"HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media

LINGYAO LI, School of Information, University of Michigan, Ann Arbor, USA
LIZHOU FAN, School of Information, University of Michigan, Ann Arbor, USA
SHUBHAM ATREJA, School of Information, University of Michigan, Ann Arbor, USA
LIBBY HEMPHILL, School of Information, University of Michigan, Ann Arbor, USA

Harmful textual content is pervasive on social media, poisoning online communities and negatively impacting participation. A common approach to this issue is developing detection models that rely on human annotations. However, the tasks required to build such models expose annotators to harmful and offensive content and may require significant time and cost to complete. Generative AI models have the potential to understand and detect harmful textual content. We used ChatGPT to investigate this potential and compared its performance with MTurker annotations for three frequently discussed concepts related to harmful textual content on social media: Hateful, Offensive, and Toxic (HOT). We designed five prompts to interact with ChatGPT and conducted four experiments eliciting HOT classifications. Our results show that ChatGPT can achieve an accuracy of approximately 80% when compared to MTurker annotations. Specifically, the model displays a more consistent classification for non-HOT comments than HOT comments compared to human annotations. Our findings also suggest that ChatGPT classifications align with the provided HOT definitions. However, ChatGPT classifies "hateful" and "offensive" as subsets of "toxic." Moreover, the choice of prompts used to interact with ChatGPT impacts its performance. Based on these insights, our study provides several meaningful implications for employing ChatGPT to detect HOT content, particularly regarding the reliability and consistency of its performance, its understanding and reasoning of the HOT concept, and the impact of prompts on its performance. Overall, our study provides guidance on the potential of using generative AI models for moderating large volumes of user-generated textual content on social media.

CCS Concepts: • **Human-centered computing → Scenario-based design**;

Additional Key Words and Phrases: Generative AI, ChatGPT, hate speech, offensive language, online toxicity, MTurker annotation, prompt engineering

ACM Reference Format:
Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web* 18, 2, Article 30 (March 2024), 36 pages. https://doi.org/10.1145/3643829

This material is based upon work supported by the National Science Foundation under grant no. 1928434.
Authors' address: L. Li, L. Fan, S. Atreja, and L. Hemphill, School of Information, University of Michigan, 105 S State St, Ann Arbor, MI 48109, USA; e-mails: lingyaol@umich.edu, lizhouf@umich.edu, satreja@umich.edu, libbyh@umich.edu.

ACM 1559-1131/2024/03-ART30
https://doi.org/10.1145/3643829

1 INTRODUCTION

Harmful behaviors, such as cyberbullying, hate speech, and harassment, can create a hostile and unpleasant online environment, which may drive users away, reduce engagement, and cause harm to individuals [1, 2]. These negative effects make it crucial to develop models that detect and moderate such textual content on social media [3, 4]. By using these models, platforms can take prompt action to remove harmful content, warn users, or provide resources to mitigate its impact. These measures not only safeguard users from harm but also foster a safer and healthier online environment. Traditionally, one prevalent method for building these models involves human annotation of potentially toxic or hateful discourse [5, 6], which can serve as an initial step in training algorithms to recognize and filter such textual content.

However, the process of annotating harmful textual content for model training poses several challenges. First, these tasks expose annotators to harmful and offensive content, which contains violence, racism, sexism, and threats that can have negative impacts on their mental health. Consequently, this limits the pool of available annotators and restricts the amount of content they can reasonably review. Second, the financial cost of annotating data is high [7]. Toxic content is relatively rare [8], and therefore, a large number of documents must be annotated to obtain representation for effective model training. This manual process requires a considerable investment of time and resources. Third, the demographics of annotators can affect the objectivity of annotation tasks [9]. Due to variations in the social and cultural backgrounds of human annotators, the annotation results can be unstable even if they are normalized.

The challenges associated with manual annotation highlight the need for alternative approaches that can accurately detect and moderate online harm. Generative AI is one approach that holds promise in this context [10]. These AI models are often trained with large datasets of existing examples to identify common patterns and features of toxic content. Moreover, generative AI models have the potential to understand and respond in ways that mimic human conversations, which can be beneficial for detecting subtle forms of toxicity such as sarcasm or irony. As a result, there is a growing trend in exploring the potential of generative AI for performing a variety of tasks, including toxicity detection [11–13].

While generative AI models, particularly the latest **generative pre-training transformer (GPT)** models for ChatGPT, have shown promise in detecting online toxicity [10], their full potential for classifying harmful textual content is not yet fully understood. Very few studies have discussed how prompts for ChatGPT could affect its ability to interpret harmful textual content. In addition, as pointed out by Davidson et al. (2017) [5], it is crucial to distinguish between various types of harmful textual content on social media, such as hate speech and offensive language. To address these research gaps, this study investigates the capabilities of ChatGPT to detect and discriminate between three prevalent and significant forms of harmful content on social media: **Hateful, Offensive**, and **Toxic (HOT)**. While HOT content on social media can include audiovisual content, our study focuses only on textual information. To achieve this goal, the study encompasses three research questions:

- **RQ1 (reliability and consistency):** How do ChatGPT's annotations compare to those produced by MTurkers in terms of reliability and consistency?
- **RQ2 (reasoning):** How does ChatGPT comprehend and discriminate between HOT concepts, as well as the reasoning behind its classification?
- **RQ3 (prompts):** How do different prompts influence ChatGPT's ability to detect HOT comments?

This research is significant and innovative for multiple reasons. First, it addresses a crucial issue: the need for harm-free data annotation. Then, this research provides several suggestions on how to interact with ChatGPT from the perspective of prompt engineering. In addition, by combining the nuanced understanding of human moderators with the processing power and speed of AI, we can create a workflow that assists in moderating large volumes of social media data. This workflow is faster, more concise, more effective, and more scalable than training a toxicity detection model based on human annotation while maintaining high quality and consistency. The implications presented in this study are valuable for high-volume environments, such as detecting HOT content from social media platforms where the amount of user-generated content can be overwhelming.

