Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations

David Hartmann
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
d.hartmann@tu-berlin.de

Amin Oueslati
Hertie School Berlin
Berlin, Germany
amin.m.oueslati@gmail.com

Dimitri Staufer
TU Berlin
Berlin, Germany
staufer@tu-berlin.de

Lena Pohlmann
TU Berlin
Berlin, Germany
Weizenbaum Institute for the Networked Society
Berlin, Germany
l.pohlmann@tu-berlin.de

Simon Munzert
Hertie School Berlin
Berlin, Germany
munzert@hertie-school.org

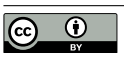
Hendrik Heuer
Center for Advanced Internet Studies (CAIS) gGmbH
Bochum, Germany
University of Wuppertal
Wuppertal, Germany
hendrik.heuer@cais-research.de

Abstract

Commercial content moderation APIs are marketed as scalable solutions to combat online hate speech. However, the reliance on these APIs risks both silencing legitimate speech, called over-moderation, and failing to protect online platforms from harmful speech, known as under-moderation. To assess such risks, this paper introduces a framework for auditing black-box NLP systems. Using the framework, we systematically evaluate five widely used commercial content moderation APIs. Analyzing five million queries based on four datasets, we find that APIs frequently rely on group identity terms, such as "black", to predict hate speech. While OpenAI's and Amazon's services perform slightly better, all providers under-moderate implicit hate speech, which uses codified messages, especially against LGBTQIA+ individuals. Simultaneously, they over-moderate counter-speech, reclaimed slurs and content related to Black, LGBTQIA+, Jewish, and Muslim people. We recommend that API providers offer better guidance on API implementation and threshold setting and more transparency on their APIs' limitations.
***Warning**: This paper contains offensive and hateful terms and concepts. We have chosen to reproduce these terms for reasons of transparency.*

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; Empirical studies in collaborative and social computing; • **General and reference** → *Measurement*; • **Social and professional topics** → **Hate speech**.



Keywords

Content Moderation APIs, Audit, AI Transparency and Accountability, Human-AI Interaction in Content Moderation, Algorithmic Bias in Hate Speech Detection

1 Introduction

Content moderation has become a widely used tool in combating online hate speech. While human moderators play an essential role in hate speech removal, human-based moderation is expensive, difficult to scale, and often exposes outsourced workers to distressing content that affects their mental health [34]. To address these challenges, companies such as Google, Microsoft, Amazon, Jigsaw, and OpenAI offer commercial, automated content moderation services. These API-based solutions are marketed as scalable, efficient solutions to tackle the growing challenges faced by social media platforms and other websites that deal with user-generated content [85]. For most major platforms, content moderation decisions are – to a large extent – partially or fully automated [21]. However, these decisions are potentially fallible. When harmful content is not moderated (under-moderation; reflected in a high False Negative Rate (FNR) of the content classifier), users are left unprotected from hate speech [23]. Conversely, when legitimate content is moderated (over-moderation; reflected in a high False Positive Rate (FPR)), this limits users' opportunities to express themselves and participate in public discourse. Both issues become aggravated when they systematically affect selected social groups, particularly those defined by protected characteristics such as gender, race, or religion.

While over- and under-moderation across different forms of hate speech has been extensively researched for off-the-shelf NLP models, research on over- and under-moderation of *commercial content moderation APIs* is limited. This is an important research gap since on off-the-shelf – meaning locally accessible – NLP models such as Sap et al. [90], Wiegand et al. [102] and Röttger et al. [87], have

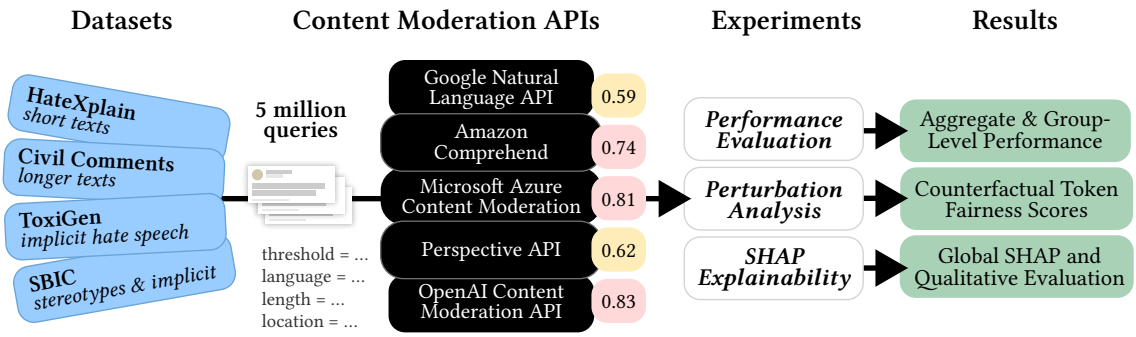


Figure 1: Our black-box audit framework to evaluate commercial content moderation APIs.

shown that these systems are prone to labeling content mentioning target groups such as Women, Jews and Muslims as more toxic. They also mistakenly flagged counter- and reappropriated speech as hateful, and misclassified content merely mentioning target identities as hate speech [23]. Additionally, Wiegand et al. [101] and ElSherief et al. [27] find that language models under-moderate variations of language. Specifically, they under-moderate implicit hate speech that uses indirect language to belittle a person or group based on protected attributes without using slurs or specific group identifiers. In contrast to off-the-shelf NLP models, content moderation APIs only grant so-called black-box access, which is limited to querying the model without actual access to the underlying algorithm, its architecture and weights (see Casper et al. [15]). In summary, as of now, there has been a lack of systematic evaluation of commercial content moderation services, leading to a concerning absence of public scrutiny.

In this paper, we introduce a framework for auditing black-box NLP systems. Using this framework, we evaluate five commercial content moderation APIs, analyzing over five million queries across four benchmark datasets and three experiments (see Figure 1). We find that commercial APIs frequently rely on group identity terms, such as "black", to predict hate speech. While OpenAI Content Moderation and Amazon Comprehend perform slightly better, all providers under-moderate implicit hate speech, which uses codified messages without identity terms, especially against *LGBTQIA+* people. Simultaneously, they over-moderate counter-speech, reclaimed slurs and content related to *LGBTQIA+*, *Black*, *Jewish* and *Muslim* people. Drawing on our findings, we recommend that API providers offer better guidance on API implementation and threshold setting and are more transparent about their APIs' limitations.

In summary, we contribute: (1) The first comprehensive audit of five widely used commercial content moderation algorithms, (2) a reproducible and query-efficient audit framework of NLP models that solely assumes black-box access, and (3) design and research recommendations on how providers should change the implementation and transparency of content moderation systems. More generally, our work provides a contribution to an AI accountability ecosystem that involves third-party auditors as an oversight mechanism [8, 84], fostering greater user trust in AI systems [46].

2 Background and Related Work

2.1 Hate Speech, Target-Groups and Linguistic Variations

Hate speech can have profound real-world effects, including the suppression of marginalized voices, social exclusion, discrimination, and violence against marginalized groups [58, 60]. At the same time, hate speech is a complex and contested concept that varies depending on the community and sender context, linguistic features, and political institutions [3, 13, 106]. In this paper, we adopt a broad conceptualization of hate speech similar to Marques [58]. We conceptualize hate speech as a discursive act of discrimination, which operates on its targets in constitutive and causal ways to effect the denial of equal opportunities and rights. Target identities, therefore, play a key role in defining hate speech, as it typically involves discriminatory acts against specific groups.

A related concept to hate speech is toxicity, a broader category that includes rude, disrespectful, or unreasonable behavior. Toxicity encompasses offensive or harmful language, even when it does not specifically target individuals or groups based on characteristics such as race, gender, or religion [6, 23]. For instance, Dixon et al. [23] defines toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Barth et al. [6] highlight the overlap between toxicity and hate speech, noting that while toxicity encompasses hate speech, it includes a wider range of offensive language that may not directly target a group based on its characteristics.

Conceptualizing and ultimately detecting hate speech is complicated by the fact that not all hate speech manifests in explicit language, and not all expressions featuring explicit vocabulary qualify as hate speech. For instance, Contrastive non-hate is language that shares linguistic features with hateful expressions [87, 98]. Distinct hate expressions span sub-categories such as explicit hate speech – offensive speech, threats, sexually abusive language, or slurs – and various forms of implicit hate speech, which do not fall under these categories yet still constitute a discursive act of discrimination [101].

Implicit hate speech is speech that conveys harmful or discriminatory intent without relying on overt slurs, profanity, or explicit hateful language as laid out by Yin and Zubiaga [105]. Instead, implicit hate speech often manifests through stereotypes, sarcasm,

irony, humor, or metaphors, which pose significant challenges for ML systems to detect [63, 90, 101]. Still, these subtle forms of hate speech can be just as damaging as explicit hate speech [2, 12], yet they frequently evade detection due to the absence of distinctive keywords, slurs and potentially clear target group identity tokens. For instance, a statement like "there is a direct correlation between the amount of melanin in a person's skin and how much they are worth as a person," found in the ToxiGen dataset [40], illustrates how implicit hate speech can convey racism without directly referencing racial identity or racial slurs.

Contrastive non-hate variations, such as re-appropriation and counter-speech, further complicate the detection of hate speech, as these nuances often blur the line between harmful speech and expressions of empowerment or resistance [70]. Re-appropriation is a sociocultural process where marginalized groups reclaim derogatory language, using it as an expression of empowerment, often among themselves [31]. What might be considered hate speech in one context could serve re-appropriation or reclamation in another. Counter-speech, on the other hand, refers to responses aimed at challenging hate speech, often using similar linguistic features to subvert or counter discriminatory narratives [70, 107].

2.2 Challenges and User Perceptions of Hate Speech Moderation

Online platforms have responded to the proliferation of hate speech by adopting extensive content moderation regimes [20] and assessing potential hateful content against so-called community guidelines by human workers, i.e., content moderators, sometimes assisted by ML-based systems. In some instances, content moderation is even conducted by algorithms entirely autonomously [37]. Content moderation by humans is often outsourced to low-paid workers employed by third-party providers or business process outsourcing companies (BPOs), outside the Global North, who are exposed to distressing content [1]. These workers face severe mental health challenges while earning as little as 1.8 USD per hour [34, 61]. As a result, automated content moderation was proposed as a scalable and cost-effective alternative. These systems potentially enable companies to remove harmful content more efficiently, but they also introduce new challenges related to transparency, fairness, and accountability [37].

Content moderation as a part of online speech governance is a wicked problem characterized by difficult trade-offs and significant contestation [24]. The contentious and political nature of concepts such as hate speech, toxicity, counter-speech, and reappropriation, combined with power asymmetries between political institutions, platforms, moderators, and users, underscores the difficulty of achieving equitable moderation outcomes [95]. These challenges apply equally to human and automated content moderation systems, which have both been shown to exhibit biases. For instance, Sap et al. [91] and Zhang et al. [108] have documented systematic biases related to stereotyping among human moderators, which can propagate to annotations and automated models [18, 91]. The work presented in this paper focuses specifically on addressing biases and functionality errors in automated content moderation systems.

Automated decision-making for content moderation can systematically err in two primary ways: first, in the systematic, algorithmic classification of non-hateful content as hateful (*over-moderation*), and second, in the systematic, algorithmic classification of hateful content as non-hateful (*under-moderation*). Over- and under-moderation are closely related to sender and target group dynamics. Under-moderation occurs when harmful content directed at certain groups is disproportionately allowed, leaving some users less protected than others online [23]. For example, the under-moderation of hate speech targeting LGBTQIA+ individuals [22] can exacerbate systemic discrimination, exclusion, and even violence against these communities. This not only contributes to their marginalization but also erodes user trust and can lead to affected users abandoning a platform altogether. Classifiers that fail to adequately protect certain user groups, particularly those defined by gender, race, or other protected characteristics, are considered biased [10], entrenching forms of representational harm against these communities [5].

Conversely, over-moderation, such as the misclassification of African-American English (AAE) [89], can silence voices and prevent counter-speech, leading to reduced participation from user groups using these speech characteristics [9]. When content moderation systems disproportionately block content from specific social groups, they not only diminish participation but also create distributive harms, where certain groups are unfairly censored. This aligns with broader discussions of algorithmic fairness [7], where the design of classifiers may disproportionately flag content from certain groups, leading to the unjust removal of their expressions. In this way, hate speech classifiers risk perpetuating inequalities by privileging certain forms of speech over others.

The Human-Computer Interaction (HCI) community has highlighted consistency, transparency, and fairness as critical elements for building user trust in content moderation systems [14, 93]. Consistency in the application of moderation policies – across different content types, users, and regions – has shown to increase trust and improve discourse. However, public perception remains that moderation practices are inconsistently implemented, leading to user dissatisfaction [94]. This is particularly true for marginalized communities, including women and racial, gender, and sexual minorities, who often experience disproportionately more moderation than others, exacerbating their sense of exclusion [41, 56]. The lack of transparency in how moderation decisions are made further alienates these users, making it harder for them to trust the platforms where they engage.

Systematic over- and under-moderation can have long-term negative effects on user trust, especially for content creators from marginalized groups. Over-moderation, such as the wrongful censorship of disability-related content, has been identified by Heung et al. [41] as a form of ableism, leading to self-censorship and disengagement from platforms. In contrast, under-moderation allows harmful content to persist, contributing to a decreased sense of safety and trust in both the platform and the broader online environment [94, 97]. Ambiguous moderation decisions, especially from ML-based systems, worsen these issues, as users are more likely to question the accountability and fairness of AI moderators when content is inherently unclear as Ozanne et al. [77] demonstrated. This perceived inconsistency can result in marginalized

users leaving platforms altogether, as platforms fail to effectively address their concerns [41].

