To Appear in the 45th IEEE Symposium on Security and Privacy, May 20-23, 2024.

# Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models

Nishant Vishwamitra*¶, Keyan Guo†¶, Farhan Tajwar Romit*, Isabelle Ondracek†,
Long Cheng‡, Ziming Zhao†, Hongxin Hu†
*University of Texas at San Antonio, †University at Buffalo, ‡Clemson University
{nishant.vishwamitra, farhantajwar.romit}@utsa.edu, lcheng2@clemson.edu
{keyanguo, ikondrac, zimingzh, hongxinh}@buffalo.edu



***Abstract*—Online hate is an escalating problem that negatively impacts the lives of Internet users, and is also subject to rapid changes due to evolving events, resulting in new waves of online hate that pose a critical threat. Detecting and mitigating these new waves present two key challenges: it demands reasoning-based complex decision-making to determine the presence of hateful content, and the limited availability of training samples hinders updating the detection model. To address this critical issue, we present a novel framework called HATEGUARD for effectively moderating new waves of online hate. HATEGUARD employs a reasoning-based approach that leverages the recently introduced chain-of-thought (CoT) prompting technique, harnessing the capabilities of large language models (LLMs). HATEGUARD further achieves prompt-based zero-shot detection by automatically generating and updating detection prompts with new derogatory terms and targets in new wave samples to effectively address new waves of online hate. To demonstrate the effectiveness of our approach, we compile a new dataset consisting of tweets related to three recently witnessed new waves: the 2022 Russian invasion of Ukraine, the 2021 insurrection of the US Capitol, and the COVID-19 pandemic. Our studies reveal crucial longitudinal patterns in these new waves concerning the evolution of events and the pressing need for techniques to rapidly update existing moderation tools to counteract them. Comparative evaluations against state-of-the-art approaches illustrate the superiority of our framework, showcasing a substantial 10.59% to 88% improvement in detecting the three new waves of online hate. Our work highlights the severe threat posed by the emergence of new waves of online hate and represents a paradigm shift in addressing this threat practically.**

**Disclaimer**. This manuscript contains harmful content, including hate speech, which has the potential to be offensive and may disturb readers.

## 1. Introduction

We live in a world with rapidly evolving events. These rapidly evolving events consequently affect the global digital landscape [1], especially Internet platforms that enable online discourse, such as Online Social Networks (OSNs). As a result, emotions of anger and anxiety, and rhetoric from these events also spill over into our global digital landscape. For example, recent polarizing events, such as the 2022 Russian invasion of Ukraine [2], the 2021 insurrection of the US Capitol [3], and the COVID-19 pandemic [4], rapidly transformed the online discourse in our cyberspaces [5]. As a consequence, the context of online hate has also rapidly changed, leading to the emergence of *new waves* of online hate. For example, during the 2021 insurrection of the US Capitol, a wave of hateful content against vulnerable groups, such as LGBTQ and minorities, was witnessed [3], [6]. During the COVID-19 pandemic, a rapid rise in online hate against Asian-Americans [4], [7], mask [8], [9], and vaccine mandates [10], [11] was reported on several OSNs. More recently, during the 2022 Russian invasion of Ukraine, we saw yet newer waves of online hate against citizens of the nations involved in the conflict [12], [13]. New waves of online hate are a crucial issue that demands immediate attention not only for the present but also for the future well-being of our digital spaces. As new waves are likely to arise in the future, it becomes imperative to address this problem proactively.

Currently, various existing tools, such as Perspective API [14], Azure Text Moderation [15], and IBM Toxic Comment Classifier [16], utilize artificial intelligence and machine learning (AI/ML) models for moderating violations of online hate policies [17], [18]. However, there is a concern regarding the effectiveness of these tools in preventing violations caused by new waves of online hate. For example, the recent Anti-Asian hate [4], [7], mask-related hate [8], [9], and vaccine-related hate [10], [11] witnessed during the COVID-19 pandemic could not be sufficiently detected by these tools, and online hate against minority communities and other vulnerable groups continued to spread unabated during this period.

A major limitation of these existing tools, which hinders their effectiveness against new waves of online hate, is their reliance on traditional AI/ML models. This poses two key challenges. *First*, the detection of new waves of online hate poses a complex decision-making challenge, significantly different from the traditional classification tasks typically addressed by AI/ML models. Online hate is inher-

¶These authors contributed equally to this work.

ently "highly subjective, ambiguous, and context-dependent, making it difficult for both humans and computers to detect" [19], and the emergence of new waves exacerbates these difficulties. For instance, during the COVID-19 pandemic, a new wave of online hate targeted emerging political identities like "antimaskers", employed novel disparaging terms like "maskhole", and utilized different stereotypes to target specific communities. Determining whether such content is hateful or not demands intricate decision-making that necessitates reasoning. This process involves exploring multiple possibilities, including carefully discerning between expressions of hateful speech, mere criticism, and ironic statements [20], [21]. *Second*, due to the abrupt occurrence of new waves, only a limited number of samples are accessible for model updates. Therefore, tools designed to detect and moderate this issue should possess the capability for rapid deployment and adaptation, utilizing either minimal or no samples of new waves. Nevertheless, existing tools face challenges in promptly adjusting to the sudden surge of a new wave, as they lack a sufficient number of training samples. Additionally, the training paradigm employed by these tools necessitates the collection of a large dataset, followed by manual labeling through human annotators, a process that typically takes months and is not practically feasible for the timely discovery and moderation of new waves of online hate [22].

In this work, we embark on addressing the practical challenges presented by new waves of online hate. To begin, we analyze several recently emerged new waves, specifically, the 2022 Russian invasion of Ukraine, the 2021 US Capitol insurrection, and the COVID-19 pandemic. To support our research, we gather a novel dataset, containing 31,549 tweets related to these three categories of new waves. We present two systematic studies examining the nature of these new waves and the necessity for novel methods to update existing moderation tools. The first study focuses on tracking the usage of hateful hashtags associated with the three new waves in our dataset, revealing significant longitudinal patterns that can be leveraged for rapid detection. Subsequently, we explore the effectiveness of techniques employed to update existing online hate moderation tools against multiple new waves of online hate. Our findings reveal that these techniques fail to address the challenges presented by the emergence of new waves.

Based on these findings, we design HATEGUARD[1], a novel framework for the discovery and moderation of new waves of online hate. HATEGUARD introduces a reasoning-based approach, capitalizing on the recent innovation of chain-of-thought (CoT) prompting [23], enabling large language models (LLMs) to undertake the complex decision-making task of identifying whether new content is hateful or not. Additionally, HATEGUARD employs an automatic strategy to generate and update prompts for zero-shot classification by only updating the newly identified hate targets and derogatory terms in the prompts rather than the model. Our approach tackles the challenge of detecting new waves of hate by carefully crafting chains of automatic reasoning through LLMs, probing, and exploring various possibilities within the content. This methodology proves more suited to the intricate decision-making requirements of detecting new waves compared to traditional binary classification approaches. Moreover, our zero-shot approach facilitates updates solely to the prompts while leaving the model untrained.

The key contributions of this paper are as follows:

- **New dataset of new waves of online hate.** To study and understand the nature of the new waves of online hate, and to demonstrate the effectiveness of our approach, we collect a new dataset of 31,549 tweets about three recent new waves of online hate: the 2022 Russian invasion of Ukraine, the 2021 US Capitol insurrection, and the COVID-19 pandemic.
- **New understanding about new waves of online hate.** We report two systematic studies on the nature of new waves of online hate and the need for techniques to rapidly update existing online hate moderation tools against the new waves. Our studies shed light on longitudinal patterns regarding the sharp rise, peak, and dissipation of these new waves with evolving events that can be leveraged to rapidly detect such new waves, and highlight the need for methods to quickly update existing moderation tools.
- **New framework for the moderation of new waves of online hate.** We design a novel framework called HATEGUARD for effectively moderating new waves of online hate. HATEGUARD incorporates a CoT reasoning approach, empowering LLMs with reasoning capabilities to determine whether new content exhibits hateful characteristics, and an automatic prompt generation and update strategy for zero-shot classification, which streamlines the update process by solely focusing on updating the prompts rather than the model. Our framework takes a first step towards practically moderating new waves of online hate, by harnessing the potency of LLMs.
- **Multi-faceted and extensive evaluation of HATEGUARD.** We showcase HATEGUARD's capability to enhance flagging of different types of new waves, achieving an impressive improvement ranging from 6.52% to 71.93% compared to baselines in the last quarter of these new waves. Additionally, we compare our framework against state-of-the-art models and demonstrate its superiority, achieving 10.59% to 88% higher accuracy than these models. We also apply our framework in a real-world scenario, where it effectively identifies and flags *all* the hateful samples within a dataset of in-the-wild samples.

## 2. Background and Related Work

In recent times, online hate has emerged as a critical threat [24] that has been the focus of both governments [25]

[1]Our code and datasets are available at https://github.com/CactiLab/HateGuard.

and institutions [17], [18]. It has been reported that in 2017, 41% of Americans reported personally experiencing varying degrees of harassment online [26], and 40% of users reported similar experiences globally [27]. Furthermore, new vectors of online hate [28] evolve fast, thus evading existing detection systems. Understanding the need to address this growing threat, concerned persons from both academia [29] and industry [30], [31] have made efforts to defend against this threat. Initially, methods that involve human moderators have been proposed [32], [33], although the practicality of scaling these methods is questionable, and these methods are also not ethical [34]. AI/ML has since emerged as a critical technology that is being explored to practically address this threat. Recent studies have employed AI/ML techniques to develop classifiers capable of moderating online hate speech [35], [36]. However, these approaches cannot be used for the discovery and moderation of new waves of online hate.

The new waves of online hate can be considered as a specific case of concept drift [37] in the domain of online hate. Concept drift is defined as the "changes in the hidden context that can induce more or less radical changes in the target concept" [38]. The concept drift problem is critical in AI/ML since changes in the target's statistical properties can render a model less effective or even useless [39]. Although the presence of the concept drift problem has been discussed in the context of online hate [40], methods that specifically address it in the online hate domain have not been yet developed. Our framework offers a potential solution for addressing the issue of concept drift in online hate.

A critical issue with online hate detection is that it is a complex decision-making problem [19], [20], [21]. While ML algorithms perform remarkably well on classification tasks, they are not well suited for decision-making tasks that demand reasoning [41]. Recent research [19] indicates that online hate is highly contextual and poses a significant challenge for moderation, even for humans. Facebook has acknowledged that human moderators are essential and always have the final say when determining if flagged posts should be removed for hate content [42]. New waves of online hate compound this issue. Since there are significant differences between the semantics of new waves, such as the use of new derogatory terms and new targets of hate, hate speech detection models' performance further deteriorates when faced with new waves of hate [4]. As a result, there is a need for reasoning-based approaches to identify new hateful content, involving decision-making to determine whether the content is hateful or not. Recently, in-context learning (ICL) [23], [43] has emerged as a method to learn a new task from a small set of examples presented within the context (the prompt) at inference time. ICL enables pre-trained LLMs to address new tasks without the need for fine-tuning. Especially, The adoption of chain-of-thought (CoT) prompting in LLMs [23], as an ICL technique, has given them reasoning capabilities, opening up a new era in decision-making AI. While a recent study has applied such a technique for explaining AI decisions [44], the specific challenges associated with using the CoT approach for moderating online hate remain unexplored.

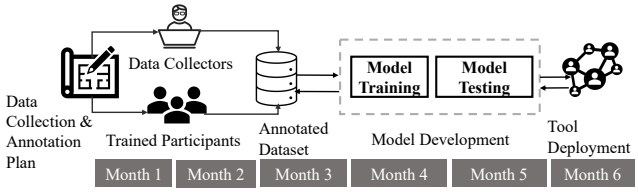



Figure 1: Conventional AI training approach.

