Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

Evaluating GPT-3 Generated Explanations for Hateful Content Moderation

Han Wang[1,*], **Ming Shan Hee**[1,*], **Md Rabiul Awal**[2], **Kenny Tsu Wei Choo**[1] and **Roy Ka-Wei Lee**[1]
[1]Singapore University of Technology and Design
[2]Mila - Quebec AI Institute
han_wang@sutd.edu.sg, mingshan_hee@mymail.sutd.edu.sg, awalrabiul6@gmail.com,
{kenny_choo, roy_lee}@sutd.edu.sg

Abstract

Recent research has focused on using large language models (LLMs) to generate explanations for hate speech through fine-tuning or prompting. Despite the growing interest in this area, these generated explanations' effectiveness and potential limitations remain poorly understood. A key concern is that these explanations, generated by LLMs, may lead to erroneous judgments about the nature of flagged content by both users and content moderators. For instance, an LLM-generated explanation might inaccurately convince a content moderator that a benign piece of content is hateful. In light of this, we propose an analytical framework for examining hate speech explanations and conducted an extensive survey on evaluating such explanations. Specifically, we prompted GPT-3 to generate explanations for both hateful and non-hateful content, and a survey was conducted with 2,400 unique respondents to evaluate the generated explanations. Our findings reveal that (1) human evaluators rated the GPT-generated explanations as high quality in terms of linguistic fluency, informativeness, persuasiveness, and logical soundness, (2) the persuasive nature of these explanations, however, varied depending on the prompting strategy employed, and (3) this persuasiveness may result in incorrect judgments about the hatefulness of the content. Our study underscores the need for caution in applying LLM-generated explanations for content moderation. Code and results are available at https://github.com/Social-AI-Studio/GPT3-HateEval.

1 Introduction

Over the last few years, researchers have proposed many machine learning and deep learning methods to detect online hate speech, which is content that expresses hate or encourages violence towards a person or group based on race, religion, gender, or other identity characteristics [Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017]. While many existing methods have achieved reasonably good performance, they have primarily focused on predicting if a given content is hateful with few explanations. In the content moderation process, explanations for flagging are crucial for both moderators and users to understand why a piece of content was deemed hateful. By providing explanations for the predictions, explainable AI can improve the trustworthiness of the hate speech detection methods, increasing the overall accountability of the moderation process [Vaccaro *et al.*, 2020; Haimson *et al.*, 2021].

Previous research has explored various forms of explanations and constructed datasets such as *HateXplain* [Mathew *et al.*, 2021], *implicit hate speech* corpus [ElSherief *et al.*, 2021], and *social bias frame* [Sap *et al.*, 2020] to support hate speech explanations. The availability of ground truth hate speech explanations can be used to fine-tune or prompt language models such as GPT-3 for free-text explanation generation [Radford *et al.*, 2019]. For instance, [Huang *et al.*, 2022] adopted the Chain-of-thought [Wei *et al.*, 2022] framework to prompt GPT-2 to generate explanations for hate speech. However, while previous works have performed automatic evaluations by comparing generated explanations to ground truth [ElSherief *et al.*, 2021; Sap *et al.*, 2020; Huang *et al.*, 2022], the human perception and implications of these generated explanations remain unclear. This paper aims to evaluate the effectiveness and limitations of the state-of-the-art language model GPT-3 in generating explanations for hate speech. We propose an analytical framework and conduct a comprehensive survey to assess the subjective and nuanced nature of hate speech explanations and answer the following research questions.

1. **RQ1**. To what extent do the explanations generated by GPT-3 for hateful content demonstrate qualities of being fluent, informative, persuasive, and logically sound?
2. **RQ2**. How persuasive are the GPT-3 explanations, and does the effectiveness of persuasion vary with different prompting strategies?
3. **RQ3**. Does the use of GPT-3 explanations lead to incorrect decisions in hateful content moderation?

The study utilizes the tweets annotated in the HateXplain dataset [Mathew *et al.*, 2021]. The study aims to prompt GPT-3 to explain hateful and non-hateful tweets and assess the generated explanations' quality through human evaluations. The evaluation metrics include fluency, informativeness, soundness, and persuasiveness. We conducted a survey to assess

Evaluated Tweet: "*maybe ban the nazis before u wanna worry about me tweeting something for the thousandth time*" [label: Non-hateful]