2.3 ML-based Over- and Under-Moderation

Machine Learning (ML)–based hate speech classification systems such as off-the-shelf[1] NLP models have been tied to over-and under-moderation. First, ML hate speech models suffer from poor generalization, as data set creators oversampled certain users in the training data. For instance, 70% of all sexist tweets and 90% of all racist tweets in a hate speech dataset by Waseem and Hovy [100] belonged to a single author [102]. Second, research revealed an over-moderation of neutral statements mentioning target groups such as "woman" [96], which was coined as systematic offensive stereotyping (SOS) bias [25]. This is fueled by the historical discrimination of groups susceptible to algorithmic disadvantage, which often makes them the target of hate speech [102]. Algorithms internalize such spurious correlations as they occur in the training data. Third, reappropriated language [90] and counter-speech [87] have been linked to algorithmic discrimination of marginalized groups. For instance, Sap et al. [90] find that surface markers of AAE are more strongly correlated with hate speech than what the authors call *White-Aligned-English*.

It is worth noting that prior research concerning biases in content moderation algorithms heavily focused on women and people of colour, but mostly disregarded the discrimination of other groups; e.g., *LGBTQIA+*, people with *Disabilities* or *Latinx* [33, 74]. Similarly, most research probed the performance of hate speech moderation algorithms in English. Tests on non-English languages pose the exception and include, for instance, Hindi in Ghosh et al. [35] or Arabic in [48]. These studies often find that content moderation algorithms perform worse in non-English languages.

2.4 Auditing and Mitigating Over- and Under-Moderation

Recognizing these challenges, researchers have called for targeted strategies to systematically evaluate and mitigate such over- an under-moderations. For instance,Yin and Zubiaga [105] emphasize the need for more rigorous evaluation of hate speech detection models. They propose models should be (1) tested on datasets not seen during training, (2) subject to detailed error analysis addressing specific challenges, and (3) evaluated using methods that account for diverse interpretations of hate speech.

In response, audits – systematic evaluations of model functionalities and problematic machine behavior [4] – have been conducted on off-the-shelf systems. One prominent example of such an audit of automated content moderation systems is HateCheck, developed by Röttger et al. [87], which evaluates functionalities of hate speech detection algorithms, including linguistic variations such as counter-speech, negation, reappropriation, and systematic offensive stereotyping bias. HateCheck findings align with prior observations, revealing biases in off-the-shelf NLP models across target groups, including women, people with disabilities, and immigrants, and functionality issues with linguistic variations such as counter speech, systematic offensive stereotyping bias, and reclaimed slurs [87].

Furthermore, in response to findings about biases and functional failures, several research papers attempted to mitigate such shortcomings. For instance, extensive research was conducted to mitigate biases against specific groups [19, 76, 78, 89]. Additionally, research aimed to reduce racial bias in relation to AAE and specific dialects [68, 103, 109], reduce systematic offensive stereotyping bias [26], and mitigating false positives with regard to reclaimed language [110].

In summary, off-the-shelf NLP hate speech detection systems suffer from biases across marginalized target groups and are tied to specific functionality failures. However, significant efforts were made to reduce such biases and failures. We collect known off-the-shelf NLP hate speech model functionality failures in Tables 14 and 15.

2.5 Auditing Commercial Content Moderation Systems

With growing awareness of errors in off-the-shelf models and ongoing mitigation efforts, it is important to understand whether these errors extend to commercial content moderation systems or if common shortfalls have been effectively mitigated in commercial applications.

Commercial content moderation systems include, for example, Perspective API, a content moderation tool developed by Jigsaw – a tech incubator and subsidiary of Alphabet Inc. (Google) [45] – and OpenAI's Moderation API. Perspective API has been extensively utilized in academic research on online toxicity [42]. Additionally, Perspective API is often employed as a benchmark for toxicity in model comparisons [62] and as a tool to identify toxic content online, serving as a form of ground truth for evaluating platforms and communities [85]. It is widely used to measure the toxic outcomes and biases of LLMs [32], and it has been suggested that Perspective API has become a cornerstone for academic research on online abuse and incivility [75]. Meanwhile, OpenAI's Moderation API functions as the hate speech filter for ChatGPT and GPT models [57].

Although research specifically addressing content moderation APIs remains limited, both Perspective API and OpenAI's Moderation API have been evaluated in certain contexts. For instance, Röttger et al. [87] assessed Perspective API alongside off-the-shelf models. They found that Perspective API exhibited fewer functional issues than off-the-shelf models but still faced significant challenges with counter-speech, negations, and reclaimed slurs. Their functionality tests further revealed that while Perspective API occasionally over-moderated specific groups, it did not exhibit systematic biases against particular target groups. However, these tests did not cover implicit hate speech or dialects, relying instead on a specific conceptualization of hate speech and focusing solely on Perspective API as a commercial system.

Mihaljević and Steffen [62] examined the potential and limitations of detecting antisemitic content with Perspective API. Their findings indicated that the API could identify antisemitism at a very basic level. However, much like off-the-shelf NLP hate speech detection systems, Perspective API struggled with understanding

[1]We refer to off-the-shelf models as pre-trained models that provide white-box access to parameters, gradients, and weights. These models are often developed for research purposes and can be accessed locally or via remote servers.

	OpenAI Moderation Endpoint	Perspective API	Azure Content Moderator	Google Natural Language API	Amazon Comprehend
Developer	OpenAI	Jigsaw (Google)	Microsoft	Google Cloud	AWS
Rate Limit	25 queries/s	1 query/s	Free: 1 query/s, Commercial: 10 queries/s	10 queries/s	20 queries/s
Cost per Unit	Free	Free	$0.40 per 1000 calls	$0.0005 per 100 characters	$0.0001 per 100 characters
Model Date	28.08.2023	×	×	01.03.2023	×
Model Version	PaLM 2	×	×	×	×
API Version	v2	2017-11-27	v1.0	text-moderation-001	v1alpha1
Model Type	×	×	×	×	Multilingual BERT-based models

Table 1: Overview of the five selected commercial content moderation APIs. '×' stands for not disclosed by developer. An extended Table is presented in the Appendix A.1.

implicit antisemitism and often misclassified legitimate counter-speech as hateful. Similarly, Dias Oliva et al. [22] demonstrated that Perspective API systematically over-moderates LGBTQIA+ communities and counter-speech.

OpenAI's Moderation API was evaluated by Mahomed et al. [57], who investigated hate speech filtering in GPT-generated content. Their study generated 3,309 synopses for 100 popular U.S. TV shows and probed GPT's content moderation system. The findings revealed that a substantial portion of synopses were flagged as content violations, with specific genres statistically linked to higher flagging rates –highlighting an instance of over-moderation.

Overall, while content moderation APIs have diverse applications, as we will discuss later, there remains limited research on the biases and failures of these systems. Beyond the evaluations of Perspective API and OpenAI's hate speech filters, little attention has been paid to the performance of other cloud-based content moderation APIs. Existing investigations tend to focus on isolated failures or disparities within specific use cases. We contribute to an AI audit ecosystem and broader evaluation of commercial content moderation systems by conducting a systematic, comparative evaluation of multiple content moderation APIs with respect to the demands by Yin and Zubiaga [105].

3 Content Moderation APIs

3.1 Selection of APIs

There is no systematic research on the usage frequency of commercial content moderation services. Based on their recurring presence and the significant role they play in moderating content across various platforms, we focus on the following five APIs: Google Natural Language API[2], Amazon Comprehend[3], Microsoft Azure Content Moderator[4], Perspective API[5], and the OpenAI Content Moderation API[6]. An overview of these APIs are provided in Table 1. Among them, the Perspective API and OpenAI Moderation API stand out for their specific use cases and widespread adoption. Perspective API, widely utilized in academic research on online toxicity [42], is also employed by prominent organizations such as the New York Times, Vox Media, and OpenWeb. In 2021, it processed 500 million requests daily [45]. Similarly, OpenAI's Moderation API, which acts as the hate speech filter for ChatGPT and GPT models [57], supports more than 200 million weekly users [86].

3.2 Integration and Application

Figure 2 shows the basic usage of these APIs for moderation of text input. A comment, post, prompt, or any other text is sent to the API. In addition, users provide information on language, sentence length, and location data, and specify a decision threshold. The length of this text can vary. However, most APIs have a maximal text length, and OpenAI specifies that for maximal performance, a text chunk's optimal length is 2,000 characters. Most of these services have prices per 1,000 tokens; as of today, only Perspective API and OpenAI Moderation are free to use. Free tiers of other APIs come with severe rate limits.

After providing the text input, the API returns an array with various confidence scores depending on the specific service. Where most earlier published classifiers and datasets have cast the detection problem of hate speech to be a binary classification problem [90], current work and all of the audited content moderation APIs classify hate speech along distinct sub-categories. An overview of the characteristics and transparency of APIs is provided in Appendix Tables 6 and A.2.

The designers and community organizers of end-user applications are responsible for determining how to handle the output, and there are various ways to manage the output of a content moderation API based on the application and integration of the API [85]. Score thresholds that define the occurrence of hate speech or one of its sub-categories can vary, just as the action taken in response. Possible sanctions do not only include deletion; instead, platforms rely on a range of interventions such as implementing age barriers, geo-blocking, or temporary holds, appending fact-checking labels and trigger warnings, and not recommending the content to anyone [36].

To illustrate, Rieder and Skop [85] describes for Perspective API how potential moderation decisions can be processed from the

[2]See https://cloud.google.com/natural-language?hl=de
[3]See https://aws.amazon.com/comprehend/features/?nc1=h_ls
[4]See https://learn.microsoft.com/en-gb/azure/ai-services/content-moderator
[5]See https://perspectiveapi.com/
[6]See https://platform.openai.com/docs/guides/moderation

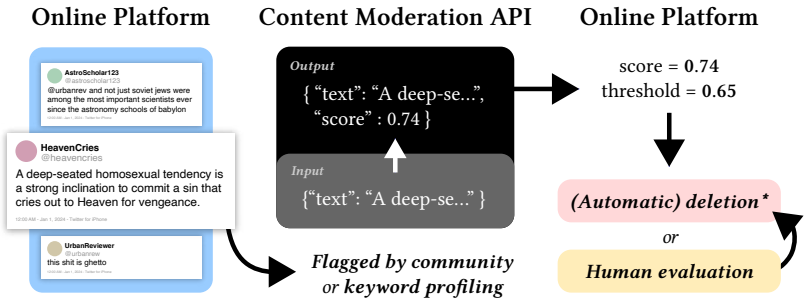


Figure 2: The pipeline of content moderation APIs, exemplary illustration with a blog post.

API's output. First, the output can be sent to a content moderator in the case of online moderation or a *human in the loop* (HINL) that is guided in their decision by the output score. Second, a hybrid or cascade solution is one where only specific content that is close to the threshold is given to a human. Other algorithms provide guidance on which content needs human review, similar to the active learning strategies in ML [54]. Third, blogs or deployers with limited resources could decide to just define a threshold and moderate content with rules for sub-categories, which would be equivalent to automatic deletion of the displayed comment in Figure 2. Jigsaw, the company behind Perspective API, discourages full automation, recognizing that ML models are prone to errors. They further recommend human oversight to mitigate potential mistakes and ensure that moderation actions remain contextually appropriate [85].

4 Audit Framework

Third-party audits are critical for ensuring accountability and transparency in ML systems [39, 83]. Unlike internal or contracted audits, independent audits offer greater scrutiny as they operate without company interference [8, 84]. However, third-party auditors commonly have limited access to the systems they evaluate [39]. This is also true for the commercial content moderation systems under study, where access to internal components like weights, gradients, and training data are restricted. This limitation makes the auditing process more challenging, as the analysis must rely on input-output behavior without insight into the model's inner workings [15, 43]. In face of these challenges, third-party audits are crucial for an AI accountability ecosystem.

We propose a black-box audit for content moderation APIs[7], as shown in Figure 1. The audit comprises a set of experiments with whom we seek to answer three research questions:

(1) **Comparative Performance:** How do the selected APIs perform in terms of hate speech detection, particularly across different target groups and linguistic variations?
(2) **Over- and Under-Moderation:** Where do these APIs tend to over-moderate or under-moderate specific target groups and linguistic variations?
(3) **Failures:** Why do these APIs over- and under-moderate in certain contexts, and what do these misclassifications reveal about their general functionality and operational design?

We apply this framework to evaluate the five mentioned commercial content moderation APIs. Overall, we collect and analyze over five million queries using four benchmark datasets.

We design three experiments addressing the research questions: (1) comparative aggregate and target-group performance, (2) perturbation sensitivity analysis (PSA) with counterfactual fairness scores, and (3) Shapley Additive exPlanations (SHAP) explainability. In total, we explore over- and under-moderation of eight marginalized groups, including (*Asian*, *Black*, *Disability*, *Female*, *Jewish*, *Latinx*, *LGBTQIA+*, *Muslim*), which can be mapped across all datasets with limited assumptions. We only use these eight to create maximum comparability.

In Experiment 5.1, we assess each API's performance across four benchmark datasets in aggregate and across the eight target groups. By comparing performance on four datasets (ToxiGen, Civil Comments, HateXplain, and SBIC; introduced below), we aim to identify areas where APIs over- or under-moderate specific groups and linguistic variations. This directly addresses **RQ1** and **RQ2** by exploring how these moderation systems function across datasets and target-groups.

