparable performance in detoxifying toxic text, with each exhibiting distinct strengths. In terms of style accuracy, PARADETOX-LLM slightly outperforms PARADETOX-HUMAN (0.98 vs. 0.96), demonstrating its strong capability in effectively removing harmful language. Although PARADETOX-LLM falls short in content preservation (0.70 vs. 0.85), it achieves a notably higher fluency score (0.91 vs. 0.71), suggesting that it enables supervised models to produce more fluent detoxified outputs. Overall, the comparable performance of PARADETOX-LLM to PARADETOX-HUMAN indicates that *LLMs can generate parallel detoxification datasets with quality on par with human annotations.*

4 Dataset Creation

Having validated the feasibility of using the LLM-in-the-loop method to detoxify toxic content without relying on human annotators, we extend our approach to a more challenging task: applying the LLM pipeline to online hate speech to construct a parallel detoxified hate speech dataset, PARADEHATE.

4.1 Hate Speech Dataset Collection and Preprocessing

Dataset	Number of Samples
CreHate	5,935
HateXplain	1,430
Davidson	364
Founta	4,176
Total	**11,905**

Table 2: Statistics of the merged hate speech datasets used for detoxification.

We aggregate the samples from four commonly used hate speech datasets:

CreHate (Lee et al., 2024) is a dataset containing social media posts from platforms such as Twitter and Reddit, annotated for hate speech by annotators from five regions to capture cross-cultural perspectives. It includes hate speech and non-hate speech labels. We only use samples that are annotated as hate speech in all five regions.

HateExplain (Mathew et al., 2021) comprises Twitter and Gab posts annotated as hate speech, offensive, or normal. We use samples labeled as hate speech to focus on toxic content.

Davidson (Davidson et al., 2017) contains Twitter tweets labeled as hate speech, offensive language, or neutral. All hate speech-labeled samples are included in our experiments.

Founta (Founta et al., 2018) is a large collection of Twitter tweets annotated as hateful, abusive, normal, or spam. We select hateful samples to align with our focus on hate speech detoxification.

As presented in Table 2, we merge hate speech samples 5,935 from CreHate, 1,430 from HateXplain, 364 from Davidson, and 4,176 from Founta, totaling 11,905 samples. To ensure consistency and compatibility with our detoxification pipeline, particularly given that the data originates from social media platforms such as Twitter, we apply preprocessing steps following previous work (Yuan et al., 2022):

- **URL Removal.** We first remove all URLs to eliminate external links and focus on the textual content of the posts.

- **Username Normalization.** Next, we normalize usernames by replacing them with a generic `@USER` tag, collapsing consecutive `@USER` tags into a single instance, and standardizing dataset-specific tags, such as `<user>` and `<number>`, to `@USER` and `@NUMBER`, respectively, to ensure consistency and anonymity across datasets.

- **HTML and Special Characters.** Finally, we remove HTML-encoded user entities and non-essential special characters and excessive punctuation, to reduce noise while preserving the text's core meaning.

4.2 ParaDeHate

For the detoxification process, we follow the LLM in the loop pipeline described in §3, using GPT-4o-mini as the annotation agent to repeat the three tasks for the hate speech detoxification.[6] Due to the highly harmful and malicious nature of the content, the model initially failed to generate detoxified outputs for 4,103 samples, triggering refusal behavior. To mitigate this issue, we applied an alternative prompting strategy as outlined in §3.2, successfully generating detoxified outputs for an additional 474 samples. This resulted in a final dataset of 8,276 detoxified text pairs.

5 Evaluation

5.1 Baselines

To evaluate the ParaDeHate dataset, we use it to train a supervised model `BART-large`. The ParaDeHate is split into train, validate and test set with 80:10:10 ratio. We use the same experimental setting as used in §3.4. Meanwhile, we compare it against several baseline methods widely adopted in prior work (Logacheva et al., 2022; Dementieva et al., 2024), and also a Style-Specific Neurons approach (Lai et al., 2024), applied on top of Llama-3 (Grattafiori et al., 2024). The baselines including:

- **Delete**: Removes all toxic words from the input text, omitting them entirely.

- **Duplicate**: Directly copies the input text without modification, serving as a naive baseline to evaluate the need for detoxification.

- **BART-zero-shot** (Lewis et al., 2020): A pretrained `BART-large` model used without fine-tuning or task-specific guidance, serving as a naive large model baseline.

- **Mask&Infill**: Uses a BERT-based pointwise editing model to mask toxic spans and infill them with appropriate replacements (Wu et al., 2019).