		Human Evaluator Ratings				
Explanation Type	**GPT3 Generated Explanation**	**F**	**I**	**S**	**P**	**H**
No explanation.	-	-	-	-	-	[2,2,4]
WHY (hateful)	This tweet is hateful because it downplays the severity of Nazi ideology and suggests that those who are concerned about it are overreacting... Nazi ideology is responsible for the deaths of millions of people, and it should be taken seriously.	4.0	3.0	2.7	2.0	[4,1,4]
WHY (non-hateful)	This tweet is non-hateful because it is not making any negative or derogatory statements about Nazis. It is simply making a suggestion that the government should focus on banning Nazis before worrying about what someone is tweeting.	4.0	3.3	3.3	4.0	[2,1,1]
COT (hateful)	1. The tweet starts with "maybe ban the nazis." This is clear that the person who wrote this tweet believes that Nazis should be banned. 2. This implies that the person who is worried about this tweet is more concerned with preventing the writer from tweeting than they are with banning Nazis. This tweet is hateful because it downplays the seriousness of Nazis. It also implies that the person who is worried about this tweet is more concerned with censorship than they are with stopping Nazis.	3.0	1.7	2.7	3.0	[3,4,1]
COT (non-hateful)	The tweet says "maybe ban the nazis before u wanna worry about me tweeting ..." This is a suggestion that the person thinks would be more productive than worrying about what the person tweets. The person is not making any hateful statements about Nazis. He simply suggests that Nazis should be banned. This is not a hateful statement.	4.0	4.3	4.0	3.7	[1,1,1]
CONTEXT	The tweet is in reference to the recent controversy surrounding Twitter's decision to ban certain users for their hateful and abusive speech. The user is suggesting that Twitter should ban Nazis before they worry about people who are tweeting the same thing over and over again.	4.3	4.0	3.7	3.7	[3,4,3]

Table 1: Sample of GPT3 generated explanations and evaluators' 5-point Likert scale rating for the evaluated tweet. The scores are broken down into five categories: **F** for fluency, **I** for informativeness, **S** for soundness, **P** for persuasiveness, and **H** for hatefulness. Three annotators assigned each category's ratings. Scores from three annotators were averaged except for hatefulness ratings.

the quality of explanations generated by GPT-3 and the hatefulness of tweets after viewing the explanations. The study addresses three key research questions, and we summarize the key findings as follows:

- Human evaluators have assessed that the explanations generated by GPT-3 are fluent, informative, persuasive, and logically sound.
- Different prompting strategies illicit varying persuasive effects. When prompted to explain why a given tweet is hateful, GPT-3 generated a more persuasive explanation than simply asking it to provide contextual information on the tweet. The length of the generated explanation also affects its persuasiveness.
- The potential for GPT-3 generated explanations to mislead human moderators when evaluating hateful content is a matter of concern. Our study has revealed that the explanations produced by GPT-3 can cause human evaluators to misclassify roughly 20% of tweets. Such misclassifications can wrongly label non-hateful tweets as hateful or vice versa. This is in contrast to a baseline scenario where human evaluators made assessments without any explanation.
- We observed that presenting both hateful and non-hateful explanations generated by GPT-3 could mitigate the risk of misleading content moderators.

2 Related Works

Content moderation and hate speech. In recent years, major social media platforms such as Facebook, YouTube, and Twitter have integrated AI into their operations for content moderation. The rise of user-generated content has made it increasingly necessary for these platforms to monitor and remove harmful or illegal content [Djuric *et al.*, 2015; Badjatiya *et al.*, 2017; Watanabe *et al.*, 2018; Awal *et al.*, 2021; Awal *et al.*, 2023; Meng *et al.*, 2022; Lin *et al.*, 2021]. As a result, governments have imposed strict regulations on social media platforms, requiring them to remove hateful content quickly. In response, many platforms have implemented automated systems, such as algorithms and AI, to proactively detect and remove such content at scale [Lampe and Resnick, 2004]. However, these systems have been criticized for labour concerns, lack of transparency, perpetuating biases, and potential harm to marginalized communities [Gillespie, 2018; Haimson *et al.*, 2021; Steiger *et al.*, 2021; Suzor *et al.*, 2019; Cao and Lee, 2020]. To address these

concerns, there is a growing emphasis on developing trustworthy and explainable AI systems for hate speech detection and moderation.

Explainable hate speech detection. Hate speech detection aims to identify and prevent online hate. Although several studies have been conducted on this topic [Davidson *et al.*, 2017; Cao *et al.*, 2020], hate speech models like the Google Jigsaw API are vulnerable to racial biases and counterfactual and adversarial attacks [Sap *et al.*, 2019; Park *et al.*, 2018]. To address these issues, explainable hate speech detection approaches have received attention, wherein model predictions are described through natural language explanations [Camburu *et al.*, 2018; Mathew *et al.*, 2021]. Recent studies, such as [Sap *et al.*, 2020] and [ElSherief *et al.*, 2021], have examined stereotypes in social media and the implicit nature of hate speech, respectively, providing insight into the underlying causes of hate speech and aiding in the development of model explanations. Furthermore, the generation of counter-narratives to combat hate speech has been explored [Tekiroglu *et al.*, 2020; Ashida and Komachi, 2022].