In Experiment 5.2, we test the robustness of each API's classification decisions with Perturbation Sensitivity Analysis. By introducing minor changes to input text that should not affect classification, we evaluate the over-and under-moderation of the target group by the API in relation to the dominant groups. Biases in the output confidence scores would not only indicate over- or under-moderation but also provide insights into the shortcomings of these models. Therefore, this experiment addresses **RQ2** and **RQ3** by examining the failures in API functionality.

SHAP in Experiment 5.3 provides an explanation about the contribution of tokens to the APIs confidence scores. We assess these contributions together with a qualitative evaluation of model failures, which helps us understand the specific causes of over- and under-moderation. By providing an interpretable visualisation of how the models weigh different features, this experiment addresses **RQ3** by offering a deeper understanding of the APIs' internal workings.

We use four benchmark datasets, each offering specific linguistic variations of hate speech:

[7]Code: https://anonymous.4open.science/r/Content-Moderation-API-Pipeline-874B/

- **Civil Comments** — Comprised of longer, human-written examples of hate speech, this dataset averages 48.3 words per sequence and contains annotations from multiple human annotators [44].
- **HateXplain** — This dataset includes human-written hate speech examples specifically labeled for target groups by at last three annotators and accompanying hate speech rationales [59].
- **Social Bias Inference Corpus (SBIC)** — Focusing on implicit hate speech and stereotypes, SBIC offers another lens for examining hate speech detection [90].
- **ToxiGen** — This dataset contains implicit and adversarial hate speech, using synthetic data (genAI) to diversify hate speech corpora [40].

The diversity in these datasets ensures that we capture both explicit and implicit forms of hate speech, and the variety of target groups and text types allows us to assess the potential limitations of the content moderation APIs. Table 8 in the Appendix offers a descriptive overview of the four datasets. All datasets are balanced on hate and non-hate speech, both at the aggregate and the group-level, to avoid distortion of performance metrics.

Conducting these experiments was resource-intensive due to the high number of API calls, some limited to one query per second. This amounted to 5 million queries, which had to be parallelized for some APIs. For instance, SHAP analyses of the Perspective API and Microsoft Content Moderator required two weeks of continuous requests. Experiments 1 and 2 were conducted between March 3, 2024, and April 15, 2024, and later repeated from June 15, 2024, to July 11, 2024, yielding consistent results. Experiment 3 was performed from July 12, 2024, to August 18, 2024. We acknowledge that these content moderation systems are dynamic, and their performance may vary over time. However, our framework provides a reusable methodology for re-assessing these APIs as they evolve.

5 Experiments

5.1 Experiment 1: Comparative Aggregate and Target-Group Performance Evaluations on Four Data Sets

5.1.1 Method. We evaluate all cloud-based content moderation algorithms across multiple datasets using both threshold-variant and threshold-invariant performance metrics established in the ML field [11, 69]. A scale-variant metric changes depending on the specific decision threshold set in a classification model, affecting metrics like Precision, Recall, and F1 score. In contrast, a threshold-agnostic metric evaluates model performance across all possible thresholds, providing a more comprehensive measure, such as the ROC AUC [25]. These metrics are assessed at the aggregate level and for specific target groups. Specifically, we compute metrics such as the F1 score, True Positive Rate (TPR), False Positive Rate (FPR), and the threshold-invariant ROC AUC Score. A key part of this evaluation is measuring these metrics not only in aggregate but also at the group level, with the aim of identifying any performance disparities that might affect certain groups. This analysis is grounded in the principle of Equality of Odds, as theorized by Hardt et al. [38] and implemented by Dixon et al. [23], which seeks to ensure that models perform consistently across all groups.

At the group level, we use a pinned ROC AUC, a metric introduced by Dixon et al. [23], to allow for more robust, scale-invariant comparisons across sub-groups. This metric works by "pinning" the data for each sub-group to the same baseline distribution, creating a dataset with 50% of the data belonging to the sub-group and 50% randomly drawn from the overall dataset. The authors acknowledge certain limitations with this approach in later research. However, it remains the most reliable scale-invariant metric for addressing group-level performance variation [11].

As noted by Fortuna et al. [30], the definitions and sub-categories of hate speech vary widely across datasets and services. These definitions do not always align with the sub-category titles, and some services fail to provide clear definitions for their categories (see Appendix Section A.2). To ensure fair and consistent comparisons, we follow a three-step process: (1) gather all available category values for each API, (2) determine which values best align with each dataset's outcomes, and (3) select the maximum value for each API to optimize comparison accuracy.

5.1.2 Results.

Aggregate-Level Performance. Table 2 presents the aggregated performance results for selected benchmark datasets. The term "aggregated" refers to a representative sample of the entire dataset, rather than focusing on specific hate-targeted groups. Our findings reveal significant performance variations across moderation APIs and datasets. Overall, Amazon Comprehend and OpenAI Moderation demonstrate the most consistent performance across the datasets. OpenAI's content moderation algorithm performs best on ToxiGen and HateXplain, generalizing well across different datasets. For Civil Comments and SBIC, Amazon Comprehend yields the most accurate results, although Perspective API also performs comparatively well on Civil Comments. In contrast, Google Natural Language API consistently exhibits the lowest performance, primarily due to a relatively high FPR, suggesting an over-moderation tendency.

We also observe notable cross-dataset performance variations. Civil Comments and HateXplain exhibit similar performance ranges, while ToxiGen and SBIC show somewhat lower results. In general, FPR is elevated across all explicit hate speech datasets, whereas the FNR is particularly high for implicit hate speech datasets. This discrepancy is especially evident with the higher FNR observed across moderation services on ToxiGen and SBIC. Interestingly, the longer text sequences in Civil Comments appear to enhance the performance of all APIs, whereas performance declines on HateXplain. Additionally, AUC scores tend to be higher than accuracy (ACC) scores, indicating that threshold settings may be contributing to misclassification, an important finding that we will discuss later.

Group-Level Performance. Table 3 performance metrics at the group level, highlighting noticeable disparities across moderation services and marginalized groups. Group *Female* consistently received the most accurate moderation across datasets, evidenced by high Pinned ROC AUC values, followed by group *Black*. In contrast, groups *LGBTQIA+*, *Disability*, and *Jewish* experience significantly

Dataset	Characteristics	Moderation Service	ACC	AUC	F1	FPR	FNR
ToxiGen [40]	Implicit / Synthetic: 7,800 samples	Amazon	70.4%	79.2%	68.9%	7.2%	51.9%
		Google	62.7%	65.0%	62.7%	39.1%	35.5%
		OpenAI	70.3%	87.2%	68.1%	5.6%	33.2%
		Microsoft	59.8%	67.0%	57.4%	16.4%	63.9%
		Perspective	61.6%	82.9%	55.5%	1.2%	75.4%
Civil Comments [11]	Explicit / Long: 50,000 samples	Amazon	92.2%	97.4%	92.2%	7.5%	8.1%
		Google	69.9%	67.4%	67.2%	58.4%	1.8%
		OpenAI	78.6%	86.9%	78.6%	17.1%	25.6%
		Microsoft	75.8%	81.7%	75.7%	20.4%	28.1%
		Perspective	87.8%	97.2%	87.7%	3.3%	20.9%
HateXplain [59]	Explicit / Short 14,000 samples	Amazon	70.2%	75.1%	69.9%	44.2%	20%
		Google	66.4%	52.2%	58.9%	76.8%	4%
		OpenAI	78%	87.3%	77.2%	41.1%	8.86%
		Microsoft	69%	66.5%	66.6%	61.2%	10.3%
		Perspective	70.6%	70.2%	75.1%	42.3 %	20.6%
SBIC [90]	Implicit / Real: 33,000 samples	Amazon	80.1%	72.7%	80.6%	12.9%	25.1 %
		Google	64.4%	66.2%	63.4%	50.34%	22.39%
		OpenAI	67.4%	81.1%	65.9%	9%	53.4%
		Microsoft	62.7%	68.1%	62.7%	32%	42%
		Perspective	60.1%	73.9%	60%	21.8%	55%

Table 2: Performance metrics by moderation service and dataset. Blue shading signals the best performance, while red shading indicates the worst performance. All datasets are balanced on toxic and non-toxic phrases.

less accurate moderation, with particularly poor performance observed in specific datasets.

At the service level, Amazon Comprehend exhibits strong, consistent performance across most marginalized groups, often achieving the highest accuracy, followed by OpenAI Content Moderation. Perspective API also ranks among the top services, particularly on the *Jewish* and *LGBTQIA+* groups. Microsoft Azure Content Moderation, however, continues to underperform, aligning with its results in table 2, and shows the weakest performance for several groups, notably *Black* and *Muslim*. Meanwhile, Google Text Moderation struggles considerably, with stark underperformance for groups *Disability* and *Jewish*, especially in datasets like CivilComments and ToxiGen.

In terms of FPR, the analysis reveals that content moderation tends to over-moderate speech related to the *Black* and *LGBTQIA+* communities. For example, on ToxiGen, Google Text Moderation severely over-moderates content targeting the *Jewish* and *Disability* groups, leading to a FPR as high as 99%. Put more tangibly, while the data contains 200 toxic and 200 non-toxic examples for group *Jewish*, Google Text Moderation predicts toxic speech in 385 instances.

On the flip side, FNRs underscore the challenge of detecting implicit hate speech, particularly on ToxiGen and SBIC. However, caution is warranted when interpreting results from ToxiGen, as it is partially synthetically created, as discussed in Section 6.5. Nevertheless, this finding is consistent with SBIC and across most services, the *Disability* group experiences significant under-moderation, with hate speech frequently going undetected. Additionally, groups like *Asian* and *Latinx* appear to be more prone to under-moderation, especially when using Microsoft Azure Content Moderation.

In summary, Amazon Comprehend and Perspective API show the most balanced performance across marginalized groups, while Google Text Moderation and Microsoft Azure exhibit significant inconsistencies, particularly for underrepresented and vulnerable groups like *Disability* and *Jewish*.

5.2 Experiment 2: Perturbation Analysis with Counterfactual Fairness

5.2.1 Method. Perturbation Sensitivity Analysis (PSA) offers an additional, arguably more robust evaluation of group-level biases by using counterfactual fairness evaluation [80]. With the target-group performance evaluation, we derive conclusions about group-level biases from variations in False Negative and False Positive Rates. This grounds on two assumptions. First, the validity of the ground truth in the data. A moderation service would appear to over-moderate or under-moderate if this over- or under-moderations is actually encapsulated in our benchmark datasets that we use for evaluation. Second, we assume that data on all groups comes from the same underlying distributions. If one group were to contain many more edge cases than other groups, this would aggravate False Positive and FNRs, compromising our ability to infer biases.

PSA avoids such pitfalls by solely exchanging group-specific identity tokens while holding the remainder of the sentence constant [81]. Table 10 displays the fundamental logic. We follow prior research in defining an anchor group against which other groups are compared [81]. Using the dominant group as baseline, Counterfactual Token Fairness scores are computed as the difference in hate speech between the baseline and the corresponding marginalized group.

	Asian	*Black*	*Disabled*	*Female*	*Jewish*	*Latinx*	*LGBTQIA+*	*Muslim*
CivilComments								
Amazon	**92%**	**86%**	**96%**	**95%**	**91%**	**95%**	**85%**	**89%**
Google	72%	66%	45%	61%	58%	73%	60%	57%
Microsoft	70%	58%	67%	73%	67%	64%	60%	63%
OpenAI	80%	73%	**83%**	79%	77%	**87%**	72 %	74%
Perspective API	**91%**	**87%**	**95%**	**94%**	**93%**	98 %	**86%**	**89%**
HateXplain								
Amazon	70%	**80%**	—	**86%**	**83%**	—	90%	**88%**
Google	**82%**	68%	—	68%	31%	—	**83** %	42%
Microsoft	**81%**	72%	—	**76%**	78%	—	75 %	78%
OpenAI	**84%**	**88%**	—	**86%**	**93%**	—	**95%**	**93%**
Perspective API	76%	**77%**	—	**89%**	**90%**	—	**89%**	**90%**
ToxiGen								
Amazon	**80%**	68%	**80%**	**86%**	77%	71%	69 %	**75%**
Google	75%	66%	41%	76%	49%	73%	57%	50%
Microsoft	65%	66%	68%	71%	64%	57%	49%	61%
OpenAI	**86%**	**87%**	**93%**	**82%**	**86%**	**86%**	**86%**	**85%**
Perspective API	**85%**	75%	**82%**	**90%**	**81%**	72%	73%	**86%**

Table 3: Pinned ROC AUC is presented per moderation service, dataset and marginalized group. ToxiGen includes 4,268 observations, HateXplain includes 1,748, Civil Comments consists of 19,228 observations, and SBIC is comprised of 5,806. For Civil Comments, target groups and hate labels were coded by human annotators. For ToxiGen we constrain our analysis to a subset of 10,000 observations, which were annotated by human reviewers. For both Civil Comments and ToxiGen, when running aggregate-level tests, we only include phrases for which all annotators are assigned the same hate label.

PSA makes two assumptions: (1) counterfactual pairs should convey the same or neutral meaning, avoiding any implicit biases or derogatory connotations. While constructing toxic counterfactuals is theoretically possible, it is methodologically demanding and exceeds the scope of this project. Instead, we construct 34 *neutral* counterfactual pairs. Importantly, each marginalized group is represented by multiple tokens, reflecting its different semantic representations. For instance, the marginalized group *female* also manifests as *woman* and *women*. Furthermore, (2) there should be no unique interactions between a particular marginalized token and the context of the sentence that would skew the analysis. This is challenging in real-world applications, as certain combinations might evoke stereotypes or specific cultural connotations. Thus, the project uses data consisting largely of short and explicit statements. Furthermore, CTF scores are calculated separately for toxic and non-toxic statements, with the latter generally supporting the assumption of counterfactual symmetry more consistently.