Furthermore, existing approaches against new waves follow a conventional AI training approach that is illustrated in Figure 1, which is not practical for the real-world discovery of new waves, and cannot support quick deployment, because this process is quite time-consuming [22]. In the conventional approach, first, a data collection and annotation plan is designed that describes what kind of posts should be deemed as hateful [29], [35]. Then, the data collection is done wherein such posts are collected via social media APIs [35], [45]. Next, humans are trained to perform annotation tasks on the collected dataset, and after this process, we get an annotated dataset that can be used to train AI models. In the next step, the AI model is trained on the annotated dataset and the performance is evaluated on a test dataset. If the performance is satisfactory, it is deployed on Internet platforms for discovery and moderation. However, a major issue with this approach is that it takes *months* to complete this process [22], which makes it impractical to address the problem of new waves of online hate that need rapid updating of the model.

Few-shot and zero-shot learning (FSL and ZSL) [46] have recently emerged as a way of developing AI models using a few or no data samples. However, these techniques have been predominantly applied for images, such as adding a new category of an object to an existing dataset to enhance a model's detection capability. Recently, the use of FSL has been explored in text-based applications [47], and some emergent studies have explored FSL for detecting hate speech in less common languages [48], [49] and task decomposition [50]. For example, [48], [49] discuss approaches to detect hate speech in rare languages using transformer-based models in a few-shot setting (*i.e.*, simulating a small number of samples). However, they don't provide specific approaches for FSL to address online hate. Recent studies have shown that LLMs can even outperform humans in zero-shot tasks [51], and in this work, we explore the zero-shot capability of LLMs in moderating new waves of online hate.

## 3. Threat Model

In this work, we address the behavior of adversaries who spread hate online, especially those who target individuals or groups based on their identity, often in reaction to evolving events. Both the adversary and the target are considered as online users. The adversary can create hateful posts using various constructs, including words and hashtags relevant to the evolving events. We focus solely on textual media and

do not consider other means of disseminating such posts, such as images, upvotes, or likes. Our study addresses posts targeted specifically against a particular user, as well as posts intended for a wider audience, such as public posts. We make no assumptions about the adversary possessing any special capabilities or employing adversarial techniques to deceive content moderation tools.

## 4. Examining New Waves of Online Hate

In this section, we present studies on the nature of the new waves of online hate considering three recent new waves. In the first study, we investigated how new waves of online hate emerged with the changes in the global digital landscape. In the second study, we examined the need to quickly update existing tools used for online hate discovery and moderation by measuring their moderation capability on new waves of online hate. Our main objectives in conducting these two studies are to find out if there are patterns in the nature of the new waves that could be utilized for their detection and motivate the need for new strategies to quickly enable existing approaches to handle new waves.

### 4.1. Data Collection and Annotation

Our data collection and annotation tasks were approved by our Institution's IRB. We carried out two dataset tasks, collection and annotation. To collect tweets related to the three recent new waves, we compiled a seed set of hashtags that were prevalent during the time these waves [52], [53], [54] were active, which consisted of diverse hashtags such as *#ChinaVirus*, *#WuhanFlu*, *#WearAMask*, *#boomerremover*, *#COVID19Vaccine*, *#MAGAMorons*, and *#F**Putin*. We expanded this set by adding new hashtags from the collected tweets until no new ones were found, ensuring a representative sample. The full list of hashtags has been provided in Appendix A.

**Collection of Tweets.** We used the official X (previously Twitter) Streaming API[2] to collect tweets during the period from December 1, 2019, to December 31, 2022, based on the hashtags. Particularly, we collected COVID-19-related tweets from December 1, 2019, to December 31, 2020, US Capitol insurrection-related tweets from November 1, 2020, to December 31, 2021, and tweets regarding the Russian invasion of Ukraine from November 1, 2021, to December 31, 2022. In total, we obtained 507 million tweets published by 38 million users. We removed tweets with only hashtags, mentions, or links in the text part, and removed non-English tweets and retweets. We were left with 31,549 tweets. Additional samples from our dataset can be viewed in Appendix C.

**Annotation of Tweets.** Two authors of this work developed a code book for labeling the samples in our collection as hate speech or not and verified it after three rounds of annotations and resolving conflicts on random samples of the dataset. The two authors independently annotated 300 random samples from our dataset in each round, followed by agreement computations and conflict resolution discussions. This process was repeated with different random samples. By the third round, the two authors achieved 100% agreement. To develop the code book, we focused on identity-based hate and hate against individuals, defined as "Hatred, hostility, or violence towards member(s) of a race, ethnicity, nation, religion, gender, gender identity, sexual orientation or any other designated sector of society" [41]. To ensure the accuracy of the annotations, we meticulously cleaned each tweet by removing URLs, mentions, and non-English characters. Additionally, we removed stacked hashtags as well as those at the beginning and end of a tweet, unless they are linked to action words or determiners like "a","an", or "the" [55]. In our code book (detailed fully in Appendix B, Table 7), our analysis began with identifying if an individual or group identity is mentioned in the text. We then assessed for any derogatory or disparaging language, followed by determining if such language was directed at the mentioned individual or identity. We used Amazon Mechanical Turk (AMT) to label a random sample of 4,000 tweets using online participants and directed workers to label a text as hate if such words were directed at the identities or individuals mentioned. We labeled a subset of our dataset since we found that our task is quite intensive since it needs a lot of human reasoning and time, which was not practical to extend to the entire dataset. Furthermore, we sampled this subset with a temporal distribution, *i.e.*, we proportionally sampled an equivalent number of tweets from each quarter. Overall, we sampled 928 tweets from Q1, 893 tweets from Q2, 1,148 tweets from Q3 and 1,031 tweets from Q4. To maximize reliable annotation, we only recruited participants with an approval rating of 90% or higher and 1,000 approved HITs to participate in our annotation task. The AMT workers on average labeled 32 tweets. Since one of the new waves, the US Capitol insurrection, is linked to US politics, we have chosen to limit the geographical scope of our study to the US and Canada. After the annotation task was completed, the two expert annotators and developers of the code book who are well-trained in using it to label the samples verified the labels and corrected any coding errors made by the workers. The experts achieved a Fleiss Kappa agreement of 0.84, which indicates a near-perfect agreement. Table 1 shows the results of our tweets collection task. Of the 4,000 tweets labeled, 1,647 were labeled as hate and 2,353 as non-hate. Additionally, we were left with a large dataset of the rest of the 27,549 tweets to support large-scale analysis.

| New Wave Type | Number of hateful tweets | Number of non hateful tweets |
|---|---|---|
| COVID-19 tweets | 1,096 | 1,600 |
| US Capitol Insurrection tweets | 314 | 390 |
| Russian Invasion of Ukraine tweets | 237 | 363 |
| Total tweets | 1,647 | 2,353 |

TABLE 1: Annotated new wave dataset with 4,000 tweets.

[2]https://developer.twitter.com/en/docs/twitter-api

### 4.2. Nature of New Waves of Online Hate

To understand the nature of online hate, we investigated the text-based tweets by studying the temporal usage pattern of the extremely hateful X hashtags (such as #WuhanFlu, #boomerremover, #F**Putin, #MAGAMorons, etc.) about the three new waves in our dataset. Furthermore, during the COVID-19 pandemic, several distinct new waves of online hate emerged. Our study specifically examined four categories prevalent in social media during this time [4], [56]: Anti-Asian, Ageism, Mask, and Vaccine. We specifically focused on a "wave" of these new categories of hate, *i.e.*, a sharp, sudden, or unprecedented increase of these tweets, and what the antecedents of this increase could be. Our observations are depicted in Figure 2 and Table 2. We utilized the pruned exact linear time (PELT) algorithm [57] to find the change points in Figure 2. Two authors then correlated these points with real-world events in Table 2. Figure 2 illustrates the temporal usage of hateful hashtags in each category, along with the corresponding dates of the events, and Table 2 displays the current events associated with each category. We observed that for each of the categories, there are three stages, *buildup*, *peak*, and *decline*, in the temporal usage (Figure 2), which are closely related to the current events.

**Buildup.** This is the stage when certain current events related to a category are building up emotions of hate, anger, and anxiety in social media. We observed that the posts in this stage were responsible for building an outbreak of hate at a later stage. In particular, awareness about a certain event also plays a major part in the buildup stage. For example, we observed that negative emotions in the Asian community were being built up due to certain events such as the US CDC screening people traveling from China [58], and the WHO issuing a Global Health Emergency [58], which added to the stress and strain due to imposed lock-downs. Events that built up negative emotions against the older individual were observed in this stage, such as reports across the world about the pandemic disproportionately affecting older individuals and subsequent tweets that used particularly offensive terms such as "BoomerRemover" for COVID-19 [59]. The imposition of mask mandates across various institutions and public spaces [60], coupled with the CDC's recommendations for mask-wearing, contributed to an increase in social media discussions and opinions about mask usage during this phase. In the case of vaccine-related hateful hashtags, certain events, such as vaccine companies starting human trials of vaccines led to a buildup of emotions regarding the use of vaccines [61]. We observed the buildup of such emotions in the case of US Capitol insurrection-related hate, wherein election day played a polarizing role [62]. Furthermore, events such as the day Russia invaded Ukraine and the siege of Mariupol [63] (an event of significance during the invasion) witnessed a buildup of strong emotions. The online activities in the *buildup* stage are crucial since they are precursors for online hate. Suitable counter-actions in this stage are thus necessary to prevent an outbreak of online hate.

**Peak.** In the following stage, an outbreak of usage of hateful hashtags was observed after the negative emotions built up in the previous stage led to a peak of the new waves of online hate. This stage depicted an uncontrolled and overwhelming usage of hateful hashtags and an apparent failure of OSNs in countering hateful activity. For example, a peak of Anti-Asian hate was observed between February 2020 and March 2020. In the case of Ageism, the peak was noted between March 2020 and April 2020. In the case of mask-related online hate, the peak was observed between June 2020 and October 2020. The peak for a vaccine-related wave of online hate was observed in May 2020. Lastly, the peak of US Capitol insurrection-related hate was observed in January 2021, and the peak of Russian invasion-related online hate was observed in September 2022. Pragmatic steps, especially in the preceding stage must be taken to avoid the *peak* stage of a new wave of online hate.

**Decline.** During this stage, there was a noticeable decrease in the use of hateful hashtags related to new waves of online hate on Twitter following their *peak* stage. This decline could occur due to OSNs becoming aware of the new wave of online hate and taking measures, such as content moderation using human moderators [34]. But by the time this stage is reached, a large number of posts had already been shared on X in all the three types of new waves considered, including the four categories of COVID-19-related hate. Besides, in some categories, such as Anti-Asian, we observed that the *decline* stage was much more gradual and prolonged than other categories (*e.g.*, Ageism, Vaccine, and Russian invasion), showing sustained publishing of tweets that used hateful hashtags despite content moderation efforts.

**Finding.** Our study reveals that real-world events can trigger new waves of online hate, leading to swift changes in hate speech dynamics. Our experiment has identified significant longitudinal patterns that can help address new waves of online hate. These new waves typically involve a *buildup* of negative emotions resulting from evolving events, followed by a *peak* stage where the outburst of the new wave is encountered, and then a *decline* stage where the new wave reduces. We are developing a framework that can be updated based on only a few samples during the *buildup* stage of a new wave, which then counters the *peak* to significantly reduce its harmful effects. By doing so, our approach offers practical moderation of new waves of online hate.

| | |
|---|---|
| 1 | COVID-19 referred to as "BoomerRemover" first time on Twitter |
| 2 | WHO informs of cases of unexplained pneumonia in Wuhan, China |
| 3 | US CDC starts screening people from China |
| 4 | WHO issues Global Health Emergency |
| 5 | Mask mandates put in place across various institutions |
| 6 | CDC recommends wearing masks |
| 7 | Companies start first vaccine trials |
| 8 | US Elections |
| 9 | Russia invades Ukraine |
| 10 | Russia seizes Mariupol |

TABLE 2: Events that engendered new waves of online hate.