2 BACKGROUND

2.1 Hateful, Offensive, and Toxic (HOT) Content and Its Detection

As the mode of communication shifts towards online platforms, there is an increasingly critical need for accurate and efficient automated methods of detecting harmful textual content on social media [14]. This need is particularly crucial for platforms such as X (formerly known as Twitter) and Reddit, where vast amounts of textual information are posted and shared daily. Harmful language can negatively affect targeted individuals and communities [15, 16]. Detecting harmful text content holds significant importance in ensuring the safety and well-being of individuals and communities using these platforms [1, 2]. Many sites often adopt manual moderation processes, which results in abusive content remaining online for prolonged periods without timely action being taken [17]. The development of detection models can alleviate the moderation burden and provide measures to address harmful content on time. Lastly, detecting and analyzing patterns of harmful content can provide valuable insights into the nature and extent of some social issues, such as discrimination, prejudice, and marginalization. This data is especially crucial for vulnerable communities disproportionately affected by these problems [18–20].

The concept of harmful content encompasses frequently used terms, such as "hate speech," "offensive," "toxic," "aggression," "abusive," and "insults," which have been extensively researched in previous studies [14, 15, 21–25]. However, understanding the subtle distinctions in defining these concepts is crucial [5], as it affects the quality of the training data, which in turn can influence the development of models for content moderation. Schöpke-Gonzalez et al. (2023) [74] introduced the HOT framework and studied these three commonly used concepts to understand the annotation variances caused by different concept definitions by comparing annotator labels across multiple concept definition conditions. They found that annotators considered the concepts distinct rather than interchangeable, and the annotations' characteristics were strong predictors of their eventual labels. Therefore, they suggested that researchers make specific choices about which concepts to analyze depending on their goals. Incorrectly flagging or missing content that harms historically marginalized groups may have severe consequences [27]. Therefore, it is crucial to consider the nuanced nature of these concepts when building detection models (see Table 1 for example definitions of the three concepts).

Researchers have used these definitions to identify HOT language on social media [31, 37, 38]. Lexicons or dictionaries were the first approaches to detecting these concepts in online discussions [39, 40]. This methodology involves creating a list of terms and phrases that are associated with HOT content, such as the Hurtlex (i.e., a multilingual lexicon of hate words) developed by Tontodimamma et al. [41]. Once this list is generated, rule-based methods, such as sentence-level subjectivity detection [39], can automatically filter out social media posts or comments containing these lexicons. However, this approach has some limitations, principally because relying solely on keyword filtering or predetermined rules does not capture the context or intent of the message [42]. It is also limited to identifying only known harmful words and may miss new or emerging forms of harmful content [43].

Table 1. Typical Definitions for the HOT Concept in Prior Studies

HOT	Author and Year	Definition	Origin
Hateful	Davidson et al. (2017) [5]	Language that is used to express hatred towards a targeted group or is intended to be derogatory, humiliate, or insult the group members.	Academic
	Nockleby (2000) [28]	Any communication that disparages a person or a group on the basis of some characteristic, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.	Academic
	Meta (2022) [29]	A direct attack against people of their race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.	Industry
	X (2023) [30]	Language that attacks other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.	Industry
	Salminen et al. (2020) [31]	Language containing either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity – in other words, rude and disrespectful comments- can result in negative online and offline consequences for the individual, community, and society.	Academic
Offensive	Wiegand et al. (2018) [32]	Hurtful, derogatory, or obscene comments made by one person to another person.	Academic
	Zampieri et al. (2019) [33]	Contains forms of non-acceptable language or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.	Academic
	Jay and Janschewitz (2008) [34]	Vulgar, pornographic, and hateful language. Vulgar language refers to coarse and rude expressions, which include explicit and offensive references to sex or bodily functions. Pornographic language refers to the portrayal of explicit sexual subject matter for the purposes of sexual arousal and erotic satisfaction. Hateful language includes any communication outside the law that disparages a person or a group on the basis of some characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, and religion.	Academic
Toxic	Google Jigsaw (2017) [35]	A rude, disrespectful, or unreasonable comment that is likely to make individuals leave a discussion.	Industry
	Kolhatkar et al. (2020) [36]	Comments that use harsh, offensive, or abusive language, which include personal attacks or insults, or which are derogatory or demeaning.	Academic

As **machine learning (ML)** and **natural language processing (NLP)** technologies advance, researchers have increasingly turned to these models to detect HOT language on social media [44–48]. These models involve using text vectorization techniques, which convert text into a numerical vector or matrix, allowing ML algorithms to process and analyze the data. Early approaches to text vectorization were based on the bag-of-words model or **term frequency-inverse document frequency (TF-IDF)** scheme. Models that are constructed using these two approaches are frequently employed as benchmark models [4, 44]. However, more recent detection models have utilized word embedding techniques, such as Word2Vec, GloVe, and **Bidirectional Encoder Representations from Transformers (BERT)**, which represent words as vectors in a high-dimensional space based on their contextual usage [49–51]. Once the text is vectorized, ML classifiers, such as **support vector machines (SVMs)** or neural networks, can be trained to detect HOT content [52, 53]. These ML classifiers learn to recognize patterns and features associated with HOT language, such as specific combinations of words, phrases, and sentence structures commonly used in HOT language.

The development of these supervised ML models depends on human annotations to build training datasets. While several widely-used datasets, such as Davidson Hate Speech and Offensive Language (DATASET) [5] and OffensEval [33], have been made available to the public, the process of developing new datasets or updating existing ones could potentially expose annotators to harmful content and require a significant amount of time and effort to generate. For instance, Riedl et al. (2020) [54] found that rating "uncivil" comments had a significant psychological toll on annotators and reduced their accuracy on subsequent tasks. In addition, Gilardi et al. (2023) [55] stated that while trained annotators incur high costs, employing crowdworkers like those on MTurk offers a more affordable alternative, albeit with potentially insufficient quality. However, Kasthuriarachchy et al. (2021) [56] pointed out that labeling data through MTurk can also become expensive when the target messages are small and have a lot of noise, such as toxic comments in social media data.