PSA experiments are conducted using two distinct datasets. First, the synthetic *Identity Phrase Templates* from Dixon et al. [23] are used. The set contains 77,000 synthetic examples of which 50% are toxic. These avoid stereotypes and complex sentence structures by design, which ensures that the symmetric counterfactual assumption is met. Mapping the dataset, which contains a broader set of identities, to the 34 tokens for marginalized identities which fit to the eight relevant marginalized groups that we use in study results in 25,738 sentence pairs (see Table 11). Second, by applying the same logic, 9,190 sentence pairs are derived from the HateXplain dataset. Specifically, we identified occurrences of these 34 tokens for marginalized identities in HateXplain sentences and replaced them with counterfactual dominant tokens. Statistics about these sentences – synthetic and non-synthetic – are presented in Table 9.

5.2.2 Results. Figure 3 displays PSA results on the synthetic and non-synthetic datasets, created from the Dixon et al. [23] Sentence Templates and HateXplain respectively. A negative CTF score indicates that, on average, tokens for marginalized identities are associated with higher hate speech scores than their counterfactual dominant tokens. Upon general inspection, two observations arise. First, differences in confidence scores by and large are more pronounced on non-hate speech than hate speech data. Intuitively this makes sense, as scores are generated non-linearly with a definite upper bound. Thus, when other elements in a sentence induce high confidence score, the marginal effect from identity tokens is comparably lower. Second, we notice greater variation in the mean Counterfactual Token Fairness scores in non-synthetic than in synthetic data. This was to be expected, as the sentences from HateXplain contain more contextual information which interacts with the tokens. Depending on the context, the difference in hate speech scores between majority and marginalized token thus varies to a greater extent. By contrast, within non-synthetic examples, the greater variation of CTF mean scores associated with non-toxic data is likely an artifact of the former's smaller sample size. This observation in itself is revealing: When analyzing HateXplain,

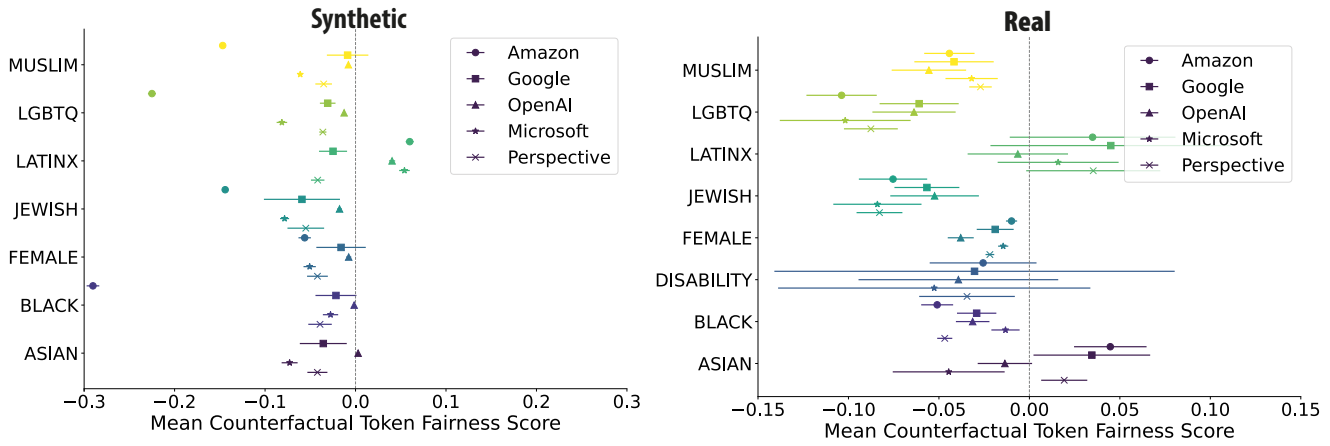


Figure 3: Perturbation Sensitivity Analysis on synthetic data from the Identity Phrase Templates in Dixon et al. [23] and non-synthetic data from HateXplain. Counterfactual Token Fairness (CTF) scores are computed as the difference in toxicity between the phrase containing the baseline dominant token and its marginalized perturbation. Counterfactual Token Fairness scores per marginalized group and service are averaged and reported for non-toxic . Besides a point estimate, the figure also includes a 95% confidence interval assuming a student-t distribution.

we found that 85% of the sentences containing neutral tokens for marginalized identities from our PSA were labeled as hate speech.

Overall, the results suggest that most minorities get associated with higher levels of hate speech scores than dominant majorities, although these effects appear relatively small for some groups, and vary in general across groups and services. Group *LGBTQIA+* is associated with the strongest negative bias, occurring for all samples and services. By contrast, we observe limited negative bias against groups *Latinx* and *Asian*. In fact, in most samples and services, tokens for marginalized identities related to group *Latinx* incur, on average, lower toxicity scores than their *White* counterfactual. Similarly, groups *Female* and *Black* are linked to comparably small negative biases. Unfortunately, our results allow for no robust inference for group *Disability*. While group *Disability* is entirely absent from the synthetic data, as no examples are included in the Dixon et al. [23] Sentence Templates, the non-synthetic data only contains very few instances, resulting in large variation for mean CTF scores.

Other patterns emerge that are more confined to particular services. For instance, Amazon Comprehend is associated with the overall strongest negative bias, linked to groups *Muslim* and *Jewish*. Notably, when applied to hate speech data, this effect vanishes. Interestingly, for Google we are presented with findings that are somewhat difficult to reconcile with the previous section's result. More specifically, we would have expected a large negative bias placed by Google Text Moderation onto group *Muslim*, which does not seem to be the case. Potentially, the extreme FPRs were not driven by *Muslim* identity terms themselves, but rather by other words co-occurring such as *Islam* or paraphrases of identity-related terms. However, we find that Google Text Moderation as higher variance also for the synthetic data set. Further research is required to explain the disparity observed. Lastly, it is worth noting that OpenAI Content Moderation seems the least biased, across groups and datasets.

5.3 Experiment 3: SHAP Explainability and Qualitative Evaluation

5.3.1 Method. To understand where APIs fail to classify hate speech and to get insights into the causes of these failures, we extend our analysis with SHAP [53]. The method is based on so-called Shapley values, which are derived from cooperative game theory and give insight into the feature-contribution to the output [88] and make the inner mechanisms of the model understandable to humans [72]. We selected SHAP since it has the advantage of having an off-the-shelf available NLP text evaluation [66], being one of the black-box access methods that can be used without access to weights or gradients [55], is theoretically grounded and, in contrary to LIME, is not susceptible to instability due to the non-random selection of perturbations [55, 66]. Furthermore, SHAP can be used for local and global explanations and is also model-agnostic [66]. However, the disadvantage of SHAP – especially in the NLP case – is that the calculation is NP-hard [52] and, thus, computationally expensive.

We, therefore, had to use two datasets, namely, HateXplain and ToxiGen, to understand where the APIs over- and under-moderate and the reason for not analyzing the whole data set but the false positives and negatives. This alone led to more than one million queries for all APIs and, thus, a lot of computational resources. We, however, tried to minimize these resources by parallelizing queries of queries dependent on rate limits per minute of each API. We represent features x as word tokens $x = \{w_1, w_2, \ldots, w_n\}$ to ensure interpretability. SHAP computes the marginal contribution of each token w_i within x to the model output $f(x)$, where $f : \mathcal{X} \rightarrow [0, 1]$ and whereby assesses each token's impact on the decision $\mathbb{P}(y = 1 \mid x) > 0.5$, indicating whether the probability of the positive class exceeds 0.5 based on $f(x)$ [66]. A BERT-based tokenizer was

API	FN/FP	Rate	ToxiGen Tokens	Rate	HateXplain Tokens
Amazon	FN	51.9%	*steel, waiting, brave, play, psychology*	20%	*source, history, corner, welcome, think*
	FP	7.2%	*dumb, asshole, homosexuality, caucasian, rites*	44.2%	*coward, uts, porn, penis, rats*
OpenAI	FN	33.2%	*lucky, between, send, health, natural*	8.9%	*words, penguin, phobia, text, international*
	FP	5.6%	*gay, kill, chinese, muslim, christians*	41.1%	*ded, traitor, n*gger, n*gga, jews*
Google	FN	35.5%	*lucky, chapter, playing, character, openend*	4%	*cat, feminine, asian, girls, illo*
	FP	39.1%	*weapon, mormon, drunk, hindu, bullying*	76.8%	*tax, christians, gay, slut, raped*
Microsoft	FN	63.9%	*ist, ple, unce, health, episode*	10.3%	*ro, st, lings, bathroom, ists*
	FP	16.4%	*dumb, crap, child, breast, holy*	61.2%	*bitch, gay, loser, violence, arabic*
Perspective	FN	75.4%	*sau, lucky, define, protect, greatest*	20.6%	*salute, rights, statistics, door, season*
	FP	1.2%	*terrorist, n*gger, queer, black, are*	42.3%	*gay, fucked, loser, coward, balls*

Table 4: ToxiGen (left) and HateXplain (right) five most contributing Global SHAP Values for False Positives (FP) and False Negatives (FN) per API. For example, identity tokens like *gay* and *jews* contribute to OpenAI misclassifying non-hate speech as hate speech (FP), while positive words like *lucky* and *health* contribute to misclassifying hate speech as non-hate speech (FN).

used to retrieve the tokens for each misclassified sentence. Shapley values were calculated through model-agnostic SHAP by using a "[MASK]" token as a perturbation.

We conduct separate analyzes for FP and FN due to the different handling of Shapley values. The process results in a SHAP dataset that includes local explanations, which are the Shapley values (s_i) for each token w_i within x for each sentence x_j in the aforementioned data. Further, we compute global explanations [65] by clustering via DBSCAN similar tokens via cosine similarity with a threshold of 0.9, averaging SHAP values within each cluster. This is done to obtain average results for words such as "women," "woman," and "female." This yields a sorted list of token clusters, providing global explanations per API for both false positives and false negatives across the datasets. These tokens can give us an explanation of which kind of tokens contribute most to the model's decisions.

To provide a comprehensive analysis addressing **RQ3**, we extended the global SHAP analysis with a qualitative evaluation of model failures. Local explanations were visualized for interpretation (see Figure 4) and qualitatively assessed by reviewing the corresponding sentences.

In Section 2.4, we examined potential functional failures of off-the-shelf models, focusing on over-moderation (false positives, FPs) and under-moderation (false negatives, FNs). These failures are rooted in the linguistic and contrastive non-hate variations outlined in Section 2.1. We developed a codebook deductively, drawing on related work that identified specific model failures in off-the-shelf models. The codebooks for over-moderation and under-moderation are presented in Tables 14 and 15, respectively. Additional codes, such as "HATE" and "UNSURE" for FPs, and "NO HATE" and "UNSURE" for FNs, were included to capture disagreements with dataset labels or uncertainty in coding.

The conceptualization of hate speech from Section 2.1 was applied to code sentences, and model failures were analyzed based on SHAP value visualizations. SHAP values were used to interpret model decisions and guide coding. For example, if a SHAP value indicated that a negation (e.g., "not") contributed significantly to the model's decision to classify a sentence as a FN, this was categorized as **Negation-FN**. Another example involves SOS bias: if only the target group identifier in a sentence was highlighted as contributing to the model's classification of hate speech (red), but the sentence was otherwise neutral and classified as a FP, this was categorized as **SOS bias**.

We randomly sampled 5% of FPs and FNs from ToxiGen and HateXplain datasets, stratified across datasets, resulting in a total of 928 FPs and 1086 FNs for coding. The stratification ensured that the varying proportions of FPs and FNs across datasets were reflected in the data. The coding process was conducted collaboratively by the first and fourth authors. Initially, 10% of the samples were coded together to discuss potential issues and refine the codebook. Subsequently, in a second round, an additional 10% of the samples were double-coded independently. Disagreements were reconciled by consensus, focusing on clarifying the definitions of codes, determining when to classify a sentence as "UNSURE," and standardizing the use of SHAP values to guide decisions. In the third round, another 10% of samples were coded independently, discussed, and compared, resulting in substantial agreement ($\kappa = 0.61$ for FPs and $\kappa = 0.59$ for FNs). To compute Cohen's Kappa [17], entries coded as "UNSURE" by one coder were excluded from the analysis but not from the discussion. The remaining 70% of the samples were coded independently by the two authors, with 35% coded by the first author and 35% by the fourth author.

5.3.2 Results. Table 4 shows the results of the global SHAP analysis. The five most contributing word token clusters are ordered from left to right and assigned to each API's FN and FP for both datasets. What we can observe across APIs is that for FPs of ToxiGen, words like *gay*, *jews*, *hindu*, *disabled*, and *homosexual* are among the most contributing tokens to the misclassifying non-hate speech as hate speech. These tokens fall under the category of identity tokens (see Dixon et al. [23]), meaning tokens signifying the identity of marginalized groups.