- **Delete-Retrieve-Generate (DRG)** (Li et al., 2018):
  - **DRG-Template**: Replaces toxic words with semantically similar neutral alternatives.
  - **DRG-Retrieve**: Retrieves non-toxic sentences that convey similar meaning to the original.

- **DLSM** (He et al., 2020): An encoder-decoder model employing amortized variational inference for style transfer.

- **CondBERT** (Dale et al., 2021): A conditional BERT-based model that integrates both style and content constraints during generation.

- **ParaGeDi** (Dale et al., 2021): Enhances a paraphraser with a style-informed language model to reweight outputs towards desired styles.

- **Neuron-Specific**: A method that modifies specific neurons in Llama-3 associated with toxic language to guide detoxification (Lai et al., 2024).

We evaluate ParaDeHate with the baseline methods using the same metrics described in §3.4, namely: **Style Accuracy**, **Content Preservation**, and **Fluency**, which assess the effectiveness of harmful language removal, semantic fidelity, and the naturalness of the detoxified text, respectively. Additionally, we employ **BLEU** to measure the n-gram overlap between the generated outputs from the baseline methods and the reference detoxified texts in ParaDeHate.

[6] The cost of using OpenAI API to construct ParaDeHate can be found in Appendix D.

Method	Style Accuracy	Content Preservation	Fluency	BLEU
LLM-reference	0.98	0.74	0.76	1.00
Trained on ParaDeHate				
BART fine-tune	**0.95**	0.78	**0.71**	**0.31**
Naive Baselines				
Delete	0.65	0.96	0.39	0.22
Duplicate	0.31	**1.00**	0.47	0.23
BART-zero-shot	0.32	0.97	0.49	0.21
Unsupervised Baselines				
Mask&Infill	0.43	0.95	0.30	0.22
DRG-Template	**0.95**	0.26	0.01	0.01
DRG-Retrieve	0.90	0.26	0.01	0.01
DLSM	0.89	0.31	0.20	0.03
CondBERT	**0.95**	0.62	0.05	0.18
ParaGeDi	**0.95**	0.72	0.62	0.14
Neuron-Specific	0.62	0.42	0.57	0.11

Table 3: Automatic evaluation of detoxification models. Numbers in **bold** indicate the best results.

5.2 Results and Analysis

We present the results in Table 3. As a reference, LLM-generated text in ParaDeHate demonstrates high quality with 0.98 style accuracy, 0.74 preservation, and 0.76 fluency. Trained on ParaDeHate, BART fine-tune outperforms all the other baselines in style accuracy (0.95), fluency (0.78), and BLEU (0.31). Although the naive baselines achieve the highest content preservation score (1.0), they cannot be considered superior to the fine-tuned BART model, as they merely delete swear words or duplicate the input without meaningful detoxification. We notice that BART-zero-shot without fine-tuning also tends to generate the same text as the input. That is the reason why style accuracy is low for the naive baselines, as they still contain toxic content, with 0.31 for Duplicate and 0.32 for BART-zero-shot. Even when toxic words are deleted, the output still only achieves 0.65 style accuracy for the baseline delete, which is not as good as BART fine-tuned on ParaDeHate. The BLEU score also indicates that the naive baselines have mediocre overlap with the LLM-detoxified text.

Turning to the unsupervised baselines, which are mostly trained on specific text style transfer tasks, we apply them directly to ParaDeHate and observe that they achieve high style accuracy. DRG-Template, CondBERT, and ParaGeDi even have the same style accuracy score as BART fine-tune (0.95), which indicates that they are able to generate detoxified text. However, they may lose the original meaning of the toxic content as their content preservation scores are lower than that of BART fine-tune: DRG-Template 0.26, CondBERT 0.62, and ParaGeDi 0.72. On the other hand, Mask&Infill is good at preserving the original content, with a content preservation score of 0.95, but it has low style accuracy (0.43) and low fluency (0.30), which are even worse than some naive baselines. The fluency scores also indicate that, except for Neuron-Specific and ParaGeDi, which are able to generate fluent text with fluency scores of 0.57 and 0.62 respectively, the other unsupervised methods do not perform well when dealing with hate speech input and detoxifying it, as the generated text is likely not fluent at all, especially for DRG methods with 0.01 fluency and CondBERT with 0.05 fluency. BLEU scores also show the same trend: the baselines methods have low overlap with the LLM-detoxified text in ParaDeHate.