2.2 Generative AI Models

Generative AI models are emerging technologies that exhibit human-like understanding and generate coherent responses to human input. Training these models involves massive amounts of data and computational resources, often following a two-step process: pretraining and fine-tuning [57, 58]. In this study, we focus particularly on **Large Language Models (LLMs)**. LLMs learn from a diverse range of text data (some publicly available, such as books, articles, websites, and social media, and some provided by data annotation companies) to acquire semantic knowledge during pre-training [59, 60]. LLMs are based on the Transformer architecture [61] and employ self-attention mechanisms to process and generate sequences of tokens, such as words, subwords, or characters, as the inputs. Pre-training typically involves training the model to predict missing tokens in a given context (masked language modeling) or to complete a partially observed sequence (causal language modeling) [62, 63]. Next, LLMs refine the pre-trained models on specific tasks or datasets, adapting to particular applications and use cases, for example, smart chatbots [64, 65].

Identification of HOT comments is one of the tasks that may benefit from generative AI models. For example, generative prompt-based inference can strengthen toxicity detection [66]. The accuracy of GPT-enabled toxicity classification has been illustrated in various languages, including Swedish [67]. Generative language models may also assist in the mitigation of HOT content. For example, Kucharavy et al. (2023) [68] demonstrated the potential of these models by using them to study cyber-defense and mitigate cyber risks. In another study, Ganguli et al. (2022) [69] built red teaming generative language models through reinforcement learning with human feedback, and they found that LLMs can reduce harmful outputs.

Comprehension is a higher level of understanding of HOT comments, in addition to correctly identifying or even mitigating them. However, comprehending controversial content often requires extensive human knowledge. Preliminary evidence suggests that LLMs can match or even outperform human annotators, including MTurk crowd workers, in various content ranking and text annotation tasks [55, 70]. However, while ChatGPT can produce well-written explanations for implicit HOT content, its explanations are often unprofessional and may contain inaccuracies and misinformation [10]. The current state of automated understanding and explaining HOT content is still based on knowledge-enhanced text generation [71]. There is an immediate need for experimenting and analyzing how reliable and consistent LLMs are in identifying HOT content

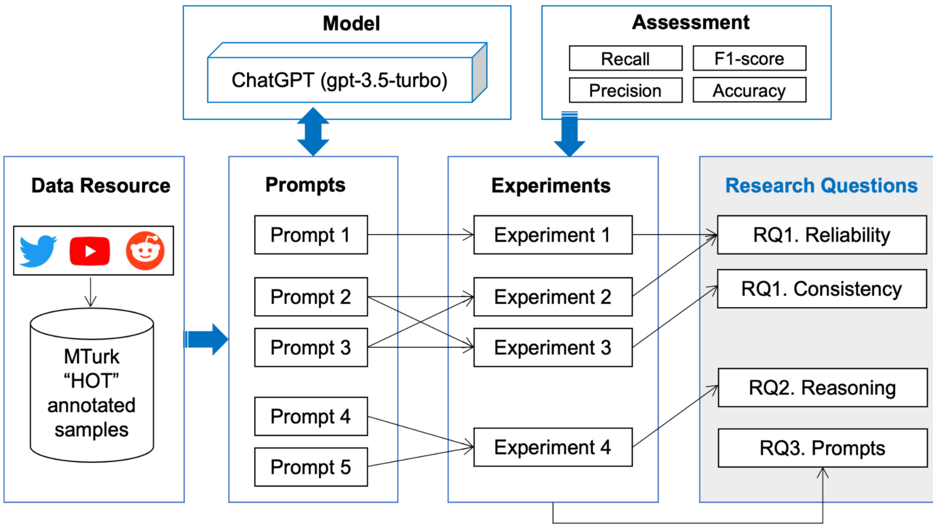


Fig. 1. Framework design for the research implementation.

and understanding the models' reasoning behind their decisions. This paper describes a study in which we experimented with various prompts to solicit HOT classifications and reasoning from one LLM, ChatGPT [72], to address this need.

3 DATA AND METHODS

Figure 1 provides a graphical representation of our research framework. In this study, we designed multiple prompts (Section 3.4) to interact with ChatGPT and assessed its performance as compared to MTurkers (Section 3.6). Our experiments aimed to answer three research questions, as documented in Section 3.4. Experiments 1, 2, and 3 aimed to address RQ1 by assessing the reliability and consistency of ChatGPT as compared to the MTurker annotations. Experiment 4 focused on RQ2 by examining the model's reasoning of HOT classifications. Finally, we assessed the effects of different prompts in multiple experiments to answer RQ3, which involved evaluating the model's overall performance.

3.1 MTurker Data Preparation

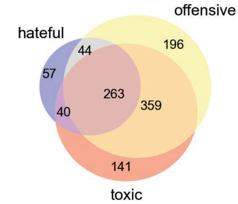
Our research employs a dataset (i.e., the HOT dataset) that was previously provided by Wu et al. (2023) [26]. We chose to use this dataset for two reasons: to evaluate the potential of ChatGPT in annotating HOT content as compared to MTurkers and to assess ChatGPT's efficiency in differentiating between HOT concepts. The HOT dataset provides labels for different HOT concepts. Using crowdsourced labels to train ML classifiers is common practice among content moderation researchers, and we were interested in how ChatGPT compares to this current practice (i.e., not to some ground-truth label).

The HOT dataset includes a diverse collection of comments sourced from three social media platforms: Reddit, X, and YouTube (Figure 1), primarily focusing on popular political news stories. As mentioned, HOT comments are not common [8]. Therefore, it was important to ensure that human annotators did not become fatigued by having to label too many comments that lacked harmful characteristics. To address this issue, Wu et al. (2023) [26] opted for purposive sampling instead of random sampling. They used a pre-trained ML model to assign a classifier score between 0 and 1 to each comment for each HOT concept. By doing so, they were able to increase the

Table 2. Annotation Result of HOT by MTurkers

Hateful	Offensive	Toxic	Count
no	No	No	2381
no	No	Yes	141
no	Yes	No	196
no	Yes	Yes	359
yes	No	No	57
yes	No	Yes	40
yes	Yes	No	44
yes	Yes	Yes	263

Venn Diagram



prevalence of HOT comments in the sample. The resulting dataset was a total of 3,481 comments, with 1,162, 1,154, and 1,165 comments collected from Reddit, X, and YouTube, respectively.