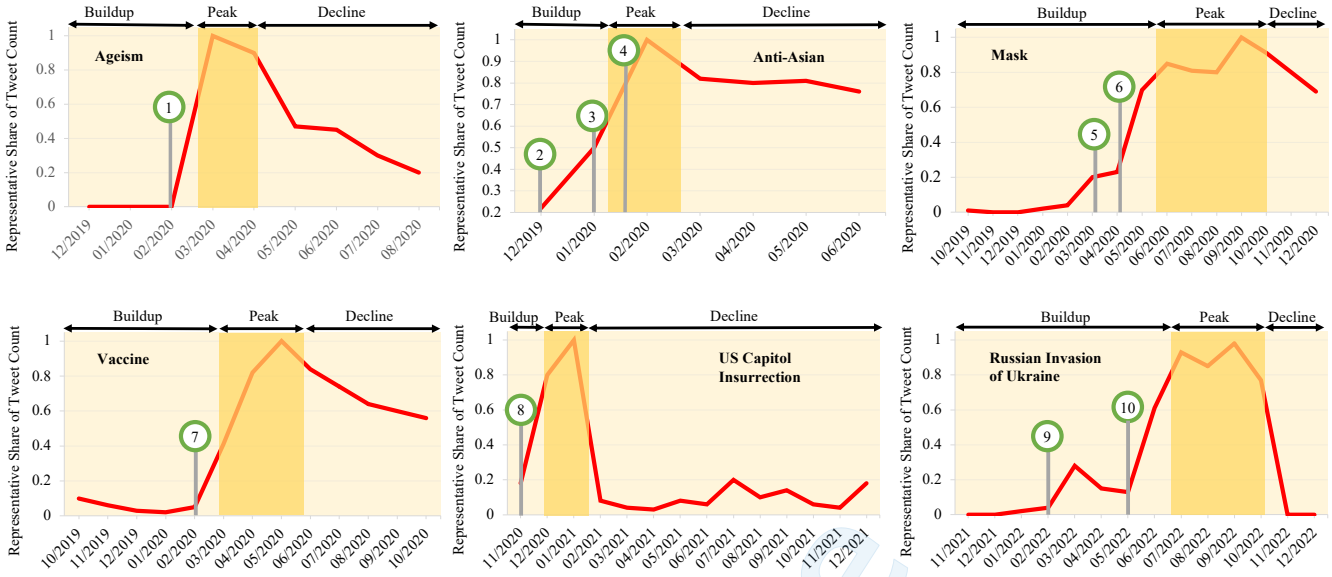



Figure 2: As current events evolve, new waves of online hate occur in the global digital landscape.

### 4.3. Using Existing Tools Against New Waves of Online Hate

Following our previous study, we wanted to investigate the need for new methods to extend the capabilities of the existing moderation tools to new waves of online hate. Our objective is not to point out that these tools are not effective against new waves, but to motivate the need to quickly update these tools. Specifically, we wanted to motivate the need for methods that can address the issue of rapid concept drifts in online hate, and study the limitations of existing methods, such as fine-tuning, which is a popular method to update such tools. Although these models are proprietary black-box, it is quite likely that they are trained with fine-tuning-based strategies. For example, Perspective API uses a multilingual BERT, which primarily uses fine-tuning that is "frequently retrained to make improvements and keep them up-to-date" [64]. Perspective API [14] needs a dataset of at least 20,000 samples of a particular label to be considered enough to re-train a model [65]. Similarly, IBM Toxic Comment Classifier [16] is based on fine-tuning a BERT-base uncased [66] model, and likely uses the same strategy of fine-tuning with a comparatively large dataset to update their model.

We measured several state-of-the-art tools (*i.e.*, Clarifai Text Moderation [67], Perspective API, Azure Text Moderation [68], and IBM Toxic Comment Classifier) against the hateful tweets in our dataset. Our objective in this measurement experiment was to study the capability of these existing systems on the new waves of online hate only, and we do not propose that these systems and models are not effective against hate in general, since they have been known to be effective against traditional hate [69]. We depict the results of this measurement experiment in terms of precision, recall, and F1-score in Table 3. We found that the existing tools are severely limited in discovering new waves of online hate observed from the low F1 scores reported by these systems. The highest F1 score was found to be just 0.38 (Perspective API), which is not sufficient for practical use.

| Detection Tools | Precision | Recall | F1-score |
|---|---|---|---|
| Clarifai Text Moderation [67] | 0.69 | 0.16 | 0.27 |
| Perspective API [14] | 0.49 | 0.31 | 0.38 |
| Azure Text Moderation [15] | 0.54 | 0.21 | 0.31 |
| IBM Toxic Comment Classifier [16] | 0.69 | 0.15 | 0.25 |

TABLE 3: Use of existing systems in detecting new waves of online hate.

**Finding.** It is evident that an existing detection tool might not perform well when new waves of online hate are presented to them. However, they can be augmented by different means such as zero-shot (or few-shot) learning to adapt to such rapid changes in the concept. We acknowledge that other factors, such as tool owners not periodically updating their models, could also limit the tools' effectiveness against new waves. However, we focus solely on the limitations of existing tools due to the rapidly evolving nature of online hate. This limitation indicates that new methods for the discovery and moderation of new waves of online hate must be developed.

## 5. HateGuard Design

### 5.1. Design Intuition

Before delving into our approach to addressing new waves of online hate, we provide a discussion about the intuitions behind the design of our approach.

**Reasoning-based decision-making for detection.** Detection of hateful content is not a simple classification task, it is a complex decision-making task that involves

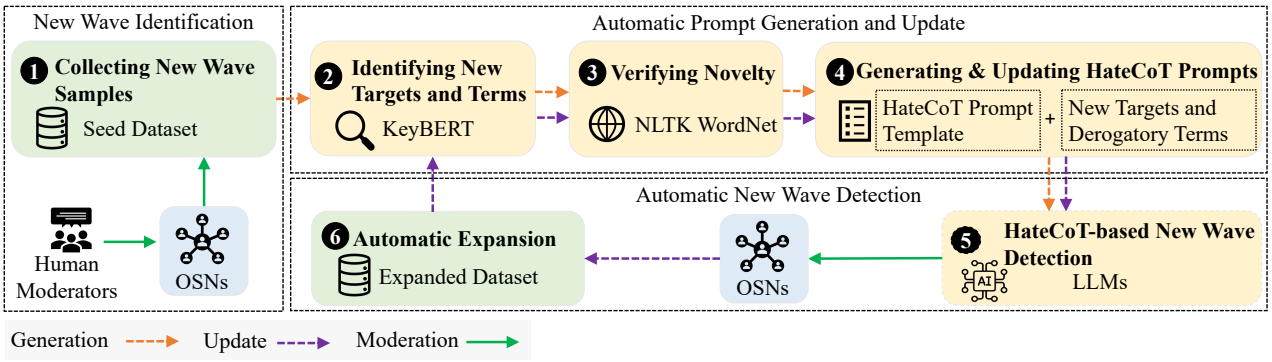



Figure 3: Overview of HATEGUARD.

reasoning [19]. A major reason for the complexity is due to its highly contextual nature. For instance, the decision whether or not new content is hateful is based on multiple factors and the interaction between these multiple factors. As an example, in identity-based hate [14], the mere mention of an identity is not sufficient for it to be decided as hateful, there needs to be an element of an attack involving derogatory words towards the mentioned identity. Furthermore, the identity could in reality be an entity, such as the US government or the United Nations, and derogatory words being used to express mere criticism, in which case it is not hateful. The task's complexity is heightened by scenarios where derogatory words are used without targeting a specific identity, resulting in posts that are offensive but not necessarily hateful [29]. This complexity is so intricate that even individuals from diverse backgrounds often find it challenging to discern whether a text is truly hateful [70]. Moreover, the emergence of new waves adds another layer of complexity, introducing fresh contexts, such as novel derogatory terms and new targets of hate, which further muddle the process of determining hatefulness.

The determination of whether new content is hateful or not is a complex and contextual decision-making process that is based on reasoning, and it is not a simple classification task. However, current ML-based techniques are based on the traditional paradigm based on training a model on a large hate speech dataset and making binary predictions. We argue that this paradigm is predominantly based on word associations in the training datasets, does not sufficiently consider context, does not consider hateful factors or the associations between them, and most importantly lacks reasoning-based decision-making.

The major challenge in performing this decision-making is how to practically control it on a large platform, such as social media. One option is to use human moderators. But using human moderators to achieve this goal is not suitable since humans' fatigue in such tasks [71] and the extremely concerning ethical issues that human moderators face, such as post-traumatic stress disorders (PTSD) after viewing such content [72]. Thus, the other option involving ML models is more suitable. However, current ML models are limited to classification tasks and are not suitable for complex reasoning-based decision-making. The recent invention of large language models (LLMs) has significantly revolutionized the landscape of NLP tasks [73]. Since these models are trained on massive amounts of data with a reinforcement paradigm, they can sufficiently capture the contextual information needed for NLP tasks and can be prompted to perform various tasks in a few or zero-shot manner. However, to develop LLMs to do reasoning-based tasks such as determination of hate, they can be prompted in several intermediate steps to arrive at the final decision based on the intermediate outputs, a recently introduced process known as *chain-of-thought* (CoT) prompting [23]. LLMs based on this prompting style have been shown to perform better on reasoning tasks, such as arithmetic problem solving [74], [75]. However, a CoT prompting process for online hate detection is yet unexplored. These intermediate prompts need to be thoughtfully designed according to hate speech definitions, allowing the model to determine if new content is hateful in a clear, step-by-step process, where each step considers the output of an intermediate step to generate the output. In this way, LLMs are not only capable of executing reasoning-based decisions for identifying online hate speech, but they also offer the advantage of scaling up this process efficiently. Additionally, they can sidestep some of the ethical challenges that social media moderators currently face. In our work, we first formulate a CoT strategy, called HateCoT (*i.e.*, Hate Chain-of-Thought) for reasoning-based decision-making of online hate.

**Learning from no or few new samples.** New waves of online hate occur suddenly. To effectively moderate them, we need AI-based discovery techniques that can be updated with no samples or only a few samples so that they can quickly adapt to an updated online hate policy and be deployed. However, training AI models with a few samples is not straightforward, as AI models need large datasets to be sufficiently trained. To effectively moderate new waves of online hate, there is insufficient time to collect and annotate large datasets for training a sufficiently effective classifier. Hence, we require methods that can be rapidly updated to moderate new waves of online hate using only a limited number of new wave samples.

In our work, we concentrate on zero-shot learning through the generation and updating of HateCoT prompts, incorporating new targets and derogatory terms associated with new waves of online hate. This is accomplished by automatically extracting the new targets and derogatory

terms using an NLP method.

## 5.2. Overview of our Framework

Our framework, illustrated in Figure 3, comprises three primary components: (1) New Wave Identification; (2) Automatic Prompt Generation and Update; and (3) Automatic New Wave Detection. Initially, human moderators from the OSN identify a small set of new wave samples at the start of the new wave. The objective here is to gather a limited dataset, which includes just a few samples of the new wave (like a few tweets), rather than a comprehensive collection of new wave samples. Subsequently, HateCoT prompts are generated automatically by identifying new targets and derogatory terms in the seed dataset, verifying their novelty, and updating the HateCoT prompt template with the new targets and derogatory terms. A pertinent example is the COVID-19 pandemic, during which Asian Americans were newly targeted with derogatory terms such as "Wuhan Virus" and "Bat Flu." These terms were then integrated into the HateCoT prompts, aiding in the detection of emerging online hate trends. Following this, the HateCoT prompts are used to perform reasoning-based decision-making to identify online hate in OSNs. An LLM is leveraged to apply these generated prompts to OSN posts, and the responses from the LLM are then examined to provide answers to the HateCoT prompts. In the final stage, the responses obtained from the LLM are scrutinized to ascertain whether the posts contain hateful content. Posts identified as hateful are used for extracting targets and derogatory terms and are subsequently flagged for moderation. Additionally, these flagged posts contribute to the automatic expansion of our new wave tweet dataset and facilitate the ongoing refinement of the HateCoT prompts at the *buildup* stage of the new waves.

## 5.3. Our Approach

**5.3.1. Collecting New Wave Samples.** A crucial need for addressing new waves is that the detection strategy should be adaptable to the new hate paradigm. One way to achieve this adaptability is by updating the targets and derogatory terms of the new waves. In our framework, we update the HateCoT prompts by continuously retrofitting them with new targets and derogatory terms in new waves.

In existing social media platforms like X, human moderators are tasked with monitoring harmful content (already part of their role [76]). Following existing approaches that propose to use human moderators to collect a limited, initial seed dataset of tweets [77], we anticipate that human moderators can pinpoint a small portion of related tweets as the initial dataset during the *buildup* phase of a new wave, and provide them to HATEGUARD for the automatic derivation of new targets and derogatory terms.