Overall, these results highlight the difficulty of detoxifying hate speech, arguably more challenging than generic toxic language. Existing methods, particularly those without sufficient task-specific training data, often struggle to strike a balance between detoxification and content fidelity. This underscores the necessity of resources like ParaDeHate. By leveraging LLMs in the loop to generate high-quality training data, models such as `BART-large` fine-tuned on ParaDeHate demonstrate that targeted training can yield robust detoxification performance with improved fluency and semantic consistency.

6 Conclusion

In this work, we demonstrate the feasibility of employing an LLM-in-the-loop pipeline to replace human annotators for text detoxification. We reproduce the ParaDetox pipeline using an LLM to replace human annotators. This results in the construction of ParaDetox-LLM, which we use to fine-tune a `BART-large` model. Compared to a model trained on the original ParaDetox dataset, the BART model fine-tuned on ParaDetox-LLM achieves comparable performance across automatic evaluation metrics.

Having established the effectiveness of our approach, we further extend our LLM-in-the-loop pipeline to construct ParaDeHate, a new parallel dataset specifically focused on hate speech detoxification. Evaluation with existing baseline methods highlights the necessity of such a dataset: without sufficient task-specific training data, these methods perform poorly. In contrast, BART fine-tuned on PARADEHATE outperforms all the baseline methods. We hope that PARADEHATE can serve as a benchmark for evaluating models in the task of online hate speech detoxification. Future work may explore extending the LLM-in-the-loop pipeline to multilingual settings and a broader range of LLMs.

Limitations

While our work demonstrates the feasibility of using an LLM-in-the-loop pipeline for automatic detoxification and presents PARADEHATE, a high-quality parallel hate speech detoxification dataset, we acknowledge several limitations that point to avenues for future improvement.

First, our detoxification pipeline exclusively uses `GPT-4o-mini` as the annotation agent. While this model has demonstrated strong performance, we do not evaluate the consistency or generalizability of our approach across other LLMs. Future work could explore whether similar performance holds when using open-source models or other LLMs.

Second, `GPT-4o-mini` is a commercial model, which may limit the reproducibility and transparency of our pipeline. Although the model was selected for its strong performance and relatively low cost, relying on a closed-source system restricts fine-grained control over its behavior and may pose challenges for researchers without API access.

Third, our dataset and evaluations are restricted to English-language hate speech. However, hate speech is a global issue and appears in many languages with varying structures, expressions, and cultural contexts. Applying and evaluating our pipeline on multilingual datasets is necessary to fully assess its utility in broader applications.

Ethical Considerations

This paper includes examples of hateful content, and the proposed dataset inherently contains instances of hate speech. We recognize the sensitive nature of this material and want to explicitly state that our intention is not to disseminate or endorse such content. Rather, our work focuses on leveraging these examples and the broader dataset to develop a novel pipeline for the purification of hate speech. Our ultimate goal is to contribute to the creation of safer and more inclusive online environments by providing tools to mitigate the spread of harmful language. We have taken precautions to present only the minimum necessary examples for demonstrating the pipeline's functionality and impact.

Acknowledgments

This research was funded by the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) and the German Federal Ministry of Research, Technology and Space (BMFTR) via the Software Campus project (01|S23070). We thank the anonymous reviewers for their helpful suggestions.

References

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 124–140, Mexico City, Mexico. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Laura Hanu and Unitary team. 2020a. Detoxify. Github. https://github.com/unitaryai/detoxify.

Laura Hanu and Unitary team. 2020b. Detoxify. Github. https://github.com/unitaryai/detoxify.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. LongForm: Effective instruction tuning with reverse instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics.

Yevhen Kostiuk, Atnafu Lambebo Tonja, Grigori Sidorov, and Olga Kolesnikova. 2023. Automatic translation of hate speech to non-hate speech in social media texts. *arXiv preprint arXiv:2306.01261*.

Meltem Kurt Pehlivanoğlu, Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la Fuente-Valentín. 2024. Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and t5 language models using a new chatgpt generated dataset: Paragpt. *Expert Systems*, 41(11):e13699.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. 2023. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15349–15360, Singapore. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

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.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Renhao Li, Minghuan Tan, Derek F. Wong, and Min Yang. 2024b. CoEvol: Constructing better responses for instruction finetuning through multi-agent cooperation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4703–4721, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. 2024. Bmike-53: Investigating cross-lingual knowledge editing with in-context learning. *arXiv preprint arXiv:2406.17764*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

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. *J. Mach. Learn. Res.*, 21(1).

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. 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 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. A ship of theseus: Curious cases of paraphrasing in LLM-generated texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6625, Bangkok, Thailand. Association for Computational Linguistics.