For the annotation task, Wu et al. (2023) [26] recruited annotators on Amazon **Mechanical Turk (MTurk)**. They set several requirements for potential annotators, including being a resident of the United States, having completed at least 1,000 **Human Intelligence Tasks (HITs)**, and maintaining a HIT acceptance rate of at least 98%. To identify the qualified annotators, Wu et al. (2023) [26] provided a qualification task that included concept definitions, labeling instructions, and three qualification questions. In order to be invited to participate in the comment-labeling task, annotators had to answer all three questions correctly. Annotators were allowed to label up to 100 comments each once qualified. Annotators were asked to select True or False for each concept to indicate whether they thought each comment was HOT or not. In total, they were able to collect annotations from five MTurkers for each of the 3,481 comments in the dataset.

In our study, we used the majority vote approach to determine the final label for each comment based on the annotations provided by the five independent MTurkers. Specifically, if a comment received at least three "True" HOT annotations and two or fewer "False" non-HOT annotations, we considered it to be a HOT comment. We refrained from using the number of "True" or "False" annotations as a probability due to two reasons. First, assessing how well the MTurkers represent the general population or the online communities in which these comments occur is challenging. Second, given the relatively small sample size, we are uncertain about the reliability of using five MTurkers' labels as a representative probability. Therefore, using five annotations may not be a reliable proxy for probability, which may compromise the validity of the result comparison. Due to these reasons, we found that the number of MTurkers who classified content as HOT did not correlate well with the probability ChatGPT provided (see Appendix A.3). Instead of comparing these proportions directly, we experimented with thresholds for the ChatGPT probability to determine a point at which ChatGPT achieves high agreement with MTurkers.

Table 2 presents the annotation results from MTurkers using the majority vote approach. The results indicate that out of the total 3,481 comments, 2,381 comments are non-HOT, and 263 are HOT. Regarding the HOT concepts, there are significant overlaps between the different concepts, with 622 out of 803 "toxic" comments classified as "offensive." However, the MTurker annotations reveal distinct patterns in the HOT comments. For the subsequent analysis, we examined ChatGPT's comprehension of these concepts. Therefore, we experimented by comparing the model's output with the MTurkers' annotations. We defined ChatGPT as "accurate" when it gave the same annotation as a majority of MTurkers for an individual item.

As the Venn diagram in Table 2 indicates, the concepts have significant overlap. Schöpke-Gonzalez et al. (2023) [74] stated that these three concepts are likely to include insults and name-calling, but they have unique distinctions that mean they are not interchangeable, as communication styles such as aggressive tone and name-calling distinguish content that MTurkers labeled with different concepts. As Schöpke-Gonzalez et al. (2023) [74] found in their paper, hate is a

Table 3. Definitions of HOT Provided to MTurkers and ChatGPT

Concept	Definition
Hateful	"expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" [5]
Offensive	"contains hurtful, derogatory, or obscene comments" [32]
Toxic	"a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion" [35]

distinct concept relative to offensive and toxic. For example, MTurkers indicated that the following comment was hateful but neither offensive nor toxic:

> *"Cool, you'll be aiding and abetting criminals, and yet your fellow Americans who did nothing wrong are rotting in jail for nothing. This should help put things into perspective for you. Democrats want to keep us all on the plantation."*

In addition, Schöpke-Gonzalez et al. (2023) [74] revealed that toxic and offensive are not interchangeable in some situations. For example, MTurkers indicated the following comment was offensive but not toxic.

> *"The problem is a lot of today's young society wants to have everything but not actually do anything for it. Sounds more like a boomer landlord to me. These women have something desirable people are willing to pay for; that's your precious capitalism for you."*

To further demonstrate the level of agreement between MTurkers, we incorporated the Krippendorff's α (i.e., an inter-coder reliability index [73]) of the HOT dataset obtained from Schöpke-Gonzalez et al. (2023) [74] into Appendix A.1. It should be noted that our task involved comparing the outputs from ChatGPT with the majority vote derived from five MTurker labels. While the inter-coder agreement was not high (see Table A.1 in Appendix A.1), our primary research objective was to compare ChatGPT's results within current practice to examine if generative AI can protect annotators from labeling harmful comments rather than determine the validity of these results against ground-truth values.

3.2 HOT Definitions

One of the research objectives was to evaluate ChatGPT's ability to capture the nuances in annotator interpretation of harmful content. As outlined in Section 2.1 (Table 1), prior studies have used different definitions of each HOT concept. However, such definition variations can potentially affect ChatGPT's understanding of HOT comments. To ensure consistency in definitions provided to both MTurkers and ChatGPT, we adopted three popular HOT definitions, as presented in Table 3. Specifically, these three widely-used definitions include Davidson et al.'s (2017) [5] definition of hateful content, a modified version of Wiegand et al.'s (2018) [32] definition that isolates its components differentiating it from hatefulness for offensive content, and Perspective API's definition for toxic content [35]. The definitions of HOT are presented in Table 3.