**5.3.2. Automatic Prompt Generation and Update.** Based on the seed dataset, HATEGUARD automatically generates prompts by identifying new targets and derogatory terms, verifying their novelty, and updating a prompt template with these new targets and terms. Then, the prompts are continuously updated by automatically expanding the set of new targets and terms to keep HATEGUARD up-to-date with the propagation of a new wave.

**Identifying New Targets and Terms and Verifying Novelty.** In HATEGUARD, we use an NLP method to identify new targets and terms, verify their novelty, and automatically expand our dataset of these elements. This approach leverages KeyBERT [78] to extract fresh targets and terms from the initial dataset and verifies their novelty using NLTK's WordNet [79]. The process iteratively broadens the scope of targets and terms by cross-checking new tweets against our existing dataset, integrating any newly identified elements to ensure comprehensive coverage.

Our NLP method has effectively pinpointed several new targets and terms in posts related to COVID-19. Notably, we uncovered terms such as "boomers" (indicating Ageism), "antimaskers", and "antivaxxers" emerging from the global debates on masks and vaccines, highlighting societal divisions. Our method also discovered derogatory terms like "BoomerRemover" (pertaining to ageism) and "Maskhole" (targeting anti-maskers).

**Generating and Updating HateCoT Prompts.** Online hate determination is quite a complex task, which can be broken down into sub-problems that can be individually addressed, and the results put back together to solve the overall problem. We use this characteristic of hate detection to design our HateCoT prompts. Solving this problem by breaking it down into sub-problems significantly improves the ability to perform complex reasoning that hate detection demands. Figure 4 depicts HateCoT prompting approach.

We craft the HateCoT prompts based on the *factors* of identity-based and individual hate, as inferred from their respective definitions. Recall the definition of hate as:

"*Hatred, hostility, or violence towards* member(s) of a *race, ethnicity, nation, religion, gender, gender identity, sexual orientation* or any other designated sector of society" [41].

From this definition, we identify four key factors of hateful content: (1) Presence of Target, (2) Derogation, (3) Direction, and (4) Incitation. Identifying online hate, therefore, is a comprehensive process that requires assessing each of these factors. We implement this through a CoT approach. This method simplifies the intricate task of detecting online hate by addressing each of the four factors as distinct sub-problems. The outcomes of these sub-problems are then integrated to assess whether content is hateful or not, forming the crucial fifth sub-problem. The specifics of these four factors are detailed below.

*Presence of Target.* A target should be mentioned in hateful content. In identity-based hate, these targets are based on several identities, such as race, nationality, political affiliation, religion, etc. In hate against individuals, the name or username of the individual is mentioned. In Figure 4, the first sub-problem is based on the presence of a target in the hateful content. This is operationalized with questions Q1a and Q1b, which respectively addressed the presence of identity targets and individual targets.

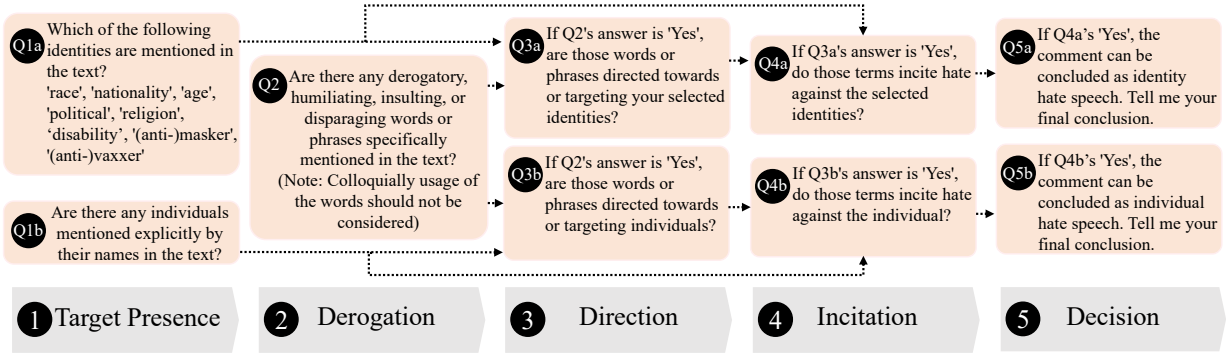



Figure 4: HateCoT prompts for new wave detection.

*Derogation.* From the definition, it can be observed that there is a presence of "hatred, hostility, or violence", that is often expressed in textual media using derogatory or disparaging words or phrases. In Figure 4, the second sub-problem is based on the presence of such words or phrases in the content. However, the language in social media platforms also consists of substantial colloquial use of derogatory words such as the f-word, and the sub-problem of detecting derogatory terms must be aware of such colloquial usages. In Figure 4, we operationalize derogation with question Q2, which addresses the mentioning of derogatory terms while being aware of the colloquial use of such terms.

*Direction.* Hate detection is complex enough that the mere presence of targets and derogatory terms are not sufficient to flag a post as hateful. An important factor of hate is that those derogatory terms must be *directed* at the target. For example, the text "lots of beautiful scenes during the *Chinese* new year, but my *stupid* camera isn't working". Although a target (Chinese) and a derogatory term (stupid) are mentioned, the term is not directed at the target. The third sub-problem is based on determining whether such terms are directed toward the target. We operationalize this factor with questions Q3a and Q3b in Figure 4, which address the direction toward identities or individuals, respectively.

*Incitation.* In addition to terms directed at the target, another factor in the detection process is whether the terms incite hate against a target. This differs from the direction of terms towards a target, since benign cases of certain terms directed towards a target can exist, such as "the f***ing Chinese are winning the space race". The fourth sub-problem is based on determining whether the detected derogatory terms in the second sub-problem are inciteful of hate toward the detected targets in the first sub-problem. We operationalize this factor with the questions Q4a and Q4b in Figure 4.

We further define the fifth sub-problem as a *decision*-making task, taking into account the context provided by the answers to all previous sub-problems. This sub-problem concludes the reasoning process and forms the final decision. As shown in Figure 4, this sub-problem is operationalized through the implementation of questions Q5a and Q5b.

### 5.3.3. Automatic New Wave Detection.

**HateCoT-based New Wave Detection.** After updating the HateCoT prompts, it's essential to employ a robust model for processing these prompts. Recent advancements have demonstrated that LLMs exhibit enhanced performance in reasoning tasks when prompts are presented in the form of a CoT [44]. The inclusion of intermediate steps in CoT enables the model to engage in more effective thinking and reasoning, thereby significantly improving its decision-making capabilities.

We leverage LLMs to execute our HateCoT prompts, ensuring their design is compatible with various LLMs. Notably, it has been observed that larger models tend to exhibit superior capabilities in CoT reasoning [75].

The LLM answers a prompt as follows. Given text input X and prompt t, the final answer is computed as,

$$\hat{y} = \text{argmax } p(y|X, t) \tag{1}$$

Instead of asking the LLM the final result $\hat{y}$, we break the problem into many sub-problems as discussed in Section 5.3.2, such that the model computes $\hat{y} \leftarrow t$ from several intermediate states $\hat{y} \leftarrow t_1 \leftarrow t_2 \ldots$. We do this as follows.
**Step 1.** First, we prompt the LLM to output the presence of identity or individual conditioned on the input text, and the identities of the new wave targets, depicted as follows:

$$\begin{aligned} A1a &= \text{argmax } p(a|X, Q1a) \\ A1b &= \text{argmax } p(b|X, Q1b) \end{aligned} \tag{2}$$

In the equation, $a, b, \ldots$ are intermediate answers that the LLM could output, such as $Yes$, $No$, and $N/A$.
**Step 2.** Next, we prompt the LLM to compute the presence of derogatory terms based only on the input sentence.

$$A2 = \text{argmax } p(c|X, Q2) \tag{3}$$

**Step 3.** Then, we prompt the LLM to compute the direction of derogatory terms based on the input sentence and the intermediate outputs from the previous steps.

$$\begin{aligned} A3a &= \text{argmax } p(d|X, A1a, A2, Q3a) \\ A3b &= \text{argmax } p(e|X, A1b, A2, Q3b) \end{aligned} \tag{4}$$

**Step 4.** Next, we prompt the LLM to output whether there is a presence of incitation based on the input sentence and the intermediate outputs from the previous steps.

$$\begin{aligned} A4a &= \text{argmax } p(f|X, A1a, A3a, Q4a) \\ A4b &= \text{argmax } p(g|X, A1b, A3b, Q4b) \end{aligned} \tag{5}$$

**Step 5.** The final decision is made by prompting the LLM to output a conclusion based on the input sentence and the previous output.

$$\hat{y1} = \text{argmax } p(h|X, A4a, Q5a)$$
$$\hat{y2} = \text{argmax } p(i|X, A4b, Q5b) \quad (6)$$

The presence of identity-based hate or individual hate is parsed based on the values of $\hat{y1}$ and $\hat{y2}$, respectively.

**Algorithm 1:** HateCoT Reasoning-based Decision-Making Algorithm

```
HateCoTPromptTemplate = S
Input: New Waves Dataset (D), Inference Function (I), Large Language Model (M)
// Extract new targets and derogatory terms
for x ∈ D do
    t_x = Targets(x)
    d_x = DerogatoryTerms(x)
    NewTargets ∪ {t_x}
    NewDerogatoryTerms ∪ {d_x}
end for
for s ∈ HateCoTPromptTemplate do
    UpdatedHateCoTPrompts = Update(s, NewTargets, NewDerogatoryTerms)
end for
                // Evaluation of new wave samples
for n ∈ D do
    Decision = I(n, UpdatedCoTPrompts, M)
    if Decision = 'Yes' then
        Enforce text control policy.
        Expand Dataset.
    end if
    else
        Share text.
    end if
end for
```

**Automatic Expansion.** Lastly, we outline the process for the practical deployment of HATEGUARD on real-world platforms, like social media, as methodically illustrated in Algorithm 1. The HateCoT prompt template under typical deployment called $HateCoTPromptTemplate$ is used to control hate speech that the platform is aware of. In the event of a new wave of hate, a small dataset (D) of these new instances is utilized to identify $NewTargets$ and $NewDerogatoryTerms$. Subsequently, the $HateCoTPromptTemplate$ is updated with these new targets and terms, rendering it ready for online deployment. The final step involves running our prompts through an LLM, which then processes the output for monitoring such content on social media platforms. If a new post adheres to the updated hate enforcement policy, it can be flagged. This decision is made by the LLM ($M$) as it evaluates the post against the updated prompts, identifying whether it contains identity hate or individual-targeted hate.

We dynamically expand our dataset of new wave samples by systematically integrating newly detected samples. This expansion is crucial for updating the HateCoT prompts with fresh targets and derogatory terms, as outlined in our NLP method (Section 5.3.2) and illustrated in Figure 4. Through this ongoing process, we ensure that the HateCoT prompts are automatically refreshed, effectively addressing new-wave instances and guaranteeing comprehensive coverage.

## 6. Implementation and Evaluation

In this section, we first discuss the implementation of our framework, followed by experiments to evaluate our approaches to all three new waves of online hate from different perspectives. Furthermore, we distinctly focus on the four categories of COVID-19-related online hate (*i.e.*, Anti-Asian, Ageism, Mask, and Vaccine) in addition to the other two new waves in our evaluation, due to the peaks of these categories occurring at different times. Our evaluation goals are summarized below.