Additionally, we find tokens like *israeli*, *chinese*, *mohammed*, and *mentally* that are not identity tokens but are spuriously correlated to these identities. This tendency is even stronger when more token

Over-Moderation (False Positives)

Category	AMZ	GPT	GOO	MIC	PER
Counter-speech	9%	6%	9%	7%	9%
Dialects	7%	7%	1%	6%	5%
Descriptive-FP	16%	12%	35%	9%	16%
OM profanity	2%	3%	4%	3%	6%
Negation-FP	0%	0%	0%	2%	5%
Non-protected	4%	5%	2%	9%	2%
OM slurs	16%	13%	5%	16%	23%
Re-appropriation	6%	10%	0%	4%	0%
SOS	20%	15%	18%	27%	6%
Unsure	10%	13%	17%	9%	11%
Hate	3%	10%	3%	2%	13%
∑ **Codes**	206	358	191	81	92

Under-Moderation (False Negatives)

Category	AMZ	GPT	GOO	MIC	PER
Counter-speech-FN	0%	0%	4%	0%	0%
Descriptive-FN	21%	15%	19%	22%	20%
Implicit hate	23%	19%	24%	28%	24%
Negation-FN	4%	4%	2%	4%	2%
Paraphrased target	15%	12%	8%	9%	8%
Positive Term	4%	6%	9%	8%	8%
Spelling variations	0%	2%	0%	3%	2%
Unsure	13%	18%	18%	12%	12%
No hate	12%	12%	11%	7%	11%
∑ **Codes**	226	312	220	181	147

Table 5: Results of the qualitative evaluation of the SHAP explanations (Experiment 3) are summarized here. The operationalizations of the applied codes are detailed in Tables 14 and 15. The code "Unsure" indicates cases where a coder was uncertain about whether a sentence is hateful, required additional context to make a decision, or could not determine why the model over- or under-moderated.

clusters than five are evaluated. For instance, among the first 50 tokens, 26 are identity tokens for Google Natural Language API.

Similar to ToxiGen, we find that HateXplain tokens demonstrate a tendency to rely on identity tokens to classify hate speech across APIs. Words like *jews*, *gays*, and *islam* in the false positive clusters signify this reliance.

Words such as *n*gga* are also included in the FP clusters, reflecting a broader issue with over-moderation. A closer examination of over-classified sentences containing terms like *n*gga*, *trans*, *queer*, or *d*ke* demonstrates the API's disproportional over-moderation of re-appropriation and counter-speech of APIs. This aligns with findings from our qualitative evaluation of failures, summarized in Table 5, highlighting similar challenges.

For FP, we observe in Table 5 that descriptive statements, neutral statements containing identity tokens (SOS bias), counter-speech, and re-appropriation are among the most over-moderated categories. These patterns are consistent across APIs, with the Perspective API demonstrating fewer issues with descriptive sentences and SOS bias. In contrast, Google and Microsoft models notably struggle with over-moderating descriptive and neutral sentences containing identity tokens.

We provide example sentences of descriptive statements with local SHAP values in Figures 4a for Amazon Comprehend and in Figure 4d for Microsoft Moderators. Notably, these sentences were misclassified as hate speech across all models except the Perspective API. Furthermore, Figure 4e illustrates issues with OpenAI Moderation in interpreting counter-speech, while Figure 4f highlights Amazon Comprehend's challenges in handling re-claimed slurs.

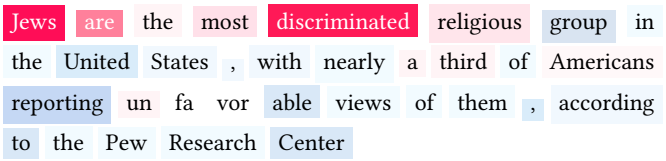
For FN, the tokens leading to misclassification as non-hate speech are semantically ambiguous or challenging to interpret. This is expected, as hate speech in the ToxiGen dataset is often implicit and lacks specific trigger words. The classifiers focus on recurring positive words like *lucky*, *health*, *strong*, *brave*, and *respect*. Here, the APIs seem to have paid too much attention to positive words and failed to understand the implicit message of the sentence, which underlines our analysis in Experiment 5.1.

FNs in HateXplain primarily include general words, with the exception of terms like *asian*, *latin*, and *girl*. The positive connotations attached to *latin* and *asian* in Microsoft and Perspective APIs aligns with Experiment 5.2, where both showed favorable bias toward these groups in real data.

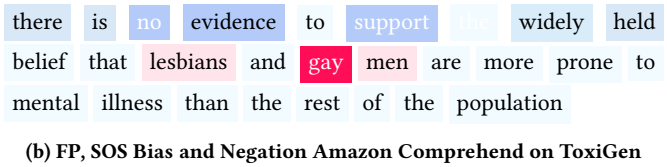
Accordingly, from the qualitative evaluation of FNs using local SHAP sentences, we find that implicit hate speech and the use of positive terms account for approximately 20% and 10% of under-moderations, respectively. Examples of such cases are provided in the Appendix. **Trigger warning:** The examples in the Appendix contain potentially offensive and upsetting language, including racist remarks. They are included solely to illustrate the findings of this research and do not reflect the views of the authors. Figures 5i, 5k, and 5g display implicit hate speech examples, while Figure 5c highlights cases where positive terms influenced predictions.

Additionally, Table 5 highlights that all models exhibit challenges in handling implicit hate speech and descriptive statements. When combined with the FNRs in Table 2, it becomes evident that Perspective API and Microsoft Moderators perform the worst in handling implicit hate speech. In contrast, spelling variations and negations appear to present fewer challenges for the APIs.

Another noteworthy observation is the difficulty commercial systems face with the paraphrasing of identity tokens. For instance, Figure 5j and Figure 5g illustrate instances where models fail to classify hate speech when the target group is referenced using novel identity terms that the systems do not recognize. This issue was



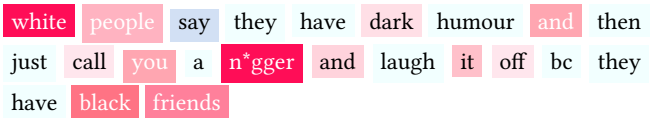
(a) FP, SOS Bias and Descriptive Statement: Amazon Comprehend on ToxiGen



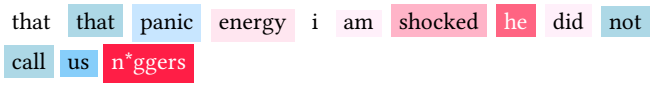
(b) FP, SOS Bias and Negation Amazon Comprehend on ToxiGen

(c) FP, Negation-FP: Microsoft Moderators on ToxiGen

(d) FP, Descriptive statement: Microsoft Moderators for sentence from ToxiGen



(e) FP, Counter-speech: Amazon Comprehend on HateXplain



(f) FP, Re-claimed slurs: OpenAI Moderation API on HateXplain

Figure 4: FP examples: SHAP value visualizations for examples from the ToxiGen and HateXplain datasets using Amazon Comprehend and OpenAI. Red indicates a strong contribution to deciding hate speech; blue indicates a strong contribution to deciding non-hate speech.

particularly evident with Amazon Comprehend, which failed to identify paraphrased tokens such as *k*ke* or *m*zrat* in over 15% of all FNs.

Overall, our findings indicate that the APIs rely disproportionately on the presence of identity tokens to classify hate speech, often leading to over-moderation. Conversely, in instances where identity tokens are absent – such as implicit hate speech or paraphrased forms of identity tokens – the APIs face significant challenges in achieving accurate classification.

6 Discussion

In this paper, we created an audit framework to evaluate black-box content moderation APIs. With this framework, we audited five widely used content moderation APIs. We focused on the comparative performance of selected APIs across different target groups (RQ1), the over- and under-moderations of APIs (RQ2), and the reasons for failures (RQ3). Most importantly, we find that content moderation APIs heavily rely on group identity terms (RQ3). As a result, APIs under-moderate implicit hate speech, which uses codified messages without identity terms, especially against *LGBTQIA+* and *people with disabilities*. Simultaneously, there is a tendency to over-moderate re-appropriation, counter-speech, dialects and descriptive content containing identity terms such as *Blacks*, *Jews*, *Gays*, and *Muslims*.

In the following, we will discuss important findings of over- and under-moderations that developers need to verify and address. Furthermore, we discuss API model transparency, transparency about the API model limitations, and guidance on implementation and application. Subsequently, we give design and research recommendations based on our findings and emphasize on the importance of independent black-box audits and their challenges.

6.1 Over- and Under-moderations

First, a clear and concerning pattern is the tendency of all moderation services to over-moderate speech related to certain marginalized groups, notably *Black*, *LGBTQIA+*, *Muslim*, *Jewish*, and particularly, the subgroup *Gay* within the *LGBTQIA+* umbrella. This over-moderation is especially surprising given the well-established nature of these biases, as demonstrated in prior research that we presented with off-the-shelf NLP systems. Experiments 2 and 3 further confirmed that commercial APIs are disproportionately prone to making misclassifications based on identity tokens and over-moderating counter-speech and re-appropriation. Specifically, our experiments revealed that words like *Gay*, *Lesbian*, *Homosexual*, as well as *Jew*, *Islam*, and *Muslim*, frequently contribute to misclassifications, resulting in inflated FPRs for these groups.

This bias is particularly prominent in groups such as *Gay*, where the mean Counterfactual Token Fairness Score is -0.24, compared to -0.05 for the broader *LGBTQ+* label. It seems that sentences containing terms like *Gay* in training datasets are overwhelmingly labeled as hate speech, a pattern that is reproduced across services. While such biases are well-known, it is surprising that developers apparently have not taken corrective measures to mitigate these effects. Similar concerns apply to the groups *Muslim* and *Jewish*, where these biases are also entrenched, with spurious correlations to identity tokens such as *Israeli* exacerbating over-moderation. The issue is even present in descriptive statements, as the local SHAP explanations are are demonstrating.

The issue extends beyond simple identity tokens. As indicated by the SHAP values, APIs frequently over-moderate content involving terms that, upon closer inspection, involve reappropriation, counter-speech, or are examples of AAE, as seen with terms like *n*gger* and others like *d*ke*, *trans*, and *queer*. These patterns of over-moderation not only suppress legitimate and even empowering speech but also

reflect a failure to account for the nuanced ways marginalized groups engage with and reclaim harmful language.

Moreover, the severe over-moderation of descriptive comments by Google Text Moderation is particularly alarming. Despite robust testing across multiple configurations, this service exhibits a consistently high FPR, especially for groups such as *Disability* and *Jewish*, with an FPR reaching 99% for ToxiGen. While this might be a deliberate design choice to prioritize caution, particularly in scenarios involving human moderation, this approach could fail with automation bias, where human moderators are prone to overly trusting algorithmic recommendations [49]. In fully automated deployments, such over-moderation could lead to unjust content removal, warranting a recalibration of Google's moderation system. We strongly oppose the use of automated deployments without safeguards and urge Google to recalibrate its content moderation model.

Systematic under-moderation is another significant issue, particularly for groups such as *Disability*, *Asian*, and *Latinx*. These groups frequently receive inadequate protection from hate speech, as reflected in high FNRs. Our findings align with previous work on off-the-shelf models suggesting that hate speech targeting these groups is underrepresented in training datasets, resulting in their comparatively lower detection rates [33]. For instance, the *Latinx* and *Asian* groups showed a positive bias in CTF scores, supporting the theory that toxic examples targeting these communities are insufficiently represented during model training. As a result, hate speech against these groups often slips through undetected.

Additionally, implicit hate speech, which tends to be more subtle and context-dependent, remains a persistent challenge across all services. The high FNRs on ToxiGen and SBIC highlight this struggle and was confirmed by the qualitative evaluation of SHAP explanations. This is particularly the case for Perspective API and Microsoft Moderators that show most problems with implicit hate speech with high FNR for these datasets and in qualitative evaluations. As previous research suggests, this issue arises from the lack of comprehensive implicit hate speech datasets available for training purposes [40]. Thus, linguistic variations such as implicit hate speech and contrastive non-hate speech are difficult to detect across commercial systems. The conducted global SHAP analysis and qualitative evaluations further demonstrated that these models disproportionately rely on identity tokens to flag hate speech, which undermines their ability to differentiate between harmful and benign content involving these groups.

6.2 Transparency of Capabilities, User Trust, and Application Guidance

These findings are particularly concerning when APIs exhibit both over- and under-moderation alongside limited transparency regarding these behaviors and the capabilities of the APIs themselves. As outlined in Table 6 , categorized by model card categories [64], most content moderation API services provide minimal information about their underlying models, training algorithms, or fairness evaluations, leaving deployers without sufficient insights to properly assess the suitability and reliability of these systems.

Transparency and guidance on the implementation of ML systems can significantly enhance user trust [50]. While documentation is available for all APIs as demonstrated in Table 6, information such as the model used, version updates, and detailed descriptions of the models' operationalization of hate speech are missing for Amazon Comprehend, Microsoft Azure Content Moderators, and OpenAI Moderation. This lack of transparency can lead to significant challenges for deployers trying to evaluate the appropriateness of a model for their specific use case and to implement it in their organizational processes. Equally concerning is the limited disclosure of the models' boundaries and limitations. As we observed in the over- and under-moderation tendencies, the models exhibit specific weaknesses, such as over-moderation of certain identity tokens (e.g., *Gay*, *Jew*, *Muslim*) and under-moderation of groups like *Latinx* and *Asian*. However, API developers have not sufficiently documented these capability limitations. We showed that Models often fail to detect implicit hate speech or nuanced content, such as satire or counter-speech, and there is no transparency regarding the model's ability to handle these more complex forms of hate speech.

Platforms may be reluctant to disclose moderation mechanisms for fear of revealing trade secrets or aiding future moderation evasion, as Schafner et al. [93] points out. However, there is a working counter-example to that. Perspective API stands out in terms of providing more insights about its training data origin, tutorials, threshold-setting, and performance metrics. As Rieder and Skop [85] describes, Perspective provides a notable level of transparency that allows researchers, users, and deployers to evaluate potential biases or blind spots within the system. This is achieved through access to some example evaluation data, model cards and tutorials though limitations persist, particularly regarding the lack of transparency about datasets used for training, open-source codes, versions and regular updates of the API, explanation of evaluation data.