3.3 ChatGPT Model

In this study, we employed the gpt-3.5-turbo model, a variant of the GPT-3.5 family of models. This model, which is the same one utilized in the ChatGPT product, was selected for its superior performance ability to generate high-quality text and its large user base [75]. OpenAI reports that the GPT-3.5 models can comprehend and generate natural language or code. Multiple evaluations

Table 4. Parameters in the Request Body of ChatGPT [72]

Concept	Definition
max_tokens	The maximum number of tokens to generate in the completion.
Temperature	A value between 0 and 2; higher values make the output more random, while lower values make the output more deterministic.
top_p	A value implies that the model considers the results of the tokens with top_p probability mass.
presence_penalty	A number between −2 and 2; positive values increase the model's likelihood to talk about new topics.
frequency_penalty	A number between −2 and 2; positive values decrease the model's likelihood to repeat the same line verbatim.

and tests have validated its capabilities in language understanding and knowledge retention [55, 76, 77]. Among the models, the gpt-3.5-turbo model is considered one of the most efficient and cost-effective LLMs as of April 2023, when we drafted this paper. This model has been optimized for chat-based applications but has also demonstrated strong performance in traditional text completion tasks [72]. We also tested several other LLMs and content moderation models (see Appendix A.5) and determined to use ChatGPT given its performance. For consistency, we used the term - "ChatGPT" to refer to the gpt-3.5-turbo model in the following writing.

The OpenAI API provides a range of parameters in the request body that can be customized to adjust the request. Some of the key parameters that can be adjusted are listed in Table 4 [72]. For our specific objective of testing the reliability and consistency of GPT models in identifying HOT content, we aimed to avoid randomness in our results. To achieve this, we set the temperature parameter to 0 in all experiments except for Experiment 3, where we varied the temperature parameter for consistency testing purposes. The top_p parameter can also be used to control randomness, but OpenAI recommends against modifying both top_p and temperature together [72]. To ensure that all available tokens in the results are accounted for, we used the default top_p value (default = 1) for our experiments.

3.4 Prompt Design

A prompt is a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities [78]. Prompt engineering has become an increasingly crucial skill set to communicate effectively with LLMs like ChatGPT. Prompts can be seen as a type of programming that enables users to personalize the generated outputs and interactions with an LLM [78]. A prompt can influence the output generated from an LLM by providing specific rules and guidelines for an LLM. For instance, when performing a data annotation task, a prompt may provide conceptual definitions to the LLM. A prompt may also specify desired output forms, such as returning a probabilistic output or a binary output. Therefore, by introducing variations in the input prompts, we can enable LLMs to perform a larger and more diverse set of annotation tasks without requiring new training data or updating the underlying model.

To generate our prompts, we used prompt patterns identified by White et al. (2023) [78]. These patterns describe effective techniques for accomplishing different interaction objectives when interacting with conversational LLMs. Although they were initially designed for software engineering tasks, these patterns are not domain-specific and can be applied in various contexts. For our annotation task, we specifically adopted the following prompt patterns that were most appropriate:

- Output Automator (binary or probability)
- Template (provide an explanation or not)

Table 5. Prompts to Interact with ChatGPT

Prompt	Prompt design	Prompt format
Prompt 1	Ask ChatGPT the same question that we ask MTurkers	Do you think this comment is <HOT>? (1) Yes, (2) No. A comment is <HOT> if you perceive that it <HOT definition>.
Prompt 2	Only ask for a binary classification without explanations.	I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is <HOT>. A comment is <HOT> if you perceive that it <HOT definition>. I want you to only respond with yes or no. Do not provide any other outputs or any explanation for your output.
Prompt 3	Only ask for a probability representing the level of <HOT> without explanations.	I want you to provide a probabilistic score between 0 and 1, where the score represents the probability of the following comment being <HOT>. A comment is <HOT> if you perceive that it <HOT definition>. A probability of 1 means that the comment is highly likely to be <HOT>. A probability of 0 means that the comment is highly unlikely to be <HOT>. Do not provide any other outputs or any explanation for your output.
Prompt 4	Ask for a binary classification with further explanations.	I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is <HOT>. A comment is <HOT> if you perceive that it <HOT definition>. I want you to respond with yes or no in the first line and provide an explanation for your output in the second line.
Prompt 5	Ask for a probability representing HOT with further explanations.	I want you to provide a probabilistic score between 0 and 1, where the score represents the probability of the following comment being <HOT>. A comment is <HOT> if you perceive that it <HOT definition>. A probability of 1 means that the comment is highly likely to be <HOT>. A probability of 0 means that the comment is highly unlikely to be <HOT>. I want you to respond with a probabilistic score in the first line and provide an explanation for your score in the second line.

We utilized these prompt patterns to design five prompts for interacting with ChatGPT (see Table 5). Prompt 1 involved posing the same question to ChatGPT as had been posed to the MTurkers. For Prompts 2 and 3, we asked ChatGPT to deliver either a binary classification or a probability regarding its perception of the HOT concept and to not provide an explanation. Prompts 4 and 5 required ChatGPT to provide an explanation for its binary or probability output. We listed the specific prompts for HOT classifications in Appendix A.2.

3.5 Experiment Design on the HOT Dataset

To evaluate the reliability and consistency of ChatGPT in comprehending HOT concepts, we carried out four experiments, which are illustrated in Figure 1. Experiments 1 and 2 were conducted to evaluate the reliability of ChatGPT, while Experiment 3 was designed to test the model's consistency, and all three aimed to answer RQ1. Experiment 4 focused on the ChatGPT's understanding and reasoning processes; specifically, we examined how it made classification decisions (RQ2). Using the five prompts we developed, we explored the impact of different prompts on ChatGPT's understanding of harmful content (RQ3). Details of each experiment are provided below.

Experiment 1: direct comparison with MTurkers. Our first experiment evaluated the performance of the GPT model in understanding the HOT concept without any instruction. To achieve this, we posed the same question to ChatGPT as posed to MTurkers. By comparing the accuracy of the labels generated by ChatGPT with those produced by MTurkers, we were able to gain a preliminary understanding of the model's ability to comprehend the HOT concepts.

Experiment 2: binary vs probability prompts. We experimented with two prompts that requested two forms of output: one that requested a binary label and another that requested a probabilistic score. By comparing the results generated by ChatGPT with MTurkers' annotations, we were able to gain insights into how the prompts affect the reliability of annotation and how ChatGPT implements different thresholds for classifying HOT comments.