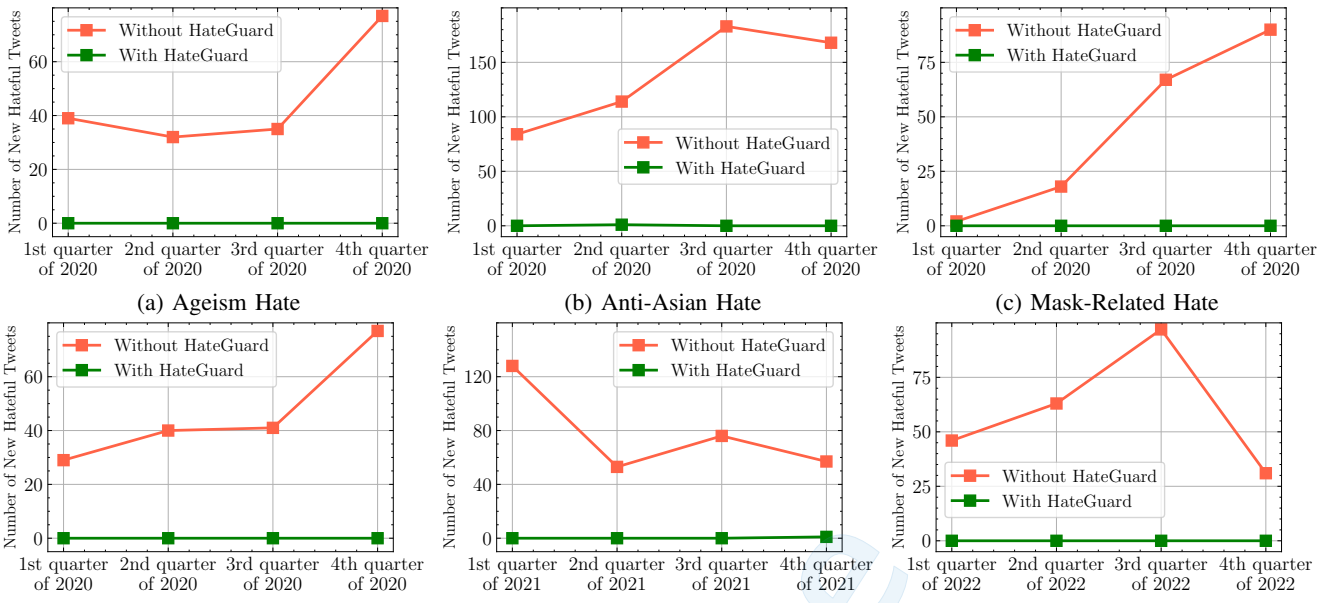
- Understanding the effectiveness of HATEGUARD by investigating the number of policy violations per quarter in the years 2020 (COVID-19 pandemic), 2021 (US Capitol insurrection), and 2022 (Russian invasion of Ukraine) (§ 6.3).
- Evaluating the effectiveness of our framework by comparing it with existing benchmarks (§ 6.4).
- Analyzing the effectiveness of HATEGUARD by comparing it with state-of-the-art ZSL, FSL, and generalized prompt strategies. (§ 6.5).
- Running HATEGUARD on "in-the-wild" unlabeled samples in our dataset. (§ 6.6).

### 6.1. Implementation

We used the GPT-4 model from the official OpenAI API endpoints to run HateCoT prompts as well as the generalized prompts [80]. We used KeyBERT version 0.8.3 and NLTK version 3.8.1 in our NLP method. Our labeled dataset was primarily utilized for the majority of our tests, while the unlabeled dataset was employed for the studies in Section 4.2 and the evaluations in Section 6.6. Furthermore, we provide detailed discussions on specific parameter settings, excluding those set to defaults, in the respective evaluation sections.

### 6.2. Baselines

To evaluate our framework, we present several key baselines for comparison: (1) two existing approaches, BERT-base [81] and Tweet-NLP [82], as discussed in Section 6.3; (2) fundamental approaches such as FSL [83] and Meta-EFL [84], covered in Section 6.5; and (3) a generalized prompt strategy for hate speech detection, derived from existing literature [80], which is elaborated in Section 6.5.




(a) Ageism Hate (b) Anti-Asian Hate (c) Mask-Related Hate

(d) Vaccine-Related Hate (e) US Capitol Insurrection Hate (f) Russian Invasion of Ukraine Hate

Figure 5: Deploying HATEGUARD in 2020 (COVID-19 pandemic), 2021 (US Capitol insurrection), and 2022 (Russian invasion of Ukraine) shows that new wave peaks are significantly reduced (green line).

### 6.3. Effectiveness of HATEGUARD in Reducing Number of Violations

In this section, we investigated the effectiveness of HATEGUARD in efficiently reducing the peaks of new waves of online hate. We were especially interested in learning whether HATEGUARD can reduce the peaks of new hate waves in a *real-world deployment simulation* scenario. We first categorized all the tweets belonging to the three new waves of online hate samples in the dataset according to different quarters and types. For COVID-19-related hate, we focused on the four quarters of 2020 since that year of the pandemic witnessed numerous waves of online hate. On a similar basis, we focused on the four quarters of 2021 for the US Capitol insurrection-related hate and 2022 for the Russian invasion of Ukraine-related hate. For these same periods, we deployed HATEGUARD for different hate categories and recorded the number of violations with our framework.

Figure 5 depicts the temporal progress of the spread of new waves without and with HATEGUARD, respectively, wherein the red line in Figure 5 indicates how many violations were made in each quarter, and the green line indicates the number of violations after deploying HATEGUARD. From Figure 5, we observed that all three new waves (including the four categories of COVID-19-related hate) reached significant peaks in at least one of the four quarters (for instance, the third quarter for Anti-Asian hate and the fourth quarter for Vaccine-related hate). However, the peaks were significantly reduced in those specific quarters, and the overall violations were reduced in each of the quarters with the deployment of our framework. In most quarters, HATEGUARD completely stopped new waves from occurring. For example, the peaks of each category of new waves were effectively reduced, demonstrating that our framework is capable of moderating various types of new waves by only updating the prompts.

### 6.4. Comparison with Existing Benchmarks

In this experiment, we studied the effectiveness of HATEGUARD in discovering new waves of online hate in comparison to existing benchmarks of transformer models [85]. We used two transformer models as benchmarks. The first, BERT-base-uncased mode, was fine-tuned using a leading dataset for hateful speech [81]. The second, Tweet-NLP model [82], is widely recognized for its effectiveness in hate speech detection on X, which is particularly relevant as our dataset of new hate waves was sourced from the same platform. For both models, we utilized the official implementations available online [83]. To clearly illustrate the effectiveness of HATEGUARD in addressing the emergence of new waves of online hate, we conducted a comparative analysis between our framework and these baseline models. This evaluation was performed quarterly, focusing on metrics such as accuracy, precision, and recall. In each quarter, we incrementally fine-tuned the benchmark models with the data from the preceding quarter. For example, in Q1, the models were applied in their original form. In Q2, we incorporated data from Q1 for further training, and this process continued sequentially. This approach was designed to mimic a real-world OSN scenario, where models are periodically updated to reflect new trends and policies in online hate speech. By following the guidelines provided in the referenced studies [81] [82], we trained both models over 50 training epochs. To evaluate HATEGUARD, we first randomly selected seed data with 10 to 20 tweets from the first month of each quarter to identify new targets and derogatory terms.

| Wave Type | Method | Quarter 1 (Jan-Mar) | | | | Quarter 2 (Apr-Jun) | | | | Quarter 3 (Jul-Sep) | | | | Quarter 4 (Oct-Dec) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | # of Tweets | Accuracy | Precision | Recall | # of Tweets | Accuracy | Precision | Recall | # of Tweets | Accuracy | Precision | Recall | # of Tweets | Accuracy | Precision | Recall |
| **- Overall Results -** | | | | | | | | | | | | | | | | | |
| Total (2020-2022) | **HATEGUARD** | 928 | **0.95** | **0.95** | **0.94** | 893 | **0.94** | **0.94** | **0.93** | 1148 | **0.94** | **0.94** | **0.93** | 1031 | **0.94** | **0.95** | **0.92** |
| | BERT-base | | 0.74 | 0.81 | 0.34 | | 0.82 | 0.76 | 0.71 | | 0.84 | 0.82 | 0.79 | | 0.83 | 0.86 | 0.8 |
| | Tweet-NLP | | 0.7 | 0.73 | 0.23 | | 0.83 | 0.79 | 0.77 | | 0.84 | 0.83 | 0.8 | | 0.83 | 0.84 | 0.8 |
| **- Category-wise Results -** | | | | | | | | | | | | | | | | | |
| Ageism (2020) | **HATEGUARD** | 186 | **0.94** | **0.91** | **0.92** | 117 | **0.95** | **0.95** | **0.95** | 114 | **0.95** | **0.95** | **0.95** | 161 | **0.95** | **0.94** | **0.96** |
| | BERT-base | | 0.82 | 0.6 | 0.44 | | 0.8 | 0.68 | 0.53 | | 0.79 | 0.68 | 0.6 | | 0.74 | 0.72 | 0.76 |
| | Tweet-NLP | | 0.79 | 0.5 | 0.15 | | 0.87 | 0.79 | 0.72 | | 0.86 | 0.74 | 0.83 | | 0.72 | 0.79 | 0.57 |
| Asian (2020) | **HATEGUARD** | 179 | **0.96** | **0.96** | **0.97** | 296 | **0.93** | **0.93** | **0.93** | 331 | **0.94** | **0.95** | **0.94** | 262 | **0.95** | **0.94** | **0.98** |
| | BERT-base | | 0.68 | 0.91 | 0.35 | | 0.84 | 0.79 | 0.8 | | 0.85 | 0.86 | 0.87 | | 0.87 | 0.88 | 0.92 |
| | Tweet-NLP | | 0.63 | 0.77 | 0.29 | | 0.84 | 0.84 | 0.72 | | 0.84 | 0.86 | 0.84 | | 0.85 | 0.91 | 0.86 |
| Mask (2020) | **HATEGUARD** | 16 | **0.99** | **0.99** | **0.99** | 64 | **0.94** | **0.96** | **0.88** | 249 | **0.98** | **0.95** | **0.97** | 199 | **0.96** | **0.97** | **0.94** |
| | BERT-base | | 0.75 | 0 | 0 | | 0.79 | 0.78 | 0.39 | | 0.85 | 0.75 | 0.66 | | 0.8 | 0.75 | 0.86 |
| | Tweet-NLP | | 0.94 | 0.67 | **0.99** | | 0.86 | 0.85 | 0.61 | | 0.87 | 0.75 | 0.78 | | 0.84 | 0.8 | 0.88 |
| Vaccine (2020) | **HATEGUARD** | 78 | **0.98** | **0.99** | **0.96** | 114 | **0.92** | **0.9** | **0.92** | 104 | **0.93** | **0.93** | **0.91** | 226 | **0.94** | **0.95** | **0.92** |
| | BERT-base | | 0.76 | 0.92 | 0.38 | | 0.78 | 0.68 | 0.7 | | 0.85 | 0.79 | 0.83 | | 0.84 | 0.75 | 0.79 |
| | Tweet-NLP | | 0.72 | 0.77 | 0.35 | | 0.75 | 0.6 | 0.83 | | 0.8 | 0.75 | 0.73 | | 0.88 | 0.83 | 0.82 |
| US Capitol (2021) | **HATEGUARD** | 311 | **0.91** | **0.91** | **0.88** | 112 | **0.99** | **0.97** | **0.99** | 158 | **0.9** | **0.89** | **0.9** | 123 | **0.9** | **0.9** | **0.9** |
| | BERT-base | | 0.68 | 0.79 | 0.31 | | 0.85 | 0.85 | 0.83 | | 0.82 | 0.85 | 0.76 | | 0.84 | 0.78 | 0.89 |
| | Tweet-NLP | | 0.63 | 0.7 | 0.16 | | 0.83 | 0.79 | 0.87 | | 0.82 | 0.87 | 0.72 | | 0.78 | 0.75 | 0.79 |
| Russia -Ukraine (2022) | **HATEGUARD** | 158 | **0.95** | **0.95** | **0.93** | 190 | **0.94** | **0.95** | **0.92** | 192 | **0.94** | **0.95** | **0.93** | 60 | **0.9** | **0.92** | **0.9** |
| | BERT-base | | 0.8 | 0.85 | 0.37 | | 0.8 | 0.72 | 0.62 | | 0.82 | 0.82 | 0.81 | | 0.83 | 0.82 | 0.87 |
| | Tweet-NLP | | 0.77 | 0.92 | 0.24 | | 0.84 | 0.82 | 0.65 | | 0.86 | 0.88 | 0.84 | | 0.85 | 0.89 | 0.81 |

TABLE 4: Comparing **HATEGUARD** against the existing benchmarks.

We utilized data from the remaining months of each quarter to assess HATEGUARD's performance.