Jigsaw's model cards for Perspective report AUC scores of 0.97 to 0.99 on a hold-out test set. This performance aligns with our results on Civil Comments, a dataset also published by Jigsaw. However, the AUC scores across the other three datasets differ significantly, particularly for implicit hate speech. This underscores the importance of evaluating models on multiple datasets that capture diverse conceptualizations of hate speech. Furthermore, it highlights the need for transparent reporting of evaluation procedures to ensure reproducibility and prevent data leakage in benchmarks, which can lead to overly optimistic results [47].

Google Natural Language API, however, disclaims that "The confidence scores are only predictions. You should not depend on the scores for reliability or accuracy. Google is not responsible for interpreting or using these scores for business decisions". It highlights a lack of accountability for applying the models in practice [16]. This absence of responsibility leaves users without the necessary guidance on properly deploying and interpreting the models' outputs. It places the burden entirely on the user to assess the model's performance and reliability without providing the tools or frameworks to do so effectively. Similarly, OpenAI only gives limited guidance on how to implement its content moderation API. The phrase "For higher accuracy, try splitting long pieces of text into smaller chunks each less than 2,000 characters", although giving some guidance

on implementation, demonstrates the API's limitations when it comes to the sociotechnical complexity of the task. There is no explanation on how a deployer should handle cases where one chunk of text might be flagged as hate speech while another might not. More specific use cases and best practice examples are essential to minimize the risks of over- and under-moderation.

6.3 Design and Research Recommendations

We argue, alongside Schafner et al. [93], that increased transparency in moderation strategies is vital for building user trust in content moderation. However, we go further in demanding that there must be increased guidance on the implementation of content moderation APIs, as well as greater transparency through tools like model cards. This should include clear decision-making criteria and openness about the APIs' limitations and capabilities to ensure responsible and effective deployment to build users' trust.

Systematic over- and under-moderations can cause users to leave the platform and reduce counter-speech. That said, we also acknowledge the inherent tensions in balancing transparency with the need to protect the integrity of content moderation and safety. Nonetheless, shedding more light on how content moderation APIs work will ultimately enhance accountability and fairness in AI-driven content moderation.

In line with our findings, such as the high FPRs for explicit hate speech datasets as well as general FPRs for Google Natural Language API, deployers and users need clear guidance on when and where these models are prone to failure. The most proactive step API developers could take is to provide better guidance for users. For example, deployers should be informed if a model was trained on explicit forms of hate speech but not on implicit forms, so that they can take necessary precautions. Currently, Perspective API offers some guidance on intended use and user groups, as well as detailed recommendations on how to interpret the model's outputs, such as the distinction between a score of 0.7 versus 0.9 for toxicity. However, most other services leave users in the dark about how to properly implement and interpret the models' results.

Threshold setting, in particular, is an area where more detailed guidance is necessary. As seen in Table 6, while Perspective API provides some guidance, other services place the responsibility entirely on the deployers. The danger of this approach lies in misinterpretations and improper use, especially in contexts where automation bias can lead to over-reliance on model outputs. Human moderators may default to the model's recommendations, even if they are prone to over-moderation [79]. Providing detailed tutorials, thresholds for specific use cases, and warnings about the model's limitations would improve fairness, accountability and, subsequently, user trust in content moderation systems. An important question that arises is whether these services should be held liable when purchased for commercial use, particularly when they result in systematic misclassifications without transparent mechanisms or guidance for proper deployment.

Future research on ML-based hate speech detection should research if over- and under-moderation can be prevented by giving models more context. While datasets like SBIC include categories like target-group and implied statements, our analysis on misclassifications suggests that context could improve classification. Incorporating sender features [67], contextual information [82], and/or both [73] has shown to improve moderation performance. This should include the sender, receiver, target-group, and situational details like the topic of the forum or blog post in response.

6.4 Independent and Black-box Audits

This paper introduces a reproducible framework for independent black-box audits. While content moderation APIs like Perspective API offer the potential for cooperative responsibility, as noted by Rieder and Skop [85], the increasing centralization of content moderation infrastructure by dominant platforms presents significant risks [37]. Smaller organizations, dependent on these cloud-based APIs, may become vulnerable to shifts in corporate priorities. A decline in public communication around Jigsaw communicated by Rieder and Skop [85] and the limited transparency in how these systems operate underlines the importance of independent audits to maintain accountability and fairness in AI-driven content moderation.

However, conducting such black-box audits still requires access to these systems, which presents its challenges. As researchers must pay for API access, systematic evaluations of ML systems become financially inaccessible for many research communities and civil society organizations [8, 84]. This financial barrier limits the ability to perform large-scale or comprehensive audits, particularly for those without substantial funding. Thus, future research must prioritize developing query-efficient approaches for auditing – minimizing costs while providing sufficient evidence of potentially problematic behavior. Active auditing methods [104], which evaluate high-evidence samples and compare them with expert analyses, offer a promising approach.

Recent political developments, such as the EU Digital Services Act (DSA), signal a shift towards supporting systematic third-party audits. Article 40(4) of the DSA grants vetted researchers the right to access data from Very Large Online Platforms (VLOPs) to investigate systemic risks [29, 39]. However, despite this promising development, content moderation services like Google's Perspective API, which are provided by external entities or spin-offs such as Jigsaw, remain outside the scope of these regulations [1]. While Google, as a Very Large Online Platform, is legally obliged to provide data access for auditing, Jigsaw is not currently subject to the same obligations [28].

This regulatory gap highlights the continued importance of third-party audits in maintaining accountability. Even without full white-box access to the inner workings of these systems, black-box audits are essential to ensuring transparency and accountability in content moderation. Future efforts must focus on developing methods that allow researchers to scrutinize these algorithms.

6.5 Limitations and Future Work

Our audit approach was primarily constrained by black-box access, meaning we lacked insight into the internal workings of the models, such as access to model weights, gradients, or detailed documentation. This limitation hindered our ability to thoroughly analyze where and why specific moderation failures occurred.

Additionally, API costs posed a significant challenge, restricting the number of queries we could perform. Future research should explore more cost-effective strategies for sampling that can still provide meaningful insights into over- and under-moderation, thereby maximizing evidence while minimizing resource usage. At the same time, there is a pressing need for increased transparency and access. API deployers and regulators should consider providing white-box or gray-box access to researchers, allowing for a more sophisticated analysis of model failures and biases [15].

Our audit framework, however, was designed to maximize comparability across APIs, target groups, and linguistic variations while minimizing resource usage. The only exception was the use of SHAP, which required a larger number of queries due to its computational complexity.

Another limitation involves the issue of ground truth and the annotation of datasets, which inevitably influence what is considered hate speech and what is not. Although we mitigated this by only using samples in which at least three annotators agreed on the label, this does not completely resolve the problem. We caution against overinterpreting results based solely on ground truth data [92]. However, our approach to using four diverse benchmark datasets aimed to reduce the dependence on any single conceptualization of hate speech, ensuring that our findings reflect a broad spectrum of hate speech definitions and linguistic variations.

Additionally, ToxiGen, a partially synthetically created and labelled dataset, comes with certain limitations [40]. These include the potential introduction of biases, which we observed during our qualitative evaluation, where there were higher rates of unsure or contrary opinions compared to its gold labels. Nonetheless, ToxiGen remains valuable for studying implicit hate speech due to its scale and the inclusion of many sentences without identity tokens. Importantly, we used four datasets for evaluation, including one real-world dataset, specifically for qualitative assessments, and found similar patterns in the distribution of codings for model failures.

It is important to note that the results of our audit represent a snapshot in time, which may change if the underlying systems are modified. We observed particular weaknesses in models when faced with novel identity tokens, underscoring the importance of our audit framework. This framework enables repeated audits, enhancing the temporal validity of the insights gained [71]. However, this requires transparency from model developers regarding the versions and updates of their APIs. Future research should aim to conduct longitudinal comparisons across different versions of these systems—a direction for which our analyses have established an initial benchmark.

Future research should explore several important areas that our current study could not fully address. One critical direction is the need for intersectional analysis. While we evaluated moderation across various marginalized groups, an intersectional approach would allow for a more nuanced understanding of how multiple, overlapping identities (e.g., race and gender) affect the likelihood of over- or under-moderation. Furthermore, our study was limited to English-language content, but there is a clear need for future work to evaluate content moderation systems across multiple languages, especially as hate speech can manifest differently across cultures and linguistic contexts.

Additionally, sample size was a constraint in our analysis due to the high costs associated with API queries. Future research should seek cost-effective methods to perform large-scale audits that remain robust while using fewer resources. While our focus was on text-based content moderation, future studies should investigate the performance of image, video, and speech moderation APIs, as these modalities are increasingly critical in online platforms but may present different challenges.

7 Conclusion

This study introduced a robust audit framework designed to evaluate black-box content moderation systems across five widely-used commercial APIs. By analyzing 5 million queries sourced from four benchmark datasets, we uncovered significant reliance on group identity terms, such as "Black", to predict hate speech. Although OpenAI Content Moderation and Amazon Comprehend performed slightly better, all providers exhibited clear tendencies to under-moderate implicit hate speech, particularly for groups like *LGBTQIA+*, where codified messages without explicit identity terms often went unnoticed. At the same time, over-moderation was prevalent in explicit content targeting *Blacks*, *Jews*, *Muslims*, and *LGBTQIA+*.

In light of these findings, it is important that content moderation providers recalibrate their models to address both over- and under-moderation issues. This recalibration process should actively involve marginalized communities and NGOs, including continuous communication and collaboration to align moderation strategies with the lived experiences of those most affected. Companies should also test their models against the specific biases identified in our findings and ensure these systems can fairly moderate content for all groups.

Additionally, content moderation providers should provide clear implementation guidance and transparency regarding their models' capabilities, particularly in relation to their limitations when moderating linguistic variations or implicit hate speech. Improved access to model internals – such as weights, gradients, and comprehensive documentation – would also empower researchers, civil society organizations, and journalists to conduct more sophisticated evaluations, helping to uncover and mitigate over- and under-moderations. These recommendations should be considered by policymakers and regulatory bodies to improve accountability and fairness in ML-driven content moderation.

Our findings showed that applying ML without sufficient transparency and oversights can lead to additional challenges for historically marginalized groups. We hope that our audit approach empowers practitioners, researchers, and civic hackers to mitigate these adverse effects.

Acknowledgments

We sincerely thank Hanhee Ra and Cristina Maurillo for their careful proofreading. We also extend our gratitude to Prof. Bettina Berendt, Dr. Milagros Miceli, and the entire research group *Data, Algorithmic Systems, and Ethics* at the Weizenbaum Institute, as well as the research group *Human-Centered Artificial Intelligence* at the Center for Advanced Internet Studies for their valuable insights and support.

References

[1] Sana Ahmad. 2023. Ground Control: Organizing Content Moderation for Social Media Platforms. https://doi.org/10.17169/REFUBIUM-40700

[2] Wafa Alorainy, Pete Burnap, Huan Liu, and Matthew L. Williams. 2019. "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web* 13, 3 (July 2019), 1–26. https://doi.org/10.1145/3324997 Retrieved 2020-01-21.

[3] Luvell Anderson and Michael Barnes. 2023. Hate Speech. In *The Stanford Encyclopedia of Philosophy* (fall 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University, Online. https://plato.stanford.edu/archives/fall2023/entries/hate-speech/

[4] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *ACM Transactions on Computer-Human Interaction* 5, CSCW1 (April 2021), 1–34.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness in Machine Learning: Limitations and Opportunities.* MIT Press, Cambridge, MA. https://fairmlbook.org/

[6] Niklas Barth, Elke Wagner, Philipp Raab, and Björn Wiegärtner. 2023. Contextures of Hate: Towards a Systems Theory of Hate Communication on Social Media Platforms. *The Communication Review* 26, 3 (2023), 209–252. https://doi.org/10.1080/10714421.2023.2208513

[7] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. *Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation.* Springer International Publishing, Oxford, United Kingdom, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

[8] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML).* IEEE, Toronto, ON, Canada, 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

[9] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. https://aclanthology.org/2020.acl-main.485

[10] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1119–1130. https://aclanthology.org/D16-1120

[11] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19).* Association for Computing Machinery, New York, NY, USA, 491–500. https://doi.org/10.1145/3308560.3317593

[12] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, Hong Kong, China, 1664–1674. https://doi.org/10.18653/v1/D19-1176 Retrieved 2020-01-20.

[13] Alexander Brown. 2017. What is Hate Speech? Part 2: Family Resemblances. *Law and Philosophy* 36, 5 (Oct. 2017), 561–613. https://doi.org/10.1007/s10982-017-9300-x

[14] Jie Cai, Aashka Patel, Azadeh Naderi, and Donghee Yvette Wohn. 2024. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)* (Honolulu, HI, USA). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613905.3650882

[15] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24).* Association for Computing Machinery, New York, NY, USA, 2254–2272. https://doi.org/10.1145/3630106.3659037

[16] Google Cloud. 2025. *Moderate Text | Cloud Natural Language API.* Google Cloud. https://cloud.google.com/natural-language/docs/moderating-text

[17] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[18] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11 (03 2023), 300–319. https://doi.org/10.1162/tacl_a_00550 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00550/2075730/tacl_a_00550.pdf

[19] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[20] G De Gregorio. 2020. Democratising online content moderation: A constitutional framework. *Computer Law & Security Review* 36 (2020), 105376.