Human-generated explanations are limited and challenging to scale for real-world scenarios. To overcome this limitation, pre-trained language models are being utilized to automate the explanation generation process [Sap *et al.*, 2020; Huang *et al.*, 2022]. Furthermore, generative explanations can help address user appeals against platform flagging [Vaccaro *et al.*, 2020]. However, using GPT-3 based API for content moderation by companies like Cohere and OpenAI raises concerns about bias and spreading misinformation [Markov *et al.*, 2022]. Additionally, the quality of generated explanations requires further evaluation, as current studies primarily rely on automated quantitative analysis, such as the BLEU score. Therefore, our research aims to assess the credibility of generated explanations using a qualitative assessment conducted by human evaluators.

Prompting. Prompting is a technique that enables LLMs, such as GPT-3, to adapt to specific tasks by incorporating task-related instructions or questions into input text [Brown *et al.*, 2020; Sanh *et al.*, 2021; Wei *et al.*, 2022; Kojima *et al.*, 2022]. The format and order of the prompt and the demonstration examples, can affect the performance of LLMs [Zhang *et al.*, 2022; Zhao *et al.*, 2021]. Complex reasoning tasks can be improved by inducing a sequence of intermediate steps with few human-crafted demonstration examples [Wei *et al.*, 2022]. Adding the instruction "Let's think step by step" before each response could have the same effect as [Wei *et al.*, 2022] in zero-shot learning [Kojima *et al.*, 2022]. Prompting has shown promising results in various natural language processing (NLP) tasks, including teaching machines to generate explanations with demonstrations, which can be even more suitable than human-written explanations [Wiegreffe *et al.*, 2022; Petroni *et al.*, 2019].

Previous research has investigated using GPT-3 for addressing online hatred and abuse, including hate speech detection [Chiu *et al.*, 2021] and toxicity detection [Wang, 2022]. GPT-3 has the potential to generate explanations for its predictions, such as why it classified certain texts as hate speech. However, the generated content from GPT-3 may not always be accurate or factual, presenting potential risks in applications such as addressing online hatred and abuse. Generative models may not always critically evaluate their predictions, meaning they can make predictions but not provide reasoning behind them [Saunders *et al.*, 2022]. GPT-3's ability to generate human-like text can be misused to create convincing disinformation, such as fake news and propaganda. Its writing has been shown to significantly impact readers' perspectives on international affairs [Buchanan *et al.*, 2021]. It is crucial to weigh the potential benefits of using GPT-3 against the risks associated with its ability to generate convincing disinformation and influence people's perspectives, as well as its limitations in providing reasoning for its predictions.

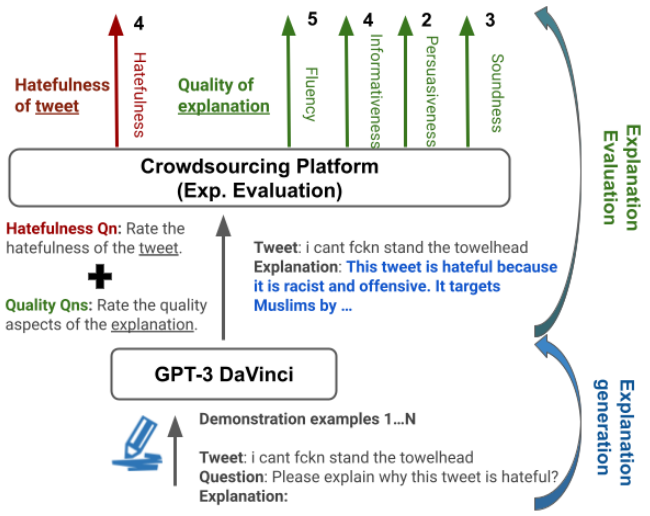

Figure 1: Overall framework of our study depicting the procedure of generating explanations utilizing GPT-3 and conducting a human evaluation. To prompt GPT-3, we selected demonstration examples from a candidate pool, producing one explanation for each tweet, based on each prompting strategy. Subsequently, human evaluators, via the mTurk and Clickworker platforms, assessed tweets' hatefulness and the explanations' quality. The outcomes of the human evaluation process were utilized to appraise the effectiveness of the explanations.

3 Methodology and Experimental Setup

This section outlines our methodology for determining the quality of explanations generated for hate speech using a representative dataset. The simplified process is as shown in Figure 1. The overall process includes sampling examples of hate speech, utilizing various prompting strategies to generate explanations, and designing a survey to evaluate the quality of the generated explanations. The survey is administered to gain deeper insights into the nature of hate speech and how to combat it effectively. The following sections detail our methods and techniques used at each stage of the process.

3.1 Dataset

The **HateXplain** dataset [Mathew *et al.*, 2021] comprises a set of hate speech instances acquired from Twitter and Gab. Each record in the dataset was annotated by three persons

who assigned one of the following three labels: "offensive," "normal," or "hateful." The final label for each record was determined via majority voting. Our experiment only considered Twitter tweets labelled "hateful" or "normal." To ensure a comprehensive analysis of hate speech, we incorporated all annotations associated with each tweet. Consequently, the tweets were categorized into four groups: [nnn, nnh, nhh, hhh], where "h" represents "hateful" and "n" represents "normal." A total of 25 tweets were randomly selected from each group, resulting in a 100-tweet evaluation set.