The outcomes of our experiment are presented in Table 4. In the summary of Overall Results, HATEGUARD significantly surpassed the existing benchmarks, demonstrating a remarkable detection rate in identifying emerging new waves of online hate throughout all quarters. This superior performance is highlighted by the bolded numbers in Table 4. Specifically, HATEGUARD achieved impressive F1 scores of 0.95 in Q1, 0.94 in Q2, 0.94 in Q3, and 0.94 in Q4, indicating that it was effective and did not overfit the training data. Upon further investigation into each type of new wave, the results (in "Category-wise Results") consistently revealed that HATEGUARD was markedly more effective in identifying and flagging the content associated with these new waves compared to benchmark methods. For instance, HATEGUARD achieved a precision of 0.91 and a recall of 0.92 on Asian hate, a major category of concern during COVID-19 in Q1, maintaining this high level of performance through to Q4. Furthermore, HATEGUARD exhibited exceptional accuracy in its performance, achieving a 91% accuracy rate in Q1, the peak period for US Capitol-related hate, and maintaining this high accuracy level at 94% in Q3, the peak quarter for Russian-Ukraine-related hate. Interestingly, our framework demonstrated notable improvement in the first quarter, even with a small number of training samples. This suggests it can effectively moderate with minimal data, akin to zero-shot learning. As we approached the final quarter, the baseline models began to catch up slightly. For example, the BERT-base-uncased model achieved a 76% recall rate for Ageism, and the Tweet-NLP model reached an 88% recall rate for Mask-related hate, indicating their increasing effectiveness as more data becomes available. While the baseline models may be adequate for identifying hate speech within larger datasets, our framework demonstrated a significantly greater capability in effectively moderating new waves.

## 6.5. Comparison with other ZSL and FSL Methods

To study the importance of our HateCoT update strategy based on the ZSL paradigm, we compared HATEGUARD with one few-shot learning model, one zero-shot model, and

| Wave Type | Method | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Ageism | ZSL | 0.69 | 0.51 | 0.33 |
| | FSL | 0.5 | 0.36 | 0.52 |
| | GP | 0.84 | 0.69 | 0.88 |
| | **HATEGUARD** | **0.95** | **0.94** | **0.94** |
| Asian | ZSL | 0.74 | 0.75 | 0.75 |
| | FSL | 0.5 | 0.53 | 0.47 |
| | GP | 0.87 | 0.82 | **0.96** |
| | **HATEGUARD** | **0.95** | **0.95** | 0.94 |
| Mask | ZSL | 0.74 | 0.65 | 0.51 |
| | FSL | 0.49 | 0.34 | 0.48 |
| | GP | 0.84 | 0.74 | 0.8 |
| | **HATEGUARD** | **0.96** | **0.96** | **0.93** |
| Vaccine | ZSL | 0.73 | 0.72 | 0.4 |
| | FSL | 0.5 | 0.37 | 0.48 |
| | GP | 0.85 | 0.73 | 0.9 |
| | **HATEGUARD** | **0.94** | **0.94** | **0.93** |
| US Capitol | ZSL | 0.69 | 0.79 | 0.42 |
| | FSL | 0.54 | 0.49 | 0.56 |
| | GP | 0.83 | 0.75 | 0.9 |
| | **HATEGUARD** | **0.92** | **0.92** | **0.92** |
| Russia -Ukraine | ZSL | 0.81 | 0.84 | 0.63 |
| | FSL | 0.49 | 0.4 | 0.49 |
| | GP | 0.87 | 0.79 | 0.93 |
| | **HATEGUARD** | **0.94** | **0.95** | **0.94** |

TABLE 5: Results of different ZSL and FSL methods.

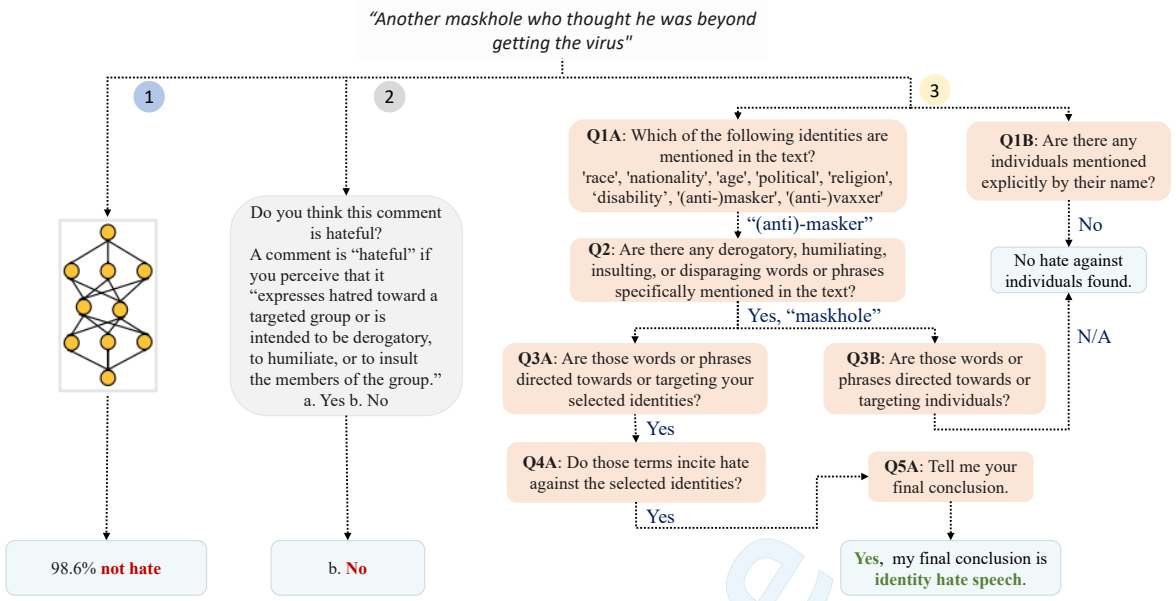



Figure 6: 3 HateCoT's prompt-based reasoning for new wave decision-making compared to 1 RoBERTa hate speech detection model [86] and 2 general prompting method [80].

a zero-shot model based on general prompting. The first method is a RoBERTa hate speech detection model fine-tuned by a benchmark dataset [86] (referred to as "ZSL" in this evaluation), the second method is the Meta-EFL model proposed by Meta [84] (referred to as "FSL" in this evaluation), and the third method is a general prompting method [80] based on ZSL (referred to as "GP"). For the first model, ZSL, we used the model by directly running it on the samples in our dataset. For the second method, we set the hypothesis for Meta-EFL as outlined in the original paper [84], where the entailment hypothesis is defined as "This tweet contains hate speech". In this experiment, we trained all the models for the same number of epochs and set the parameters as mentioned in their original fine-tuned state. For GP, we used the prompt mentioned in the original paper [80].

Table 5 presents the results of this experiment, depicting a detailed comparison of the accuracy, precision, and recall metrics of these methods with our framework. Notably, our framework outperformed the other three models based on ZSL/FSL. The primary aim of this experiment was to assess the effectiveness of the HateCoT prompts and our update strategy. First, we note that our approach significantly outperformed the ZSL and FSL models by a large margin, which could be attributed to the lack of any contextual detection offered by these models. While the GP strategy performed better than the other models, our prompting strategy vastly outperformed GP, especially in terms of precision. These results shed light on the contribution of chain-of-thought reasoning in the domain of hateful content detection, especially in new waves. To further clarify our approach, we display a specific example in Figure 6, featuring the sentence "Another maskhole who thought he was beyond getting the virus," analyzed using the ZSL model, the GP method, and our approach. This comparison highlights how our model's effectiveness is enhanced by its responses to a series of chained prompts. By breaking down the problem into a decision-making task involving sub-problems, such as identifying identities, derogatory terms, and their intended direction, the LLM is prompted to consider a range of possibilities, a crucial aspect in hate speech detection. This methodical, reasoning-based examination of a sample leads to an improvement in both precision and recall, as the final decision is derived through these intermediate steps rather than an immediate classification. More examples demonstrating our approach are available for review in Table 8, Appendix C.

### 6.6. Running HATEGUARD on "In-the-Wild" Samples

In this study, we tested our approach on unlabeled samples from our dataset to create a real-world, "in-the-wild" scenario. The goal of the experiment was to spontaneously flag potential hate content using HATEGUARD, and do a post-fact verification on whether the flagged content is indeed hateful.

In our experiment, we first randomly sampled 1,500 unlabeled samples from our dataset and ran HATEGUARD on these random samples. Two experts who are authors of this paper independently labeled the 1,500 samples based on the code book developed for labeling the new waves of online hate. We considered the two expert's annotations as ground truth and then analyzed the performance of HATEGUARD on the same samples. HATEGUARD was successfully able to flag 100% of the samples labeled as hate in the random in-the-wild samples, which indicates that it could be deployed as a main defense strategy on posts in real-world applications. Furthermore, in this experiment, HATEGUARD achieved a precision of 0.97 and a recall of 0.99, indicating that it is capable of achieving a very low False Positives (FP) rate. We hold the belief that HATEGUARD could be a poten-

tially potent defense mechanism in times of crisis, such as elections, where unchecked new waves proliferate. In such cases, it remains crucial to maintain the balance between curtailing hate speech and safeguarding free speech and legitimate criticisms, which are integral to the foundation of democratic societies. Given the notably low FP rate achieved by our approach, we hold confidence that it can effectively identify hateful posts while avoiding inadvertent censorship of harmless content. This, in turn, contributes to alleviating the burden and exhaustion endured by human social media moderators in the task of hate content moderation.

Additionally, to test the efficacy of our NLP method (Section 5.3.2), we conducted an assessment, utilizing 10 tweets from the COVID-19 pandemic dataset at the beginning of the buildup phase of Mask-hate. Our method effectively identified 1 target ("antimaskers") and 2 terms ("maskhole" and "maskoff") about Mask-hate based on the initial dataset, expanding to 39 targets/terms by the end of the buildup phase of Mask-hate. Our analysis indicated a strong alignment between the new targets and terms identified through our NLP method and those gleaned from the authors' manual analysis.

## 7. Discussion

**Broader Impacts of Chain-of-Thought Reasoning on LLM-based Security Applications.** To the best of our knowledge, our work is the first to leverage CoT reasoning within LLMs to address a critical cybersecurity challenge. We advocate the CoT strategy for scenarios requiring intricate decision-making, as many cybersecurity applications based on LLMs could significantly benefit from this approach, as evidenced in our study. A case in point is Microsoft's introduction of the Security Copilot [87], a cutting-edge tool that integrates ChatGPT for sophisticated threat analysis. While our current focus is on new waves of online hate, the versatility of our methodology is evident in its potential applicability to analogous challenges [88], [89], [90]. For instance, our approach could expediently enhance LLM-based malware detection systems [91], enabling them to swiftly adapt to zero-day malware with minimal initial samples. CoT-based reasoning might also streamline the training processes for transformer-based network intrusion detection systems [89], minimizing their dependency on extensive data. Additionally, CoT could be instrumental in generating varied fuzzing inputs across multiple programming languages [92]. We encourage cybersecurity professionals and researchers to further investigate the application of CoT principles in cybersecurity solutions.

**Limitation.** In our study, we primarily focus on text modality, the most common medium for disseminating online hate. However, emerging research highlights the increasing use of other modalities like images [93], videos [94], [95], and speech [96] in propagating online hate. Incorporating these modalities into our evaluations could offer a more comprehensive assessment of our framework. Additionally, our current analysis is confined to English-language posts. Expanding our scope to include other languages would provide a deeper insight into the effectiveness of our framework in diverse linguistic contexts. Moreover, we envision the application of HATEGUARD by content moderators, especially during critical crisis events, to identify and manage online hate content. However, there is a potential risk of unintended consequences, such as misclassifying benign comments as hate speech, if HATEGUARD is employed without proper supervision or review.

**Ethical Considerations.** We used workers from AMT to annotate the new waves of online hate dataset. Our data collection task was approved by IRB. We also warned workers about potential hateful content before they agreed to work on our task. In our paper, we have taken steps to minimize depicting samples of hate from our dataset, and we have carefully censored words that are extremely hateful or derogatory. We ensured the removal of mentions to user accounts so that user accounts could not be traced via public social media.