Experiment 3: consistency. In our third experiment, we conducted multiple iterations of labeling the dataset with Prompts 2 and 3. We posed the same task using the same prompt to ChatGPT multiple times to see whether it returned the same results on repeated requests. By comparing the results generated by ChatGPT across different iterations, we were able to assess the consistency of ChatGPT's performance in terms of data annotation.

Experiment 4: annotation reasoning. Our last experiment examined the understanding of ChatGPT's decision-making process in the context of harmful content. We used the prompts from Experiment 2 and asked ChatGPT to provide additional explanations for its classifications. By doing so, we could understand how the reasoning affects ChatGPT's classifications of HOT content. In addition, we analyzed the semantic patterns present in ChatGPT's explanations by comparing the n-grams present in the reasoning and manually reading a sample of comments. Using Venn diagram and n-gram analysis, we gained insights into ChatGPT's ability to recognize the nuances in HOT content and their associated reasoning patterns.

3.6 Performance Measures

Given that the HOT dataset has unbalanced classes, we used Precision, Recall, and F1-score plus accuracy to compare ChatGPT's comprehension of harmful concepts to estimate its performance relative to MTurkers. It is worth noting that in our evaluation of ChatGPT's classification performance, we used MTurker annotations as the "ground-truth" values. However, we acknowledge that determining the actual "ground-truth" requires expert knowledge and further justification for HOT concepts.

In our research context, Precision measures the fraction of true positive cases (i.e., HOT comments identified by both MTurkers and ChatGPT) over the total number of predicted positive cases (i.e., HOT comments identified by ChatGPT). Precision represents the model's ability to identify relevant cases correctly. In contrast, Recall measures the fraction of true positive cases (i.e., HOT comments identified by both MTurkers and ChatGPT) over all the relevant cases (i.e., HOT comments identified by MTurkers). Recall represents the model's ability to identify all relevant cases. In addition, the F1-score is a commonly used metric that combines both Precision and Recall into a single score. It provides an overall assessment of the model's performance for each class [79]. In our study, a higher F1-score indicates that ChatGPT shows more agreement with MTurker annotations for a given class. Accuracy measures the overall number of classification agreements as compared to MTurkers. These metrics are defined as below.

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

$$F1\text{-}score = 2 \times Precision \times Recall/(Precision + Recall) \tag{3}$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{4}$$

where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

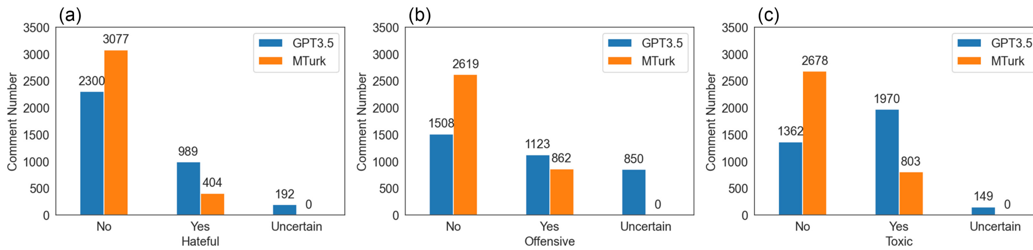


Fig. 2. HOT classification results based on Prompt 1. (a) Hateful classification results of ChatGPT and MTurker annotations. (b) Offensive classification results of ChatGPT and MTurker annotations. (c) Toxic classification results of ChatGPT and MTurker annotations.

4 RESULTS

We conducted four experiments, as outlined in Section 3.5. In response to RQ1, we assessed two crucial attributes of ChatGPT, namely reliability and consistency. To evaluate reliability, we reported Precision, Recall, and F1-scores compared to MTurker annotations. We also applied chi-square tests in Experiments 1 and 2 to compare ChatGPT's outputs with MTurker annotations. To evaluate consistency across different iterations and temperatures, we calculated the Krippendorff's α, an inter-coder reliability index [73] of the achieved agreement between independent experiments (i.e., ChatGPT outputs). In response to RQ2, we first utilized Venn diagrams to illustrate ChatGPT's ability to comprehend different HOT concepts and highlight the differences between them. We then analyzed n-grams to investigate how ChatGPT arrived at its HOT classification for the comments. In response to RQ3, we generated heatmaps to compare the performance using five prompts and explored how different prompts can affect the model's ability. Examples of ChatGPT's outputs using different prompts are presented in Table 6.

4.1 Results of Experiment 1 – Direct Comparison

Figure 2 displays the classification output of ChatGPT for HOT categories based on 3,481 comments. Our findings indicate ChatGPT's answers were more varied, including uncertain responses that leaned towards either yes or no. When ChatGPT is confident in its classification, it returns a clear yes or no. However, when it is uncertain about the classification, it often provides reasoning. It explains why some people may perceive the content as HOT or not. In this experiment, we classified all comments on which ChatGPT was unclear or ambiguous (not definitively "yes" or "no") as "uncertain."

The results presented in Figure 2 indicate that ChatGPT exhibits higher confidence levels in classifying comments as "hateful" and "toxic" as opposed to "offensive." Specifically, 850 items are classified as uncertain in the Offensive category by ChatGPT. The model displays a greater inclination towards labeling comments as HOT, as evidenced by the larger number of "yes" outputs in all three categories, compared to MTurker annotations. ChatGPT shows a higher tendency to label comments as "toxic," which may indicate a lower threshold for identifying toxic language. Table 7 presents the classification performance, wherein we excluded the "uncertain" classifications and focused solely on the "certain" results to determine the GPT model's reliability.