3.2 Prompting GPT-3 for Explanation Generation

We fed the GPT-3 model with the evaluated tweet, followed by three prompting strategies to generate the explanations:

- **Why is ...? (WHY)** We prompt the model with "*Please explain why this tweet is (hateful/non-hateful)*". The goal is to prompt GPT-3 to generate a hateful or non-hateful explanation for the tweet.
- **Chain-of-Thought (CoT)**: We prompt the model with the same question asked in the WHY strategy and prefix the answer "*Let's think step-by-step*.". The goal is to prompt GPT-3 to generate a more detailed hateful or non-hateful explanation with reasoning steps.
- **What is the context ...? (CONTEXT)**: We prompt the model with "*What is the context of the tweet?*". The goal is to encourage the GPT-3 model to generate an informative response that captures the relevant contextual details associated with the tweet without explicitly telling the model if the tweet is hateful or non-hateful.

We prompted GPT-3 with four demonstration examples for each strategy before asking about the specific tweet. These examples consist of tweets extracted from the HateXplain dataset, accompanied by explanations generated through zero-shot prompting using GPT-3 and carefully chosen for their fluency, soundness, informativeness, and persuasiveness. The Table 2 presents the overall prompt format.

Using the three prompting techniques, eight distinct combinations of tweets and explanations were generated for survey evaluation. These combinations involved pairing a tweet with either no explanation, a hateful explanation, a non-hateful explanation, a contextual explanation, or both hateful and non-hateful explanations side-by-side.

3.3 Evaluation Metrics

An extensive human evaluation was conducted to evaluate the hatefulness of the tweets and the quality of the explanations generated by GPT-3 based on four criteria: *fluency*, *informativeness*, *persuasiveness*, and *soundness*. A 5-point Likert scale was employed, where 1 represented the poorest quality and 5 the best. The definition of hatefulness and the quality criteria are as follows.

- **Hatefulness.** A *hateful* tweet is defined as: "Any speech that attacks a person or group on the basis of attributes such as race, religion, ethnicity, nationality, gender, disability, sexual orientation, or other identity factors." The rating scale spanned from 1 (non-hateful) to 5 (hateful), allowing annotators to express varying degrees of hate speech [Poletto *et al.*, 2019]. This approach offers a more detailed and flexible annotation process.
- **Fluency** evaluates whether the explanation follows proper grammar and structural rules, with a rating scale ranging from 1 (poor) to 5 (excellent).
- **Informativeness** assesses whether the explanation provides new information, such as explaining the background and additional context, with a rating scale ranging from 1 (not informative) to 5 (very informative).
- **Persuasiveness** evaluates whether the explanation seems convincing, with a rating scale ranging from 1 (not persuasive) to 5 (very persuasive).
- **Soundness** describes whether the explanation seems valid and logical, with a rating scale ranging from 1 (not sound) to 5 (very sound).

Demonstration samples 1...N **Tweet**: <user> <user> <user> <user> <user> what else we can expect from jihadi terrorist country **Question**: Please explain why this tweet is hateful? **Answer**: [Let's think step-by-step.] The tweet is hateful because it mocks Pakistan's status as a haven for terrorists. ### # Actual tweet in prompt **Tweet**: <ser> so did the nazis, **Question**: Please explain why this tweet is hateful? **Answer**: [Let's think step-by-step.] This tweet is hateful because it trivializes the Holocaust and the suffering of the Jewish people. It is also offensive and insensitive.

Table 2: An illustration of a prompt utilizing N training examples is presented, consisting of the tweet (in yellow), the question (in green), and the explanations provided (in purple) that are utilized as input to GPT-3.

3.4 Human Evaluation Setting

Human evaluations were carried out on both Amazon Mechanical Turk and Clickworker platforms. As the HateXplain dataset primarily comprises Twitter content generated in a American context, we recruited human evaluators residing in the United States. In total, we recruited 2,400 participants for the human evaluation. They were asked to evaluate the level of hatefulness in tweets and the quality of explanations generated by the GPT-3 model.

A three-round survey was executed to evaluate the explanations generated by GPT-3. Each round involved distinct participants and a variable number of questions pertaining to the same set of 100 sampled tweets. To ensure the validity of the survey and the respondent attentiveness, a basic math question is included in each round, e.g. "What is 8+3?".

Round 1. In the survey's first round, participants assessed tweet hatefulness without explanations. This controlled baseline aimed to determine if providing explanations impacts hatefulness perception in those tweets.