## 8. Conclusion and Future Work

In this work, we conducted a large-scale experiment to study the nature of new waves of online hate and showed how a new wave of online hate reaches a peak of activity, during which online hate spreads unabated. Then, we examined the capabilities of the existing moderation tools and found that they are significantly limited when used against the new waves of online hate. We proposed a novel framework HATEGUARD to practically address the problem of new waves of online hate. Our evaluation shows that HATEGUARD can significantly reduce the number of violations caused by new wave samples, and help in practically addressing this critical problem.

In the future, we aim to expand our framework to accommodate multilingual scenarios. This is crucial given the diverse linguistic landscape of global OSN platforms. To realize this, we plan to investigate multilingual encoders [97] that seamlessly integrate into our framework, primarily focusing on updating the content encoding step. Additionally, we plan to broaden our framework to include multimodal scenarios, particularly those combining image and text modalities, such as memes. Given the recent rise of memes as a medium for propagating online hate [93], we will explore how CoT prompting can be used to control new waves of multimodal online hate. Additionally, we plan to investigate the efficacy of alternative reasoning approaches, such as tree-of-thought (ToT) [74] and graph-of-thought (GoT) [98], in combating online hate. Finally, while our current work involves manually crafted prompts, we are keen on examining other prompt engineering techniques [99] that could further refine and enhance the efficacy of our prompts.

## 9. Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2245983, 2129164, 2114982, 2228617, 2120369, 2237238, 2239605, 2228616, and 2114920, and a National Centers of Academic Excellence in Cybersecurity grant No. H98230-22-1-0307. We thank Dr. Bimal Viswanath from

Virginia Tech for his valuable suggestions on enhancing the evaluation of HATEGUARD.

## References


[1] D. Valdez, M. Ten Thij, K. Bathina, L. A. Rutter, J. Bollen *et al.*, "Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of Twitter data," *Journal of medical Internet research*, vol. 22, no. 12, p. e21418, 2020.

[2] "Warning incitement of racial, religious hatred can trigger atrocity crimes, Special Adviser stresses states' legal obligation to prevent genocide," https://reliefweb.int/report/ukraine/warning-incitement-racial-religious-hatred-can-trigger-atrocity-crimes-special-adviser-stresses-states-legal-obligation-prevent-genocide, accessed: 2023-12-18.

[3] "Violence at Capitol and beyond reignites a debate over America's long-held defense of extremist speech," https://www.cnn.com/2021/01/19/us/capitol-riots-speech-hate-extremist-first-amendment/index.html, accessed: 2023-12-18.

[4] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, "Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 90–94.

[5] "A Tsunami Of Hate: The Covid-19 Hate Speech Pandemic," https://www.humanrightspulse.com/mastercontentblog/a-tsunami-of-hate-the-covid-19-hate-speech-pandemic, accessed: 2023-12-18.

[6] A. Prabhu, D. Guhathakurta, M. Subramanian, M. Reddy, S. Sehgal, T. Karandikar, A. Gulati, U. Arora, R. R. Shah, P. Kumaraguru *et al.*, "Capitol (pat) riots: A comparative study of twitter and parler," *arXiv preprint arXiv:2101.06914*, 2021.

[7] P. P. Chiang, "Anti-Asian racism, responses, and the impact on Asian Americans' lives: a social-ecological perspective," in *COVID-19*. Routledge, 2020, pp. 215–229.

[8] H. A. Choi and O. E. Lee, "To mask or to unmask, that is the question: Facemasks and anti-asian violence during covid-19," *Journal of human rights and social work*, vol. 6, no. 3, pp. 237–245, 2021.

[9] S. Wang, M. Schraagen, E. T. K. Sang, and M. Dastani, "Public sentiment on governmental covid-19 measures in dutch social media," 2020.

[10] C. Wardle and E. Singerman, "Too little, too late: social media companies' failure to tackle vaccine misinformation poses a real threat," *bmj*, vol. 372, 2021.

[11] D. A. Broniatowski, M. Dredze, and J. W. Ayers, ""First Do No Harm": Effective Communication About COVID-19 Vaccines," pp. 1055–1057, 2021.

[12] M. Wypych and M. Bilewicz, "Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among Ukrainian immigrants in Poland." *Cultural diversity and ethnic minority psychology*, 2022.

[13] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, P. Pratikakis, and S. Ioannidis, "Twitter Dataset on the Russo-Ukrainian War," *arXiv preprint arXiv:2204.08530*, 2022.

[14] "Perspective API," 2020, https://www.perspectiveapi.com.

[15] "Azure Text Moderation," https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/text-moderation-api, accessed: 2023-12-188.

[16] "IBM Toxic Comment Classifier," https://www.ml-exchange.org/models/max-toxic-comment-classifier/, accessed: 2023-12-18.

[17] "Hate Speech," https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fm.facebook.com%2Fcommunitystandards%2Fhate_speech%2F&refsrc=deprecated, 2021.

[18] "Hateful conduct policy," https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, accessed: 2023-12-18.

[19] T. Davidson, "Machine learning and the sociology of automated hate speech detection," 2022.

[20] K. Gelber, *Speaking back: The free speech versus hate speech debate*. John Benjamins Publishing, 2002, vol. 1.

[21] J. W. Howard, "Free speech and hate speech," *Annual Review of Political Science*, vol. 22, pp. 93–109, 2019.

[22] "Harmful content can evolve quickly. Our new AI system adapts to tackle it." https://ai.facebook.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/, accessed: 2023-12-18.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[24] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 247–267.

[25] "Hate Crime Laws," https://www.justice.gov/crt/hate-crime-laws, accessed: 2023-12-18.

[26] M. Duggan, "Online harassment 2017."

[27] "Microsoft, "Civility, Safety & Interaction Online"," https://news.microsoft.com/wp-content/uploads/prod/sites/421/2020/02/Digital-Civility-2020-Global-Report.pdf, accessed: 2023-12-18.

[28] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, and S. Zannettou, "On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning," *arXiv preprint arXiv:2212.06573*, 2022.

[29] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[30] "How Facebook uses super-efficient AI models to detect hate speech," https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/, accessed: 2023-12-18.

[31] "Twitter Says AI Flags Over Half of Tweets Violating Terms of Services," https://www.emergingtechbrew.com/stories/2020/07/20/twitter-says-ai-flags-half-tweets-violating-terms-services, accessed: 2023-12-18.

[32] "Facebook Community Standards Enforcement Report," https://transparency.fb.com/data/community-standards-enforcement/, accessed: 2023-12-18.

[33] "A healthier twitter: Progress and more to do," https://blog.twitter.com/en_us/topics/company/2019/health-update, accessed: 2023-12-18.

[34] "Content moderators at YouTube, Facebook and Twitter see the worst of the web and suffer silently," https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/, accessed: 2023-12-18.

[35] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 3, 2011, pp. 11–17.

[36] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*, 2012, pp. 19–26.

[37] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[38] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, pp. 69–101, 1996.

[39] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.

[40] M. Omar and D. Mohaisen, "Making Adversarially-Trained Language Models Forget with Model Retraining: A Case Study on Hate Speech Detection," 2022, pp. 887–893.

[41] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PloS one*, vol. 14, no. 8, p. e0221152, 2019.

[42] Billy Perrigo, "Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch," 2019. [Online]. Available: https://time.com/5739688/facebook-hate-speech-languages/

[43] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan *et al.*, "Show your work: Scratchpads for intermediate computation with language models," *arXiv preprint arXiv:2112.00114*, 2021.

[44] F. Huang, H. Kwak, and J. An, "Chain of explanation: New prompting method to generate higher quality natural language explanation for implicit hate speech," *arXiv preprint arXiv:2209.04889*, 2022.

[45] N. Vishwamitra, R. R. Hu, F. Luo, L. Cheng, M. Costello, and Y. Yang, "On analyzing covid-19-related hate speech using bert attention," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 669–676.

[46] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[47] L. Yan, Y. Zheng, and J. Cao, "Few-shot learning for short text classification," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 799–29 810, 2018.

[48] L. Stappen, F. Brunn, and B. Schuller, "Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL," *arXiv preprint arXiv:2004.13850*, 2020.

[49] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-lingual few-shot hate speech and offensive language detection using meta learning," *IEEE Access*, vol. 10, pp. 14 880–14 896, 2022.

[50] B. AlKhamissi, F. Ladhak, S. Iyer, V. Stoyanov, Z. Kozareva, X. Li, P. Fung, L. Mathias, A. Celikyilmaz, and M. Diab, "ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection," *arXiv preprint arXiv:2205.12495*, 2022.

[51] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[52] "2020 Time Capsule #5: The 'Chinese Virus'," https://www.theatlantic.com/notes/2020/03/2020-time-capsule-5-the-chinese-virus/608260/, 2020.

[53] "Urban Dictionary has a new word for coronavirus screw-ups: COVIDIOT," https://nypost.com/2020/03/24/urban-dictionary-has-a-new-word-for-coronavirus-screw-ups-covidiot/, 2020.

[54] "Coronavirus outbreak: What is "COVIDIOTS" trending on Twitter?" https://www.financialexpress.com/lifestyle/coronavirus-outbreak-what-is-covidiots-trending-on-twitter/1907432/, 2020.

[55] Y. Bao, C. Quan, L. Wang, and F. Ren, "The role of pre-processing in twitter sentiment analysis," in *Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10*. Springer, 2014, pp. 615–624.

[56] S. Taylor and G. J. Asmundson, "Negative attitudes about facemasks during the covid-19 pandemic: The dual importance of perceived ineffectiveness and psychological reactance," *PLoS One*, vol. 16, no. 2, p. e0246317, 2021.

[57] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[58] "Cdc museum covid-19 timeline," https://www.cdc.gov/museum/timeline/covid19.html, accessed: 2023-12-18.

[59] B. Lichtenstein, "From "coffin dodger" to "boomer remover": Outbreaks of ageism in three countries with divergent approaches to coronavirus control," *The Journals of Gerontology: Series B*, vol. 76, no. 4, pp. e206–e212, 2021.

[60] "Mask Rules Expand Across U.S. as Clashes Over the Mandates Intensify," https://www.nytimes.com/2020/07/16/us/coronavirus-masks.html, accessed: 2023-12-18.

[61] "A Timeline of COVID-19 Vaccine Development," https://www.biospace.com/article/a-timeline-of-covid-19-vaccine-development/, accessed: 2023-12-18.

[62] David French, "It's Clear That America Is Deeply Polarized. No Election Can Overcome That," 2020. [Online]. Available: https://time.com/5907318/polarization-2020-election/

[63] E. Winter, "The Russian Siege of the Azovstal Steel Plant in Ukraine: An International Humanitarian Law Perspective," 2022.

[64] Jigsaw, "Machine learning can help reduce toxicity, improving online conversation," 2023. [Online]. Available: https://jigsaw.google.com/the-current/toxicity/countermeasures/

[65] Perspective API, "Contribute Feedback," 2023. [Online]. Available: https://developers.perspectiveapi.com/s/docs-contribute-feedback?language=en_US

[66] IBM, "IBM Toxic Comment Classifier," 2023. [Online]. Available: https://www.ml-exchange.org/models/max-toxic-comment-classifier/

[67] "Clarifai," 2020, https://www.clarifai.com/.

[68] "Azure Content Moderator," 2022, https://azure.microsoft.com/en-us/services/cognitive-services/content-moderator/.

[69] B. Vidgen, S. Hale, S. Staton, T. Melham, H. Margetts, O. Kammar, and M. Szymczak, "Recalibrating classifiers for interpretable abusive content detection," in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 2020, pp. 132–138.

[70] Emily Ekins, "82% Say It's Hard to Ban Hate Speech Because People Can't Agree What Speech Is Hateful," 2017. [Online]. Available: https://www.cato.org/blog/82-say-its-hard-ban-hate-speech-because-people-cant-agree-what-speech-hateful

[71] A. Mao, E. Kamar, and E. Horvitz, "Why stop now? predicting worker engagement in online crowdsourcing," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 1, 2013, pp. 103–111.

[72] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support," in *CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[73] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, 2021.

[74] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

[75] Y. Fu, L. Ou, M. Chen, Y. Wan, H. Peng, and T. Khot, "Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance," *arXiv preprint arXiv:2305.17306*, 2023.