[21] Daria Dergacheva, Vasilisa Kuznetsova, Rebecca Scharlach, and Christian Katzenbach. 2023. *One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database.* Universität Bremen. https://doi.org/10.26092/ELIB/2707

[22] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (2021), 700–732. Issue 2. https://doi.org/10.1007/s12119-020-09790-w

[23] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).* Association for Computing Machinery, New York, NY, USA, 67–73.

[24] Evelyn Douek. 2021. Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability. *Columbia Law Review* 121, 3 (2021), 759–833.

[25] Fatma Elsafoury, Stamos Katsigiannis, and Naeem Ramzan. 2023. On Bias and Fairness in NLP: How to have a fairer text classification? http://arxiv.org/abs/2305.12829 arXiv:2305.12829 [cs].

[26] Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic Offensive Stereotyping Bias in Word Embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1263–1274. https://aclanthology.org/2022.coling-1.108

[27] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana, Dominican Republic, 2021). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 345–363. https://doi.org/10.18653/v1/2021.emnlp-main.29

[28] European Commission. 2024. Supervision of the designated very large online platforms and search engines under DSA. https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses#ecl-inpage-google Last accessed 2024-04-29.

[29] European Parliament. 2022. Regulation (EU)2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065 Last accessed 2024-04-29.

[30] Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 6786–6794. https://aclanthology.org/2020.lrec-1.838

[31] Adam D Galinsky, Kurt Hugenberg, Carla Groom, and Galen V Bodenhausen. 2003. The reappropriation of stigmatizing labels: Implications for social identity. In *Identity issues in groups.* Vol. 5. Emerald Group Publishing Limited, Leeds, England, 221–256.

[32] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (09 2024), 1097–1179. https://doi.org/10.1162/coli_a_00524 arXiv:https://direct.mit.edu/coli/article-pdf/50/3/1097/2471010/coli_a_00524.pdf

[33] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *Comput. Surveys* 55, 13s (2023), 1–32.

[34] Fasica B. Gebrekidan. 2024. *Content moderation: The harrowing, traumatizing job that left many African data workers with mental health issues and drug dependency.* DAIR Institute. https://data-workers.org/fasica

[35] Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting Cross-Geographic Biases in Toxicity Modeling on Social Media. In

Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). Association for Computational Linguistics, Online, 313–328.

[36] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (2022), 20563051221117552. https://doi.org/10.1177/20563051221117552 arXiv:https://doi.org/10.1177/20563051221117552

[37] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 205395171989794. http://journals.sagepub.com/doi/10.1177/2053951719897945

[38] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[39] David Hartmann, José Renato Laranjeira de Pereira, Chiara Streitbörger, and Bettina Berendt. 2024. *Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society.* AI and Ethics. https://doi.org/10.1007/s43681-024-00595-3

[40] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3309–3326. https://doi.org/10.18653/v1/2022.acl-long.234

[41] Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2024. "Vulnerable, Victimized, and Objectified": Understanding Ableist Hate and Harassment Experienced by Disabled Content Creators on Social Media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, USA). ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3641949

[42] Manoel Horta Ribeiro. 2024. Content Moderation in Online Platforms. https://doi.org/10.5075/EPFL-THESIS-10387

[43] Juliane Jarke and Hendrik Heuer. 2024. *Reassembling the Black Box of Machine Learning: Of Monsters and the Reversibility of Foldings.* Amsterdam University Press, Amsterdam, Netherlands, 103–126. http://www.jstor.org/stable/jj.11895528.7

[44] Jigsaw. 2019. Jigsaw toxic comment classifi- cation challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[45] Jigsaw. 2021. Google's Jigsaw Announces Toxicity-Reducing API Perspective is Processing 500M Requests Daily. https://www.prnewswire.com/news-releases/googles-jigsaw-announces-toxicity-reducing-api-perspective-is-processing-500m-requests-daily-301223600.html

[46] Amba Kak and Sarah Myers West. 2023. *Algorithmic Accountability: Moving Beyond Audits.* AI Now Institute. https://ainowinstitute.org/publication/algorithmic-accountability

[47] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Article Volume 4, Issue 9* 4, 9 (September 08 2023), 100804. https://doi.org/10.1016/j.mlops.2023.100804 Open access.

[48] Ramzi Khezzar, Abdelrahman Moursi, and Zaher Al Aghbari. 2023. arHateDetector: detection of hate speech from standard and dialectal Arabic Tweets. *Discover Internet of Things* 3, 1 (March 2023), 1.

[49] Clara Lachemaier, Eleonore Lumer, Hendrik Buschmeier, and Sina Zarrieß. 2024. Towards Understanding the Entanglement of Human Stereotypes and System Biases in Human-Robot Interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 646–649.

[50] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[51] Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (Eds.). Association for Computational Linguistics, Seattle, Washington (Hybrid), 29–39. https://doi.org/10.18653/v1/2022.woah-1.4

[52] Scott Lundberg. 2018. *An Introduction to Explainable AI with Shapley Values — SHAP Latest Documentation.* Scott Lundberg. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

[53] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[54] Thodoris Lykouris and Wentao Weng. 2024. Learning to Defer in Content Moderation: The Human-AI Interplay. arXiv:2402.12237 [cs.LG] https://arxiv.org/abs/2402.12237

[55] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics* 50, 2 (06 2024), 657–723. https://doi.org/10.1162/coli_a_00511 arXiv:https://direct.mit.edu/coli/article-pdf/50/2/657/2457495/coli_a_00511.pdf

[56] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How Do Users Experience Moderation?: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 278 (oct 2023), 30 pages. https://doi.org/10.1145/3610069

[57] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 660–686. https://doi.org/10.1145/3630106.3658932

[58] Teresa Marques. 2023. The Expression of Hate in Hate Speech. *Journal of Applied Philosophy* 40, 5 (2023), 769–787. https://doi.org/10.1111/japp.12608 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/japp.12608

[59] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14867–14875. https://ojs.aaai.org/index.php/AAAI/article/view/17745 Number: 17.

[60] Mari J. Matsuda, Charles R. Lawrence III, Richard Delgado, and Kimberlé W. Crenshaw. 1993. *Words That Wound: Critical Race Theory, Assaultive Speech, and The First Amendment.* Faculty Books, New York. https://scholarship.law.columbia.edu/books/287 Accessed: date-of-access.

[61] Milagros Miceli, Paola Tubaro, Antonio A. Casilli, Thomas Le Bonniec, and Camilla Salim Wagner. 2024. *Who Trains the Data for European Artificial Intelligence?: Report of the European Microworkers Communication and Outreach Initiative (EnCOre, 2023-2024).* Technical Report. European Parliament; The Left. 1–40 pages.

[62] Helena Mihaljević and Elisabeth Steffen. 2022. How toxic is antisemitism? Potentials and limitations of automated toxicity scoring for antisemitic online content. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis* (2022-09-12). CPSS-2022, Hochschule für Technik und Wirtschaft Berlin, Potsdam, Germany, 1–12.

[63] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. http://arxiv.org/abs/1908.06024 Retrieved 2020-02-04.

[64] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[65] Christoph Molnar. 2022. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book

[66] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4593–4603. https://aclanthology.org/2022.coling-1.406

[67] Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and Interpreting the Impact of User Context in Hate Speech Detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Online, 91–102. https://doi.org/10.18653/v1/2021.socialnlp-1.8

[68] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.

[69] Andreas C Müller and Sarah Guido. 2016. *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.", Delaware, USA.

[70] Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 742, 22 pages. https://doi.org/10.1145/3613904.3642025

[71] Kevin Munger. 2019. The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity. *Social Media + Society* 5, 3 (2019), 2056305119859294. https://doi.org/10.1177/2056305119859294 arXiv:https://doi.org/10.1177/2056305119859294

[72] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1900654116

[73] Shubhanshu Nagar, Faysal A. Barbhuiya, and Koushik Dey. 2023. Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining* 13, 47 (2023), 1–14. https://doi.org/10.1007/s13278-023-01051-6

[74] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 26–34. https://doi.org/10.18653/v1/2023.trustnlp-1.3

[75] Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2025. Toxic Bias: Perspective API Misreads German as More Toxic. In *Proceedings of the 19th AAAI International Conference on Web and Social Media (ICWSM'25)*. AAAI Press, Copenhagen, Denmark, 12–23. Please check and cite the published version of this paper.

[76] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence* (Thessaloniki, Greece) *(WI '19)*. Association for Computing Machinery, New York, NY, USA, 149–155. https://doi.org/10.1145/3350546.3352512

[77] Marie Ozanne, Ameya Bhandari, Natalya N. Bazarova, and Dominic DiFranzo. 2022. Shall AI Moderators Be Made Visible? Perception of Accountability and Trust in Moderation Systems on Social Media Platforms. *Big Data & Society* 9, 2 (2022), 1–13. https://doi.org/10.1177/20539517221115666

[78] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2799–2804. https://doi.org/10.18653/v1/D18-1302

[79] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Technical Report MSR-TR-2022-12. Microsoft. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[80] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5740–5745. https://aclanthology.org/D19-1578

[81] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5740–5745.

[82] Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE Access* 11 (2023), 30575–30590. https://doi.org/10.1109/ACCESS.2023.3258973

[83] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Honolulu HI USA, 429–435.

[84] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. http://arxiv.org/abs/2206.04737 arXiv:2206.04737 [cs].

[85] Bernhard Rieder and Yarden Skop. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 8, 2 (2021), 20539517211046181. https://doi.org/10.1177/20539517211046181 arXiv:https://doi.org/10.1177/20539517211046181

[86] Emma Roth. 2024. *ChatGPT's weekly users have doubled in less than a year*. The Verge. https://www.theverge.com Accessed: 2024-12-10.

[87] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

[88] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The Shapley Value in Machine Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 5572–5579. https://doi.org/10.24963/ijcai.2022/778 Survey Track.

[89] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. https://aclanthology.org/P19-1163

[90] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5477–5490. https://aclanthology.org/2020.acl-main.486

[91] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

[92] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, United States). Association for Computational Linguistics, Seattle, United States, 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

[93] Brennan Schafner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L. Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, USA). ACM, New York, NY, USA, 16. https://doi.org/10.1145/3613904.3642333

[94] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 485, 16 pages. https://doi.org/10.1145/3544548.3581020

[95] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. https://doi.org/10.1145/3544548.3581538

[96] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339

[97] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's Common and a Part of Being a Content Creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA). ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501879

[98] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1667–1682. https://doi.org/10.18653/v1/2021.acl-long.132

[99] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 78–84.

[100] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of*

the NAACL Student Research Workshop, Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[101] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 576–587. https://aclanthology.org/2021.naacl-main.48

[102] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 602–608. https://doi.org/10.18653/v1/N19-1060

[103] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, Lun-Wei Ku and Cheng-Te Li (Eds.). Association for Computational Linguistics, Online, 7–14. https://doi.org/10.18653/v1/2020.socialnlp-1.2

[104] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Baltimore, Maryland, USA, 24929–24962. https://proceedings.mlr.press/v162/yan22c.html

[105] Wenqi Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598. https://doi.org/10.7717/peerj-cs.598

[106] Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, Antske Fokkens and Vivek Srikumar (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 27–39. https://doi.org/10.18653/v1/2022.conll-1.3

[107] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. arXiv:2206.06423 [cs.CL] https://arxiv.org/abs/2206.06423

[108] Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4920–4932. https://doi.org/10.18653/v1/2023.emnlp-main.300

[109] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 3143–3155. https://doi.org/10.18653/v1/2021.eacl-main.274

[110] Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate Speech Detection and Reclaimed Language: Mitigating False Positives and Compounded Discrimination. In *Proceedings of the 16th ACM Web Science Conference* (Stuttgart, Germany) *(WEBSCI '24)*. Association for Computing Machinery, New York, NY, USA, 241–249. https://doi.org/10.1145/3614419.3644025

A Appendix

A.1 Model Overview Documentations

Criteria	Natural Language API	Amazon Comprehend	Content Moderators	Moderation API	Perspective API
Model Details					
Developer	Google	Amazon	Microsoft Azure	OpenAI	Jigsaw (Google)
Documentation available	✓	✓	✓	✓	✓
Model Date	28.08.2023	×	×	01.03.2023	×
Model Version	PaLM 2	×	×	×	×
API Version	v2	2017-11-27	v1.0	text-moderation-001	v1alpha1
Model Type	Fine-tuned language Model	Proprietary NLP model	Proprietary NLP model	Proprietary moderation model	Multilingual BERT-based models
Information about Training Algorithms, Fairness Evaluations, Parameters	×	×	×	×	Training Data origin discussed and model performans provided
Operationalization of Hate Speech	Definitions of subcategories given, no further explanations	×	×	×	The likelihood that a reader interprets the comment provided in the request as embodying the specified attribute.
Intended Use and Usage Guidance/Tutorials					
Primary intended use	×	×	×	×	Provides specific intended uses
Primary intended users	×	×	×	Provided	Inteded usecases linked to inteded users
Information how to embed the Model	×	×	×	Proprietary moderation model	Currently working in including context
Guidance on how to choose tresholds	Defined and tested by the user, Google refuses to take responsibility	×	×	No guidance, but it is mentioned that thresholds have to be carefully selected	Give guidance
Tutorials for implementation	×	×	×	✓	✓
Factors and **Metrics**					
Relevant Factors and Evaluation Factors	×	×	×	×	✓
Model performance measures	×	×	×	✓	✓
Ethical Considerations, Caveats and Recommendations					
Ethical considerations	×	×	×	×	✓
Caveats	×	×	×	×	✓
Recomendations	Difference between probability and severity is explained.	×	×	×	Difference between probability and severity is explained ("score of 0.9 is not necessarily more toxic than a comment with a TOXICITY score of 0.7")

Table 6: Transparency comparison of commercial content moderation API services with a focus on model card categories by Mitchell et al. [64]

A.2 Content Moderation API Configurations

OpenAI	Microsoft	Amazon	Google	Perspective
Harassment	Sexually explicit or adult	Profanity	Death, Harm & Tragedy	Threat
Harassment threatening	Sexually suggestive or mature	Hate speech	Toxic	Insult
Hate	Offensive	Insult	Insult	Toxicity
Hate threatening	Profanity	Graphic	Health	Identity attack
Self harm		Harassment or abuse	Violent	Severe toxicity
Self harm instructions		Sexual	Illicit drugs	Profanity
Self harm intent		Violence or threat	Finance	
Sexual			Profanity	
Sexual minors			Firearms & weapons	
Violence			Religion & belief	
Violence graphic			Legal	
Self-harm			Public safety	
Sexual/minors			Politics	
Hate threatening			Derogatory	
Violence graphic			Sexual	
Self-harm intent			War & conflict	

Table 7: Hate Speech Sub-category Configurations across APIs.