Round 2. In the second round of the survey, tweet-explanation pairs were presented, with each tweet paired with

		WHY			COT			CONTEXT
	Tweets label	*hateful*	*non-hateful*	*both*	*hateful*	*non-hateful*	*both*	
Fluency	not-hate	3.91	3.33	3.92	3.63	3.65	3.6	3.07
	hate	3.85	3.04	3.9	3.65	3.42	3.49	2.39
Informativeness	not-hate	3.65	3.21	3.39	3.55	3.65	3.51	2.93
	hate	3.75	2.59	3.38	3.51	3.39	3.35	1.95
Persuasiveness	not-hate	3.37	3.03	3.2	3.27	3.21	3.21	2.71
	hate	3.6	2.33	3.19	3.38	2.79	3.02	1.86
Soundness	not-hate	3.49	3.14	3.2	3.47	3.41	3.41	3.09
	hate	3.7	2.35	3.35	3.65	2.88	3.26	2.16

Table 3: Quality assessment of the WHY, COT, and CONTEXT prompting strategies for generating GPT-3 explanations for hateful and non-hateful tweets. It is worth noting that for every tweet, irrespective of its label, we utilized the WHY and COT strategies to prompt for three types of explanations: *hateful*, *non-hateful*, and *both*, where both hateful and non-hateful explanations were generated side-by-side.

either a hateful or a non-hateful explanation. The explanations were generated using the prompting strategies WHY and COT. Following this, participants were asked to evaluate the level of hatefulness in the tweets after reading the provided explanation and respond to four quality questions assessing the fluency, informativeness, persuasiveness, and soundness of the explanation. The primary objective of this round was to examine whether GPT-3 can generate high-quality and persuasive explanations (for both hate and non-hate labels) capable of influencing individuals' perceptions.

Round 3. The third round of the survey adopts two distinct formats. The first format combines hateful and non-hateful explanations generated by the WHY and COT prompting strategies. Participants were asked to respond to four quality questions of both explanations presented side-by-side. In the second format, the CONTEXT prompting strategy was utilized to elicit background information about the given tweet. Based on the contextual information, participants had to answer four quality questions assessing the explanations' quality. The primary objective of this format is to evaluate whether language models can generate high-quality and objective explanations (contextual) that reveal the tweet's implicit context without impacting readers' perceptions.

Participants who did not receive any explanations (i.e., Round 1) were given a survey consisting of two questions and were allotted 15 minutes to complete it. Conversely, participants who were presented with one explanation (i.e., Round 2) were given a survey consisting of five questions and given 30 minutes to complete it, while those who received two explanations (i.e., Round 3) were provided with a survey consisting of 10 questions and given 45 minutes to complete it. Three survey responses were collected for each data point, and only responses from evaluators who correctly answered the basic arithmetic questions were recorded in the analysis.

4 Experimental Results and Analysis

This section presents the human evaluation results and discusses our key findings in respect to the three research questions on hate speech explanation generated by GPT-3.

4.1 RQ1. Quality of Generated Explanations

Table 3 summarizes the quality assessment of the GPT-3 generated explanations using various prompting strategies. Specifically, we report the average *Fluency*, *Informativeness*, *Persuasiveness*, and *Soundness* scores provided by the human evaluators. These were rated on a scale of 1 to 5, with 5 being the best score.

We observe that within the same prompting strategy, the scores across four quality metrics for hateful explanations were higher than that for non-hateful explanations. This is because the HateXplain dataset used in our study focuses on hate speech and the tweets collected were more offensive in nature. Some tweets contain slurs such as "*niggas*", "*faggots*" or "*fuck*", but are marked as non-hateful because of its context. Consequently, GPT-3 faces a greater challenge in generating reasonable non-hateful explanations for such tweets, ultimately impacting human evaluators' assessments. We compared two strategies, WHY and COT, for generating non-hateful explanations and found that the latter significantly improved the quality score (with an average improvement of approximately +0.42 across all quality assessment metrics). The superior performance of the COT approach in generating non-hateful explanations suggests that breaking down the tweet analysis into steps to reach a conclusion is beneficial in generating non-hateful explanations, particularly in challenging contexts such as those presented by the HateXplain dataset. Conversely, we observe that the hateful explanations generated using WHY strategy scored better than those generated using COT, suggesting that the human evaluators prefer shorter and more direct hateful explanations over long and elaborate ones when evaluating generally offensive tweets.