Based on the chi-square test results presented in Table 7, there is sufficient evidence to conclude that the distribution of ChatGPT's annotation is the same as the distribution of MTurker annotations across all three categories. ChatGPT shows better performance in all three categories regarding F1-scores for non-HOT than HOT comments, especially non-hateful ones. Compared to MTurkers, ChatGPT exhibits higher agreement for the Offensive and Hateful categories, while the

Table 6. Examples of ChatGPT's Outputs based on Different Prompts

Comment	Prompt	ChatGPT's Response by Concept		
		Hateful	Offensive	Toxic
Lmao, why did the CDC stop tracking all breakthrough cases in May? Well how else are you gonna get fda approval if you can't fudge the numbers? You can't be this stupid....	Prompt 1	Yes.	Yes, the comment is offensive.	Yes, this comment is toxic.
When is enough enough???? Everyone needs to stand up and run these people out of our country. F for our children and there future. There is no future for them as of right now. I beg you you stand up for what is right and protect your country and kids.	Prompt 1	No. While the comment may express strong opinions and a desire for action, it does not express hatred towards a specific group or intend to insult or humiliate anyone.	As an AI language model, I cannot perceive emotions or feelings. However, the comment can be considered offensive to some people as it suggests the removal of a group of people from a country.	Yes, this comment can be perceived as toxic as it uses aggressive language and encourages violence towards a group of people. It may also make readers feel uncomfortable and want to leave the discussion.
@nprpolitics They all need mental health resources to help them recover from the years of masking and isolation.	Prompt 2	No.	No.	No.
I hate masks, but I am about to take a trip me be around thousands of other people. Bought a respirator and big boy filters for it. Fuck masks, but fuck covid more.	Prompt 3	0.2	0.7	0.8
@TypeOutThought @geralt_of_rivi @CNN And what happened to, you're not talking about vaccinations? Your level of basic intelligence is extremely low.	Prompt 4	No. The comment does not express hatred towards a targeted group or intend to be derogatory, humiliate, or insult any members of a group. It is a personal attack on the intelligence of an individual, but it does not target a specific group based on their race, ethnicity, religion, gender, or any other characteristic.	Yes. The comment includes a derogatory statement about the intelligence of the person being addressed, which can be considered hurtful and offensive.	Yes. The comment contains personal attacks and insults towards the recipient's intelligence, which is likely to make readers feel disrespected and want to leave the discussion. Additionally, the use of laughing emojis to mock the recipient's previous statements adds to the toxic nature of the comment.

(Continued)

Table 6. Continued

Comment	Prompt	ChatGPT's Response by Concept		
		Hateful	Offensive	Toxic
Any pastor pushing the shot needs to be double checked, and you should probably look at a new church.	Prompt 5	0.3 While the comment may be seen as critical of pastors who promote vaccination, it does not necessarily express hatred or insult towards a targeted group. The comment does not specifically target any particular group and does not use derogatory language. However, the use of the phrase "double checked" and the suggestion to find a new church may be seen as somewhat confrontational, which could be interpreted as mildly derogatory. Overall, the comment is not highly likely to be considered hateful, but there is some potential for it to be perceived as such.	0.3 While the comment may be seen as critical of pastors who promote vaccinations, it does not contain any explicit derogatory or obscene language. However, the suggestion to leave one's church could be perceived as hurtful to those who hold their faith and religious community in high regard. Overall, the comment is more critical than offensive.	0.8 The comment contains a disrespectful and potentially harmful statement towards pastors who advocate for the COVID-19 vaccine. The use of the phrase "double checked" implies a suspicion of wrongdoing, which can be seen as rude and unreasonable. Additionally, suggesting that someone should leave their church based on their stance on a medical issue can be seen as disrespectful. Overall, the comment has a high probability of being toxic.

Table 7. Classification Performance of HOT based on Prompt 1

Category	Class	Support	Precision	Recall	F1-score	Accuracy	Chi-square statistic
Hateful	Yes	397	0.30	0.74	0.43	0.76	410.7***
	No	2883	0.95	0.76	0.85		
Offensive	Yes	726	0.55	0.85	0.67	0.77	740.6***
	No	1905	0.93	0.74	0.82		
Toxic	Yes	795	0.39	0.96	0.55	0.63	588.7***
	No	2537	0.98	0.52	0.68		

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Toxic category shows lower agreement. F1-scores for the "yes" or "no" classification of the Offensive category are more balanced. ChatGPT demonstrates high precision in identifying non-HOT comments but displays lower precision in identifying HOT comments, which could be attributed to its lower threshold. Compared to MTurkers, while the distribution of annotations is similar, ChatGPT classified more comments as HOT for the HOT dataset.

4.2 Results of Experiment 2 – Binary vs. Probability Prompts

Our second experiment comprised two prompts, as described in Section 3.4. Prompt 2 was utilized to elicit a binary response (i.e., yes or no) from ChatGPT. The resulting outputs were then compared to the annotations provided by MTurkers and are presented in Figure 3. However, a few cases were observed in which ChatGPT inaccurately identified a comment as containing multiple responses, thereby producing multiple results. Such cases were classified as "wrong class."

Table 8. Classification Performance of HOT based on Prompt 2

Style Tag	Class	Support	Precision	Recall	F1-score	Accuracy	Chi-square statistic
Hateful	Yes	404	0.45	0.34	0.39	0.87	356.5***
	No	3077	0.92	0.79	0.93		
Offensive	Yes	862	0.67	0.55	0.61	0.82	851.3***
	No	2616	0.86	0.91	0.89		
Toxic	Yes	801	0.47	0.86	0.61	0.75	816.6*
	No	2675	0.94	0.71	0.81		

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

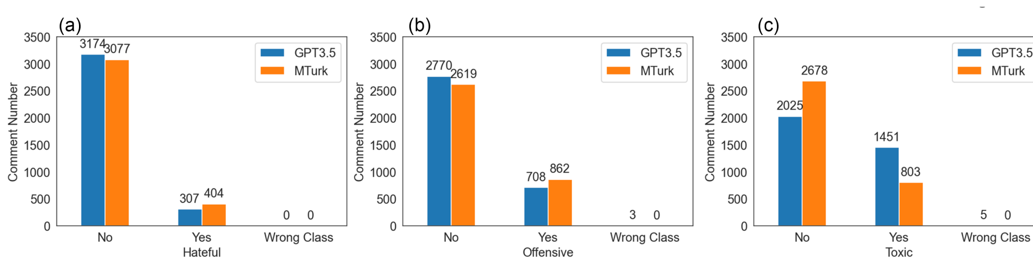


Fig. 3. HOT classification results based on Prompt 2. (a) Hateful classification results of ChatGPT and MTurker annotations. (b) Offensive classification results of ChatGPT and MTurker annotations. (c) Toxic classification results of ChatGPT and MTurker annotations.