A.3 Audit Sample Size

Group	ToxiGen		Civil Comments		HateXplain	
	N	Avg. Words	N	Avg. Words	N	Avg. Words
Aggregate	7,846	18	50,000	48	50,000	19
Asian	654	17	284	70	1,666	20
Black	360	22	4,230	74	10,126	19
Disability	882	17	1,048	58	898	19
Female	202	16	6,170	72	16,722	19
Jewish	400	20	714	72	898	20
Latinx	470	17	130	63	1,260	20
LGBTQIA+	480	21	2,674	70	484	19
Muslim	820	18	3,978	67	2,730	19

Table 8: Descriptive statistics across ToxiGen, Civil Comments and HateXplain: Sample size and average number of words in a sentence, at the aggregate level and by marginalized group. All datasets are balanced on toxic and non-toxic speech, both at the aggregate and the group-level.

Group	Synthetic (Dixon et al. [23])			Non-Synthetic (HateXplain)		
	N	Share Toxic	Avg. Words	N	Share Toxic	Avg. Words
Aggregate	25,738	0.5	5	9,190	0.85	20
Asian	1,514	0.5	5	122	0.63	19
Black	1,514	0.5	5	1,053	0.71	20
Disability	0	0	0	31	0.45	22
Female	1,514	0.5	5	2,294	0.71	21
Jewish	1,514	0.5	5	953	0.89	20
Latinx	4,542	0.5	5	45	0.76	18
LGBTQIA+	13,626	0.5	5	552	0.76	19.7
Muslim	1,514	0.5	5	4,140	0.97	20

Table 9: Descriptive statistics of datasets for Perturbation Sensitivity Analysis, reporting sample size, share of toxic statements and average words per sentence. The *Identity Phrase Templates* from Dixon et al. [23] were synthetically created for the purpose of Perturbation Sensitivity Analysis, while the non-synthetic dataset is derived from HateXplain. The Identity Phrase Templates contains 1,514 unique sentences which are repeated for all their tokens for marginalized identities. Among the groups defined for the purpose of this project, some map onto several tokens from Dixon et al. [23], resulting in sample sizes in excess of 1,514. While the synthetic dataset was designed to be balanced in terms of toxicity and contains very short sentences, the real dataset is skewed towards toxic speech and is made up of comparable longer phrases.

A.4 Pertubation Sensitivity Analysis with Counterfactual Fairness Tokens and Performance Overview

Marginalized Reference	Toxicity	Dominant Reference	Toxicity	PSA
Being **Asian** is fun	0.21	Being **White** is fun	0.33	**0.12**
Being **Black** is fun	0.31	Being **White** is fun	0.33	**0.02**
Being **Female** is fun	0.16	Being **Male** is fun	0.18	**0.02**
Being **LGBTQIA+** is fun	0.31	Being **Straight** is fun	0.25	**-0.06**
Being **Jewish** is fun	0.40	Being **Christian** is fun	0.36	**-0.04**
Being **Muslim** is fun	0.42	Being **Christian** is fun	0.36	**-0.06**
Being **Latinx** is fun	0.24	Being **White** is fun	0.33	**0.09**

Table 10: In Perturbation Sensitivity Analysis, valid counterfactuals are constructed by solely exchanging marginalized and corresponding dominant tokens in a sentence. Counterfactual Token Fairness scores are computed by taking the difference in toxicity score, revealing any groups-specific biases. In this illustrative example, we report the mean toxicity score across all moderation services. Negative Counterfactual Token Fairness scores indicate potentially negative biases associated with a marginalized's common identity tokens.

Group	Marginalized	Dominant	Group	Marginalized	Dominant
asian	asian	white	**black**	black	white
disability	disabled	abled	**female**	woman	man
disability	disability	ability	female	women	men
disability	disabilities	abilities	female	female	male
disability	handicapped	abled			
jewish	jewish	christian	**latinx**	latino	white man
jewish	jew	christian	latinx	latinos	white men
jewish	jews	christians	latinx	latina	white woman
jewish	judaism	christianity	latinx	latinas	white women
			latinx	latinx	white person
LGBTQIA+	lesbian	straight	**muslim**	muslim	christian
LGBTQIA+	lesbians	straight women	muslim	muslims	christians
LGBTQIA+	gay	straight	muslim	islam	christianity
LGBTQIA+	gays	straight men	muslim	arabic	christian
LGBTQIA+	bi	straight			
LGBTQIA+	bisexual	straight			
LGBTQIA+	trans	straight			
LGBTQIA+	transexual	straight			
LGBTQIA+	queer	straight			
LGBTQIA+	lgbt	straight			
LGBTQIA+	LGBTQIA+	straight			
LGBTQIA+	homosexual	heterosexual			

Table 11: 34 marginalized token across eight marginalized groups are employed in Perturbation Sensitivity Analysis. Per marginalized token we identified the substantively and semantically appropriate dominant token to construct valid counterfactuals.

Group	Marginalized Reference	Dominant Reference
LGBTQIA+	Being **gay** is great	Being **straight** is great
Latinx	Aki is a filthy **latino**	Aki is a filthy **white**
Black	Dishad is a dumb **black**	Dishad is a dumb **white**
Muslim	you are a shitty **muslim**	you are a shitty **christian**
Female	destroy all **women**	destroy all **men**

Table 12: Synthetic data examples for Perturbation Sensitivity Analysis, based on the *Identity Phrase Templates* from Dixon et al. [23]. Semantically, these are very stripped down and short. Per sentence we create a valid counterfactual by solely exchanging the marginalized token for the appropriate dominant token.

Group	Marginalized Reference	Dominant Reference
Muslim	oh well actually this town has large amounts of **muslim** immigration the **muslim** world is ripping the christians of af	oh well actually this town has large amounts of **chrisitan** immigration the **christian** world is ripping the christians of af
Asian	so we went from owning a mud **asian** lady to her having to go back to where she came from	so we went from owning a mud **white** lady to her having to go back to where she came from
Female	Immediately following a presentation by three **women** okay sure they made up the next show	Immediately following a presentation by three **men** okay sure they made up the next show
LGBTQIA+	youre a **transsexual** lol get over yourself	youre a **straight** lol get over yourself
Jewish	the same way **jews** run the government they want to control the masses and impose their cultural values	the same way **christians** run the government they want to control the masses and impose their cultural values

Table 13: Non-synthetic data examples for Perturbation Sensitivity Analysis, based on HateXplain. While still relatively short, this are on average longer than the synthetic examples and involve more contextual information. Per sentence we create a valid counterfactual by solely exchanging the marginalized token for the appropriate dominant token.

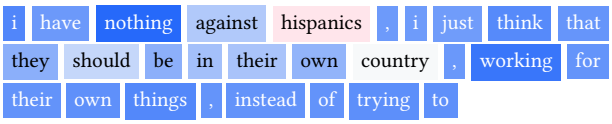
A.5 Qualitative SHAP Analysis Codes of Model Failures

Code	Operationalization	Source
CS: Counter speech	The comment appears to criticize or oppose a hateful statement made by someone else. This includes (1) Denouncements of hate that quote it and (2) denouncements of hate that make direct reference to it	[22, 87, 98]
RE: Re-appropriation, reclaimed slurs	A discriminatory word or phrase is reclaimed and used by a member of the marginalized group in a non-hateful context. This includes (1) Non-hateful homonyms of slurs and (2) Reclaimed slurs	[22, 87, 98]
OS: Over-moderating slurs and profanity	A slur or profanity is used in a non-hateful or exaggerated context but is flagged as hate speech.	[87, 98]
LS: Long Sentences	Statements which are long, often involving multiple clauses and sentences.	[98]
DA: Dialects	The comment uses a dialect (e.g., African American English) that is misclassified as hate speech.	[9, 89, 109]
SOS: Systematic Offensive Stereotyping Bias	A marginalized group is mentioned in a neutral or informational context but is misinterpreted as hate speech.	[25, 87, 109]
NE: Negation	Hate expressed using negated positive statement	[87, 98]
NP: Abuse against non-protected target	The comment targets a group or individual that is not a protected target.	[87]
DE: Descriptive comment or Non-hate group identity	The comment describes hateful topics (e.g., political situations or events) without endorsing them. It may contain misinformation or conspiracy theories but does not qualify as hate speech against a protected group.	[87]

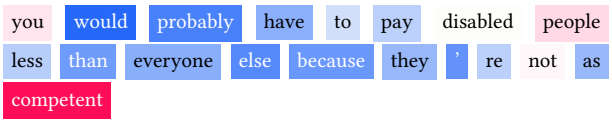
Table 14: Categories for coding false positive sentences

Code	Operationalization	Source
IMP: Implicit hate speech	The comment contains implicit hate speech, which is not overtly expressed but still offensive. This includes (1) jokes and sarcasm, (2)	[90, 99, 101, 102]
POS: Terms with positive sentiment	Statements which contain positive phrases and terms but still express hate.	[98]
NE-FN: Negation	Non-hate expressed using negated hateful statement	[87, 98]
PAR: Paraphrased target	A novel way of describing or paraphrasing of a marginalized group or protected group is used, which the system fails to detect as hateful.	[51]
DE-FN: Misread as description or political statement	A hateful comment is misclassified as a neutral or descriptive statement.	[87]
CS-FN: Misread as Counter speech	A comment that is flagged as non-hate and is detected as counter speech but is hateful in its content.	[22]
SPELL: Spelling Variations	Swaps of adjacent characters, missing characters, missing word boundaries, added spaces between chars and leet speak spellings	[87]

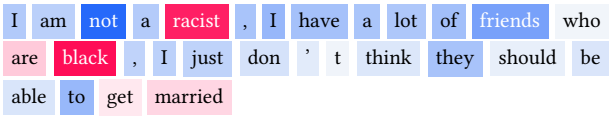
Table 15: Categories for coding false negative sentences



(a) FN, Negation-FN Latinos: Perspective API on ToxiGen

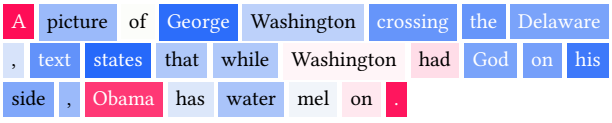


(c) FN, Positive Terms: Perspective on ToxiGen



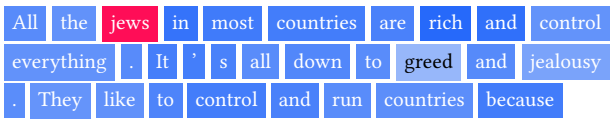
(e) FN, Implicit hate speech and Negation Perspective on ToxiGen

(g) FN, Implicit Bias and Paraphrasing Target Identity Microsoft on HateXplain.

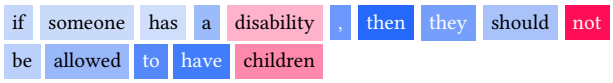


(i) FN, Implicit and Paraphrasing Identity Token Microsoft Moderators on ToxiGen.

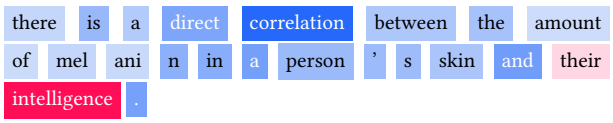
(k) FN, Implicit Perspective on ToxiGen.



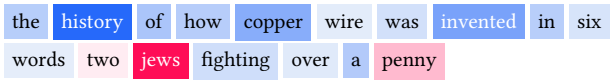
(b) FN, SOS Bias Perspective on ToxiGen



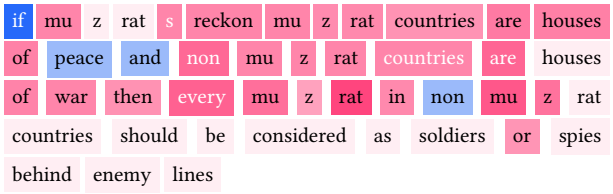
(d) FN, SOS Bias Perspective on ToxiGen



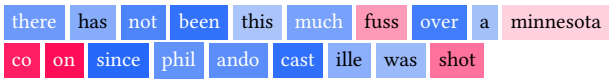
(f) FN: SOS Bias against black people. Perspective for sentence from ToxiGen



(h) FN, Implicit Amazon Comprehend on HateXplain



(j) FN, Paraphrased target and Implicit GPT on HateXplain



(l) FN: Implicit Perspective on HateXplain

Figure 5: FN examples: SHAP value visualizations for examples from the ToxiGen and HateXplain datasets. Red indicates a strong contribution to deciding hate speech; blue indicates a strong contribution to deciding non-hate speech. For visualization, we added some tokens together and averaged contribution of both.