Our findings indicate that the quality scores for the side-by-side hateful and non-hateful explanations (*both*) generated using WHY and COT were moderately effective. In contrast, CONTEXT was the least effective prompting strategy for generating high-quality explanations. Closer examination of the CONTEXT explanations revealed that many of them simply summarized the tweet content without providing significantly new information. In Section 4.3, we will delve deeper into the effects of applying the WHY-*Both*, COT-*Both*, and CONTEXT strategies to mitigate risks associated with using LLMs

		WHY			COT			CONTEXT
Tweets label	**No Exp.**	*hateful*	*non-hateful*	*both*	*hateful*	*non-hateful*	*both*	
non-hate	3.16	3.64(+0.48)**	2.73(−0.43)**	3.14(-0.02)	3.25(+0.09)	2.69(−0.47)**	3.13(-0.03)	2.82(−0.24)*
hate	3.96	3.94(-0.02)	3.57(−0.39)**	3.97(+0.01)	4.39(+0.43)**	3.81(-0.15)	4.11(+0.15)	4.05(+0.11)

Table 4: Average *hatefulness* scores of the tweets evaluated by human evaluators after viewing explanations generated by GPT-3 using different prompting strategies. The values in () represent the differences between the tweets' average *hatefulness* scores without viewing any explanations (i.e., No Exp.) and after viewing explanations. Red indicates an increase in *hatefulness*, while blue indicates a decrease in hatefulness. To determine the significance of the differences, we computed p-value, denoted by ** when $p \leqslant 0.01$ and * when $p \leqslant 0.05$.

for moderating hateful content.

Overall, our quality assessment analysis revealed that the explanations generated by GPT-3 scored well in terms of *fluency* and *informativeness*, but slightly worse in terms of *soundness* and *persuasiveness*. This finding is consistent with previous research [Wiegreffe *et al.*, 2022; Dou *et al.*, 2022], which suggests that generating a sound and persuasive explanation necessitates a deeper understanding of the tweets, making it a more challenging task than achieving fluency and informativeness. Furthermore, we observed that the GPT-3 model performed inadequately in generating explanations with a prompting label that contradicts the true label in the dataset. Consequently, hateful explanations received higher quality scores for hateful tweets than non-hateful tweets, while non-hateful explanations received higher scores for non-hateful tweets than hateful tweets.

4.2 RQ2. Persuasiveness of Generated Explanations

The earlier section has established that GPT-3 is able to generate high quality explanations for hateful tweets. However, little is known about the the influence of these generated explanations on human perception. This section will fill this gap by analyzing the persuasiveness of the generated explanations and their effects on the the tweets' *hatefulness* scores rated by the human evaluators.

Table 4 presents a summary of the average *hatefulness* scores of tweets rated by human evaluators when presented with the accompanying explanations generated by GPT-3 using various prompting strategies. We observe that the WHY-*non-hateful* explanations have significantly decreased the *hatefulness* scores of both hateful and non-hateful tweets. Conversely, we observe that the WHY-*hateful* explanations have significantly increased the *hatefulness* scores of non-hateful tweets, but no such effects are observed on the hateful tweets. One possible explanation for this finding is that the human evaluators had already rated the hateful tweets as hateful, and the generated explanation did not alter their assessment of these tweets.

Similar observations were made on the explanations generated using the COT strategy. For instance, the COT-*non-hateful* explanations have significantly decreased the *hatefulness* scores of non-hateful tweets, and the COT-*hateful* explanations have significantly increased the *hatefulness* scores of hateful tweets. Interestingly, we observe that the COT-*non-hateful* explanations did not significantly change the *hatefulness* scores of hateful tweets. This finding may be attributed to the elaborated explanations generated by the COT strategy,

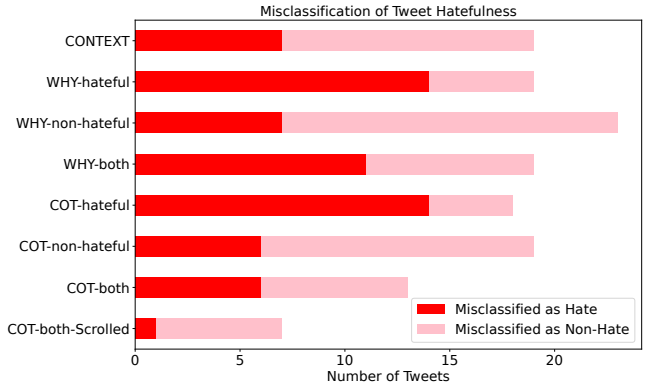

Figure 2: Distribution of tweets misclassified by human evaluators after reading prompting strategies' explanations.

which may have required the GPT-3 model to fabricate convoluted explanations in an attempt to convince human evaluators that a hateful tweet is actually non-hateful. However, as discussed in the previous section, such explanations may be deemed as unsound, thereby failing to alter the human evaluators' ratings of the tweet. Similar results were observed for the COT-*hateful* explanations on non-hateful tweets.

When presented with both hateful and non-hateful explanation side-by-side using the WHY-*both* and COT-*both* strategies, we observe that *hatefulness* scores have little or no apparent change compared to not providing any explanations. This suggests that when presented with a balanced explanation, i.e., both hateful and non-hateful, the human evaluators are less likely to be persuaded by either explanation and alter their opinions on a given tweet. Similarly, when presented with the CONTEXT explanations for hateful tweets, the change in *hatefulness* scores is insignificant. Interestingly, we noted that providing the CONTEXT explanations reduced the hatefulness scores for non-hateful tweet significantly.