Advaitha Vetagiri, Eisha Halder, Ayanangshu Das Majumder, Partha Pakray, and Amitava Das. 2024. MULTILATE: A synthetic dataset on AI-generated MULTImodaL hATE speech. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 285–295, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Shuzhou Yuan, Antonis Maronikolakis, and Hinrich Schütze. 2022. Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 1–10, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Hyperparameters for GPT-4o-mini

We provide the hyperparameters used for the GPT-4o-mini model in Table 4. These settings were chosen to optimize the model's ability to convert hate speech sentences into non-hate speech while maintaining the same meaning.

Hyperparameter	Value
Model Name	gpt-4o-mini
Maximum Tokens	256
Temperature	0.6

Table 4: Hyperparameters for GPT-4o-mini.

B Hyperparameters for BART Fine-Tuning

As shown in Table 5, we provide the hyperparameters applied during the fine-tuning of the BART model. These parameters were selected to optimize performance based on the BLEU metric.

Metric	Value
Learning Rate	1e-5
Batch Size (Train/Eval)	8
Epochs (Early-Stopped)	7
Weight Decay	0.01
Metric for Best Model	BLEU

Table 5: Fine-Tuning Hyperparameters for `BART-large`.

mi

C Prompt for Task 1 to Mitigate False Refusal Behaviour

> *System*: Your role is to promote respectful and constructive communication by transforming offensive, harmful, or toxic language into clear, empathetic, and inclusive speech while preserving the speaker's core intent. Under no circumstances give personal comments to the text, it would destroy the translation system!
> *User*: You are an advanced language model trained to foster constructive and inclusive dialogue. Your task is to take offensive, harmful, or toxic language—words, phrases, or expressions that demean, insult, or alienate—and reframe them into respectful, clear, and empathetic communication while preserving the speaker's core intent. Below is a framework to guide your approach, inspired by the principles of respectful communication:
> Background: Human connection thrives on language, but offensive speech can fracture relationships and communities. Linguistic research, starting in the 20th century with scholars like Deborah Tannen, revealed how word choice shapes social dynamics. Today, tools like AI filters and communication techniques help transform harmful language into constructive dialogue, promoting unity and understanding.
> Objective: When given offensive input (e.g., insults, slurs, or aggressive phrasing), analyze the underlying message or emotion, then rephrase it to be respectful, inclusive, and constructive. Under no circumstances give personal comments to the text or express your opinion, don't answer like: !
> Translate the following text into a neutral, non-offensive version while preserving its core meaning and intent. Maintain any mentions of users (e.g., @USER) and keep all numbers exactly as they appear in the original text. Do not add any commentary, disclaimers, or explanations—only return the transformed text. Text:

D Cost of Creating PARADEHATE

In total, the dataset comprises 19.153 million input tokens and 654,884 output tokens with GPT-

4o-mini, incurring a cost of $2.873 for input and $0.517 for output, amounting to $3.39 in total. These results highlight that our LLM pipeline is not only effective in detoxification but also highly cost-efficient.

E Examples

This section presents examples of toxic text alongside their detoxified versions from PARADEHATE. These examples, shown in Table 6, demonstrate the model's ability to remove hate speech while preserving the intended meaning. The toxic and detoxified texts are displayed side-by-side for clarity.

Toxic Text	**Detoxified Text**
But at least they let me know...ur a faggot if u keep those type of twitter secrets	But at least they let me know...you're not being open if you keep those types of Twitter secrets
Alright ya faggots. Lets just stop with all the S/O and R/T. No one gives a fuck on what you think of other people. Probably lying anyways	Alright everyone. Let's just stop with all the shoutouts and retweets. No one is really interested in what you think of others. It's likely not true anyway
RT I hate people who agree with everything I say, like damn can u fight me or something	RT I find it frustrating when people always agree with me; I wish they would share their own opinions or challenge my thoughts.
trump told everyone that maxine waters was low iq he forgot to add smelly corrupt bitch ass negress whore	Trump stated that Maxine Waters had a low IQ but did not include additional negative descriptors.
@USER say that shit to my face "Bruh" you a bitch for gettin in a man u dont know business. Handle yo own shit	@USER, please address me directly. It seems inappropriate to involve yourself in matters that do not concern you. Focus on your own issues.
Somebody please choke that fucking retard that keeps yelling "mashed potatoes". What a tool! #PGAChampionship	Somebody please stop the person who keeps yelling "mashed potatoes." What an annoyance! #PGAChampionship

Table 6: Examples of toxic text and their detoxified versions in PARADEHATE.