[76] X, "X Rules," 2023. [Online]. Available: https://help.twitter.com/en/rules-and-policies/x-rules

[77] P. Paudel, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Lambretta: learning to rank for Twitter soft moderation," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 311–326.

[78] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT." 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265

[79] A. Farkiya, P. Saini, S. Sinha, and S. Desai, "Natural language processing using NLTK and wordNet," *Int. J. Comput. Sci. Inf. Technol*, vol. 6, no. 6, pp. 5465–5469, 2015.

[80] L. Li, L. Fan, S. Atreja, and L. Hemphill, ""HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media," 2023.

[81] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[82] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[83] "Hugging Face: The AI community building the future." https://huggingface.co/.

[84] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as few-shot learner," *arXiv preprint arXiv:2104.14690*, 2021.

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[86] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection," 2021.

[87] Microsoft, "Microsoft Security Copilot," 2023. [Online]. Available: https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot

[88] S. Mahdavifar and A. A. Ghorbani, "Application of deep learning to cybersecurity: A survey," *Neurocomputing*, vol. 347, pp. 149–176, 2019.

[89] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "Flowtransformer: A transformer framework for flow-based network intrusion detection systems," *arXiv preprint arXiv:2304.14746*, 2023.

[90] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, pp. 1–22, 2018.

[91] "LLM-assisted Malware Review: AI and Humans Join Forces to Combat Malware," https://shorturl.at/loqT4, accessed: 2023-12-18.

[92] C. S. Xia, M. Paltenghi, J. L. Tian, M. Pradel, and L. Zhang, "Universal fuzzing via large language models," *arXiv preprint arXiv:2308.04748*, 2023.

[93] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[94] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr, and V. Almeida, "Analyzing right-wing youtube channels: Hate, violence and discrimination," in *Proceedings of the 10th ACM conference on web science*, 2018, pp. 323–332.

[95] Y. Yang, C. Noonark, and C. Donghwa, "Do YouTubers Hate Asians? An Analysis of YouTube Users' Anti-Asian Hatred on Major US News Channels during the COVID-19 Pandemic," *Global Media Journal-German Edition*, vol. 11, no. 1, 2021.

[96] J. L. Sullivan, "The platforms of podcasting: Past and present," *Social Media+ Society*, vol. 5, no. 4, p. 2056305119880002, 2019.

[97] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[98] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, L. Gianinazzi, J. Gajda, T. Lehmann, M. Podstawski, H. Niewiadomski, P. Nyczyk *et al.*, "Graph of thoughts: Solving elaborate problems with large language models," *arXiv preprint arXiv:2308.09687*, 2023.

[99] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

# Appendix A.

We provide the complete list of hashtags used for data collection in this work in Table 6.

| Category | Hashtags |
|---|---|
| Anti-Asian | '#chinesevirus', '#chinavirus', '#wuhanflu', '#batmaneatingflu', '#yellowmanflu', '#fuckchina', '#bombchina', '#ChinaLiedPeopleDied', 'boycottchina' |
| Ageism | '#boomerremover', '#boomerentomber', '#okboomer', '#boomerdeath', '#oldaf', '#boomermoober' |
| Mask | '#NoMask', '#NoMasks', '#MasksOff', '#MasksDontWork', '#WearAMask', '#WearADamnMask', '#MaskUp' |
| Vaccine | '#covid19vaccine', '#covidvaccine', '#pfizercovidvaccine', '#modernacovidvaccine', '#astrazenecacovidvaccine', '#biontechcovidvaccine', '#covidiots', '#iwillnotcomply' |
| US Capitol Insurrection | '#MAGARioters', '#MAGAMorons', '#MAGATerrorists', '#TrumpCrimeFamily', '#TrumpIsALaughingStock', '#TrumpCrimeSyndicate', '#TrumpInsurrection', '#TrumpCrimeFamilyForPrison' |
| Russian Invasion of Ukraine | '#NaziRussia', '#FkPutin', '#Rushit', '#getoutrussia', '#BanRussia', '#RussiaUkraineConflict', '#RussiaIsATerroristState', '#SanctionRussiaNow', '#PutinWar', '#Zelenskyclown', '#ZelenskyJoker', '#KillPutin' |

TABLE 6: List of hashtags.

# Appendix B.

We introduce our code book, devised by two experts and utilized for the Mturk annotation task in Table 7.

| |
|---|
| Please read the short instruction, then answer questions according to the following sentences: ${text} |
| Q1A: Which of the following identities are mentioned in the text? (select one or multiple options, "No identity" if there's not)<br>☐ Race ☐ Nationality ☐ Age ☐ Political ☐ Religion ☐ Disability ☐ Anti-masker<br>☐ Anti-vaxxer ◯ No identity |
| Q1B: Are the any individuals mentioned explicitly by their name?<br>◯ Yes ◯ No |
| Q2: Are you sure that the options you selected in Q1A are indeed identities and NOT just entities such as "her/she/they", CCP, Chinese army, USA government, etc.?<br>◯ Yes, I'm sure. ◯ No, I need to revise my Q1A's answer. |
| Q3: Are there any derogatory, humiliating, insulting, or disparaging words or phrases mentioned in the text?<br>◯ Yes ◯ No |
| Q4A: If Q3's answer is "Yes", are those words or phrases directed towards or targeting your selected identities? Select "No" if Q3's answer is "No".<br>◯ Yes ◯ No |
| Q4B: If Q3's answer is "Yes", are those words or phrases directed towards or targeting the mentioned individuals? Select "No" if Q3's answer is "No".<br>◯ Yes ◯ No |
| Q5: If Q4A or Q4B's answer is "Yes", are you sure that the identities or individuals selected are really being attacked? Or is it possible that the text is mere criticism of an entity?<br>◯ Yes, I'm sure. ◯ No, I need to revise my Q4's answer. |

TABLE 7: Code book developed by two experts.

# Appendix C.

Table 8 depicts additional samples from our dataset, including the corresponding HateCoT answers and their respective ground truth labels.

| Input Text | HateCoT Answers | Ground Truth |
|---|---|---|
| when i print that russians are debils it gets always corrected to devils actually not big difference | A1: The identity mentioned in the text is 'nationality' (Russians).<br>A1B: No, there are no individuals mentioned explicitly by their name.<br>A2: Yes, there are derogatory words used in the text ("debils" and "devils").<br>A3A: Yes, the derogatory words are directed towards the Russians.<br>A3B: Not applicable, no individual is mentioned by name.<br>A4A: Yes, the comment seemingly attempts to incite hate against Russians.<br>A4B: Not applicable, no individual is mentioned by name.<br>A5A: Yes, considering the words used and the nationality they are directed towards, the comment can be concluded as identity hate speech.<br>A5B: Not applicable, no individual is mentioned by name | hate |
| yeh sala china communist leftist maoist pig terrorist country and their dirty pig terrorist chinese people r dangerous to the whole world and humanity these chinese pig terrorist always spreading chinese virus and again a new chinese virus | A1: The identities mentioned in the text are 'nationality' and 'political'.<br>A1B: No individuals are mentioned explicitly by their name.<br>A2: Yes, there are several insulting or disparaging words specifically used in the text such as "sala", "communist", "leftist", "maoist", "pig", "terrorist", "dirty", "dangerous", "chinese virus".<br>A3A: Yes, these words or phrases are directed towards or targeting the selected identities which are 'nationality' and 'political'.<br>A3B: No, there are no words or phrases directed towards or targeting individuals.<br>A4A: Yes, these terms incite hate against the selected identities which are Chinese people and they are referred to as communists, leftists, and maoists.<br>A4B: Not applicable as there are no individuals targeted.<br>A5A: Since the terms used in the comment incite hate against the nationality and political identities, it can be concluded as 'Identity Hate'.<br>A5B: Not applicablet. | hate |
| every antimasker in public is a drunk driver killing Canadians treat them with the same respect you treat any drunk driving through a school zone they don't care if they destroy your life by killing your loved ones | A1: The identity mentioned in the text is '(anti-)masker'.<br>A1B: No individuals are mentioned explicitly by their names.<br>A2: Yes, 'drunk driver killing Canadians' and 'they don't care if they destroy your life by killing your loved ones' are derogatory and insulting phrases.<br>A3A: Yes, those words or phrases are directed towards the '(anti-)masker' identity.<br>A3B: No, words or phrases are not directed towards named individuals.<br>A4A: Yes, these terms promote negative attitudes and insights hate against the '(anti-)masker' identity<br>A4B: Not applicable, as no named individuals are insulted or disparaged.<br>A5A: Considering the above answers, the comment is identity hate speech.<br>A5B: Not applicable, as no named individuals are insulted or disparaged. | hate |
| i like a good conspiracy theory jfk moon landings roswell even boris trump and putin but aren t even real conspiracies because they can be proven basically it s a load of idiots who are just plainly dumb | A1: The identities mentioned in the text are 'political'.<br>A1B: Yes, individuals mentioned explicitly by their names are JFK, Boris, Trump and Putin.<br>A2: Yes, the phrases "a load of idiots" and "just plainly dumb" can be considered derogatory or insulting.<br>A3A: No, these phrases are not explicitly directed towards the political identity.<br>A3B: Yes, these phrases could be seen as directed towards the individuals, especially since it is in the context of conspiracy theories involving those individuals.<br>A4A: Not Applicable<br>A4B: No, the comment does not necessarily incite hate against these individuals.<br>A5A: Not Applicable<br>A5B: No, the comment cannot be concluded as individual hate speech since it does not purposely incite hate against the individuals mentioned. | Non-hate |
| the only people whit fealings are chinese people | A1: The identity mentioned here is 'nationality' (Chinese).<br>A1B: No, there aren't any individuals mentioned explicitly by their name.<br>A2: No, there aren't any derogatory, humiliating, insulting, or disparaging words or phrases used in the text.<br>A3A: Not applicable as A2's answer is 'No'.<br>A3B: Not applicable as A2's answer is 'No'.<br>A4A: Not applicable as A3A's answer is 'No'.<br>A4B: Not applicable as A3B's answer is 'No'.<br>A5A: Not applicable as A4A's answer is 'No', so the final conclusion is 'Non-hate'.<br>A5B: Not applicable as A4B's answer is 'No', so the final conclusion is 'Non-hate'. | Non-hate |
| hey i know how to engage teens in our synagogue have them be the purell dispensers since theyre likely to not get covid19 and also nobody wants them around anyway its genius mandated by an okboomer if i had to guess or like try something else maybe | A1: The identities mentioned in the text are 'age' and 'religion'.<br>A1B: No, there are no individuals mentioned explicitly by name.<br>A2: Yes, the phrase "nobody wants them around anyway" can be seen as derogatory or insulting.<br>A3A: Yes, the derogatory phrases are directed towards the identities 'age' (teens).<br>A3B: No, the derogatory words or phrases are not directed toward individuals.<br>A4A: The terms are not exactly inciting hate but rather belittling the 'age' group, which is not necessarily the same as hate speech.<br>A4B: Not applicable, there are no derogatory terms towards individuals.<br>A5A: Given that the comment is more belittling and does not necessarily incite hate, the final conclusion would be 'Non-hate'.<br>A5B: Not applicable, as there are no derogatory terms towards individuals. | Non-hate |

TABLE 8: Additional samples from our dataset.

## Appendix D.
## Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

### D.1. Summary

Online hate speech is hard to detect due to its rapid changes. This problem only gets amplified by the fact that deploying new detection models can take a long time as relevant data is data collection and a new model is trained. To bridge the gap, the authors proposed the HateGuard framework that depends on LLMs/Chain-of-Though (CoT) reasoning and updates in policy, rather than model, as a rapid response to hate speech in news events. With this framework, the spread of online hate speech can be reduced while a more robust model is developed.

### D.2. Scientific Contributions

- Provides a New Data Set For Public
- Use Creates a New Tool to Enable Future Science
- Provides a Valuable Step Forward in an Established Field

### D.3. Reasons for Acceptance

The paper provides a valuable step forward to a known problem. Hate speech detection is challenging due to its evolving content and temporal dependencies. Unlike prior approaches, which largely depend on training models on new data, the CoT-based approach reduces the time-to-deployment to help mitigate the spread of hate speech. Though the approach does not solve the hate speech detection problem, it does help minimize the spread of hate speech while more robust solutions are being developed.