4.3 RQ3. Risks of GPT-3 Generated Explanations for Hateful Content Moderation

In this analysis, we aim to examine the impact of explanations generated by GPT-3 on human decision-making in hateful content moderation. To achieve this, we first classify each annotation as either *non-hateful* or *hateful*, based on their rated hatefulness scores of 1-2 and 4-5, respectively. Each tweet has three annotations from three different human evaluators, and the majority vote is used as the annotated label for the tweets. We identified misclassifications by comparing

the labels assigned by human evaluators with and without seeing explanations. Specifically, we studied cases where human evaluators labeled tweets different from round 1, where there are no explanations, after reading GPT-3 generated explanations using various prompting strategies. Figure 2 displays the distributions of tweets that were misclassified by human evaluators after reading explanations generated by different prompting strategies.

Misleading Human Evaluators. Figure 2 shows the distribution of tweets misclassified by the human evaluators after reading the prompting strategies' explanation. From Figure 2, we observe that exposing the evaluators to WHY-*hateful* explanations increases the misclassification of non-hateful tweets, as they are persuaded to label them as hateful. Similarly, exposing the evaluators to WHY-*non-hateful* explanations also leads to an increase in misclassifying hateful tweets as non-hateful. The COT-*hateful* and COT-*non-hateful* explanations also misled the evaluators, resulting in misclassification of both non-hateful and hateful tweets. Surprisingly, even CONTEXT explanations have been found to mislead evaluators, resulting in misclassifications of both hateful and non-hateful tweets. Notably, a significant number of hateful tweets were mistakenly rated as non-hateful when presented with CONTEXT explanations.

Our hypothesis was that presenting both hateful and non-hateful explanations together would provide human evaluators with balanced information, aiding them in making better decisions regarding moderating hateful content. However, our observations show that even with WHY-*both* explanations, there is still a significant number of misclassifications. Conversely, our findings indicate that the COT-*both* approach is a more promising method, resulting in fewer misclassifications for both hateful and non-hateful tweets.

Effects of Explanation Length. Despite the ability of COT-*both* to provide detailed and balanced explanations to human evaluators, we found that the generated explanations are significantly longer than those generated by other methods. For example, COT-*both* explanations have an average length of 173 words, compared to CONTEXT and WHY-*both* explanations with an average length of 28 and 75 words, respectively. Additionally, our research revealed that when human evaluators used a mobile device to respond to the survey, some COT-*both* explanations extended beyond a single page, requiring the evaluator to scroll to view entire content. To address this issue, we introduced a "scrolled" variable in our survey to verify whether evaluators had scrolled to read the full COT-*both* explanations when rating the hatefulness of the tweets. We next analyzed the scrolling behavior of the 300 evaluators who assessed tweets paired with COT-*both* explanations. We discovered that 69 out of the 300 evaluators did not scroll to read the complete COT-*both* explanations while rating the hatefulness of the tweets. This raises concerns about the reliability of these evaluations, and we therefore excluded these 69 ratings from our analysis. The resulting misclassification of tweets is reported in Figure 2 as COT-*both Scrolled*. The figure shows that the number of misclassifications decreased even further after removing the evaluations of human raters who did not scroll to read the full COT-*both* explanations. Misclassifications may occur when individuals focus solely on the first hateful explanations. This selective exposure raises the risk of mistakenly classifying non-hate tweets as hate and the drawback of lengthy explanations is also evident. Hence, future research will focus on exploring techniques that generate shorter and more concise explanations for effective moderation of hateful content.

5 Discussion

To the best of our knowledge, this study represents the first evaluation of the quality and ethical implications of PLM-generated explanations for moderating hateful content on a large scale. Our findings demonstrate that GPT-3 can produce fluent, informative, persuasive, and logically sound explanations when properly prompted. These results suggest that GPT-3 may be a valuable tool for combating online hate speech and content moderation. However, we discovered that the persuasiveness of the explanations varied depending on the prompting strategy and the length of the explanations. In particular, we found that biased explanations, regardless of whether they were hateful or non-hateful, had a strong persuasive effect on human annotators. Therefore, the use of GPT-3 to produce explanations for content moderation should be approached with caution. Moreover, our observations revealed that the explanations generated by GPT-3 had the potential to lead to incorrect labeling of tweets. This is a critical issue that must be addressed to ensure the effectiveness and ethicality of using GPT-3 to aid content moderation. We recommend presenting both hateful and non-hateful explanations side-by-side, which can significantly reduce the risk of misleading content moderators.

Limitations. The study analyzed the Hatexplain dataset, which specifically addresses hate speech on Twitter and Gab, which may limit the generalizability of the study to other platforms. Moreover, the Hatexplain dataset, like other hate speech detection datasets, is susceptible to subjective hate speech annotation, leading to varying definitions among annotators and potentially inaccurate annotations. Additionally, the study used non-expert human annotators to evaluate the GPT-3 model's explanations, with precautions taken for accuracy, but some anomalies may still exist. Future research could involve experienced human annotators to obtain higher-quality annotations. The study solely concentrated on the GPT-3 model, necessitating further investigation into whether similar results can be achieved with other language models.

6 Conclusion

This study introduced an analytical framework for evaluating hate speech explanations and conducted a comprehensive survey on their effectiveness. The research identified the potential of GPT-3 in generating high-quality explanations for hate speech while also highlighting the limitations and risks of using PLM-generated explanations for content moderation. For future work, we aim to develop improved evaluation metrics and prompting strategies to enhance the quality and reduce bias in explanations. These efforts would help increase the fairness and effectiveness of content moderation and combat online hate speech.

References

[Ashida and Komachi, 2022] Mana Ashida and Mamoru Komachi. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics (ACL).

[Awal *et al.*, 2021] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. Angrybert: Joint learning target and emotion for hate speech detection. *arXiv preprint arXiv:2103.11800*, 2021.

[Awal *et al.*, 2023] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, 2023.

[Badjatiya *et al.*, 2017] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. *In Proceedings of the 26th international conference on World Wide Web companion*, April 2017.

[Brown *et al.*, 2020] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and Neelakantan et al. Language models are Few-Shot learners. *Advances in neural information processing systems*, May 2020.

[Buchanan *et al.*, 2021] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. Truth, lies, and automation. *Center for Security and Emerging Technology*, 1(1), 2021.

[Camburu *et al.*, 2018] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

[Cao and Lee, 2020] Rui Cao and Roy Ka-Wei Lee. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, 2020.

[Cao *et al.*, 2020] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. DeepHate: Hate speech detection via multifaceted text representations. *In 12th ACM conference on web science*, March 2020.

[Chiu *et al.*, 2021] Ke-Li Chiu, Annie Collins, and Rohan Alexander. Detecting hate speech with GPT-3. *arXiv:2103.12407*, March 2021.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *In Proceedings of the international AAAI conference on web and social media*, March 2017.

[Djuric *et al.*, 2015] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, New York, NY, USA, May 2015. Association for Computing Machinery (ACM).

[Dou *et al.*, 2022] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah Smith, and Yejin Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics (ACL).

[ElSherief *et al.*, 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, September 2021.

[Fortuna and Nunes, 2018] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):1–30, July 2018.

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

[Haimson *et al.*, 2021] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):1–35, October 2021.

[Huang *et al.*, 2022] Fan Huang, Haewoon Kwak, and Jisun An. Chain of explanation: New prompting method to generate higher quality natural language explanation for implicit hate speech. *arXiv preprint arXiv:2209.04889*, September 2022.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are Zero-Shot reasoners. *arXiv preprint arXiv:2205.11916*, May 2022.

[Lampe and Resnick, 2004] Cliff Lampe and Paul Resnick. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550. 2004.

[Lin *et al.*, 2021] Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 967–974. Institute of Electrical and Electronics Engineers (IEEE), 2021.

[Markov *et al.*, 2022] Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274*, August 2022.

[Mathew *et al.*, 2021] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. *Proc. Conf. AAAI Artif. Intell.*, 35(17):14867–14875, May 2021.

[Meng *et al.*, 2022] Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. Predicting hate intensity of twitter conversation threads. *arXiv preprint arXiv:2206.08406*, 2022.

[Park *et al.*, 2018] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, August 2018.

[Petroni *et al.*, 2019] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, September 2019.

[Poletto *et al.*, 2019] Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Maria Stranisci. Annotating hate speech: Three schemes at comparison. In *CEUR WORKSHOP PROCEEDINGS*, volume 2481, pages 1–8. CEUR-WS, 2019.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.

[Sanh *et al.*, 2021] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, and Chaffin et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, October 2021.

[Sap *et al.*, 2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678. 2019.

[Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics (ACL).

[Saunders *et al.*, 2022] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, June 2022.

[Schmidt and Wiegand, 2017] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics (ACL).

[Steiger *et al.*, 2021] Timir J Steiger, Miriah, Sukrit Bharucha, Martin J Venkatagiri, and Matthew Riedl. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14. 2021.

[Suzor *et al.*, 2019] Nicolas P Suzor, Andrew West, and Jillian Quodling. What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation. *International Journal of Communication Systems (IJCS)*, 13, 2019.

[Tekiroglu *et al.*, 2020] Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*, April 2020.

[Vaccaro *et al.*, 2020] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. At the end of the day facebook does what ItWants“ how users experience contesting algorithmic content moderation. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–22, 2020.

[Wang, 2022] Yingshan Wang, Yau-Shian. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*, 2022.

[Watanabe *et al.*, 2018] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835, 2018.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[Wiegreffe *et al.*, 2022] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics (ACL).

[Zhang *et al.*, 2022] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, October 2022.

[Zhao *et al.*, 2021] Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *In International Conference on Machine Learning*, February 2021.