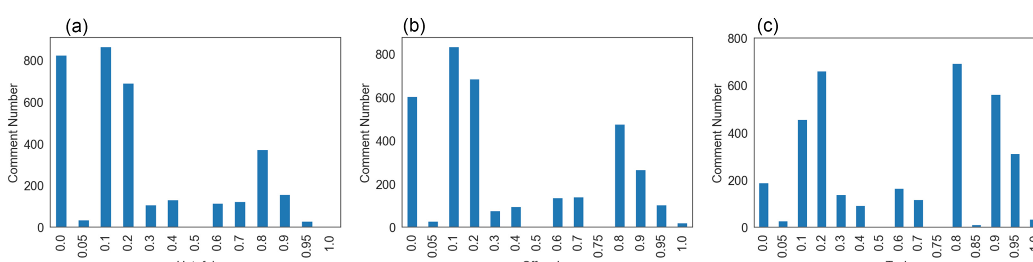


Fig. 4. HOT classification results based on Prompt 3. (a) Hateful classification results. (b) Offensive classification results. (c) Toxic classification results.

The results in Figure 3 demonstrate that ChatGPT exhibits a similar distribution with MTurker annotations, particularly concerning the Hateful and Offensive categories. The chi-square test presented in Table 8 also shows that the distribution of ChatGPT's annotation is significantly similar to the distribution of MTurker annotations across the three categories. Further, we used Precision, Recall, and F1-scores to evaluate the model's comprehension of HOT concepts, as presented in Table 8. Compared to the classification performance based on Prompt 1, the performance based on Prompt 2 shows clear improvement, particularly for the non-HOT class, as evidenced by higher F1-scores. Despite this improvement, ChatGPT still displays low F1-scores for the HOT concept, especially with its classification of "non-hateful" comments exhibiting significant disagreement with MTurker annotations. In line with the first experiment's results, ChatGPT displays a greater tendency to classify comments as "toxic" compared to MTurkers.

When using Prompt 3, ChatGPT generated a probability to indicate the level of HOT given a comment. Figure 4 presents the classification outcomes obtained from the model. We observed several interesting findings. First, ChatGPT tends to avoid assigning a probability of 0.5, which is often associated with a completely neutral stance. This output could suggest that the model struggles to make a confident determination of whether a comment should be classified as HOT or non-HOT when the probability is precisely in the middle.

Second, the model exhibits a conservative approach towards extremely HOT classifications, as indicated by the low number of "1" classifications. This output is also consistent with previous experiments for the binary classification from Experiments 1 and 2. We noticed that the model tends to lean towards "no" when asked for a binary classification with no other explanation, as demonstrated by the substantial reduction in the number of "yes" classifications in Experiment 2 compared to Experiment 1.

Third, we observed that comments with a probability between 0.3 and 0.7 were relatively rare. We hypothesize that ChatGPT tends to classify comments as either HOT or non-HOT strongly, with less emphasis on the intermediate probability. This result is likely due to the subjective nature of determining whether a comment falls under the HOT category, as a probability between 0.3 and 0.7 can be interpreted as HOT or not HOT depending on the individual's perspective.

Last, we noticed that ChatGPT occasionally produces classifications that are not rounded to one decimal place (e.g., 0.95 and 0.05), possibly implying a high degree of confidence in classifying HOT contents but not necessarily complete certainty. In contrast to the "non-hateful" (Figure 4(a)) or "non-offensive" (Figure 4(b)) classifications, our result shows that extreme "non-toxic" classifications are relatively rare in the selected dataset (Figure 4(c)). This output may be due to ChatGPT's lower threshold for "toxic" classification as compared to "hateful" or "offensive."

A critical question when working with probability is determining the appropriate threshold for classification. In the case of the Perspective API [35], researchers are advised to experiment with thresholds between 0.7 and 0.9 to classify harmful content. Similarly, in our study, we were interested in finding the threshold that aligns with the HOT classifications made by MTurkers. To this end, we calculated Precision, Recall, and F1-scores for the HOT classifications given thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9, which are reported in Figure 5.

For the Hateful category, a threshold of 0.7 produced the highest F1-score for "hateful" while still maintaining a respectable F1-score of 0.9 for "non-hateful." However, choosing a threshold of 0.9 to improve the F1-score for "non-hateful" resulted in a significant drop in the F1-score for "hateful," given that our data is imbalanced. The Offensive category followed a similar pattern, with F1-scores for thresholds of 0.5, 0.6, 0.7, and 0.8 being close. The thresholds of 0.7 and 0.8 produce the overall best F1-scores among the four, and a threshold of 0.9 leads to a substantial decrease in the F1-score for "offensive." Unlike these two categories, the Toxic category yielded an interesting result, with F1-scores for both "toxic" and "non-toxic" classes increasing as the threshold increases.

4.3 Results of Experiment 3 – Consistency

The third experiment was designed to test the consistency of ChatGPT's performance in annotating HOT comments. We aimed to study how stable the annotation results are across different rounds of experiments of Prompts 2 and 3 based on different Temperature parameters. Different Temperature settings in the ChatGPT API are defined as a float number from 0 to 1, which represents the level of randomness or entropy present in the model's output. Lower entropy (temperature 0) corresponds to more deterministic outputs, while higher entropy (temperature 1) results in more varied and less predictable text.

In particular, we analyzed the variation of the model's outputs using Krippendorff's α [73]. That is, for each setting, e.g., prompt 2 with temperature 0, we ran the experiment twice and used Krippendorff's α to assess consistency between these two attempts with the same setting. As suggested, Krippendorff's α above 0.8 is considered a very good agreement.

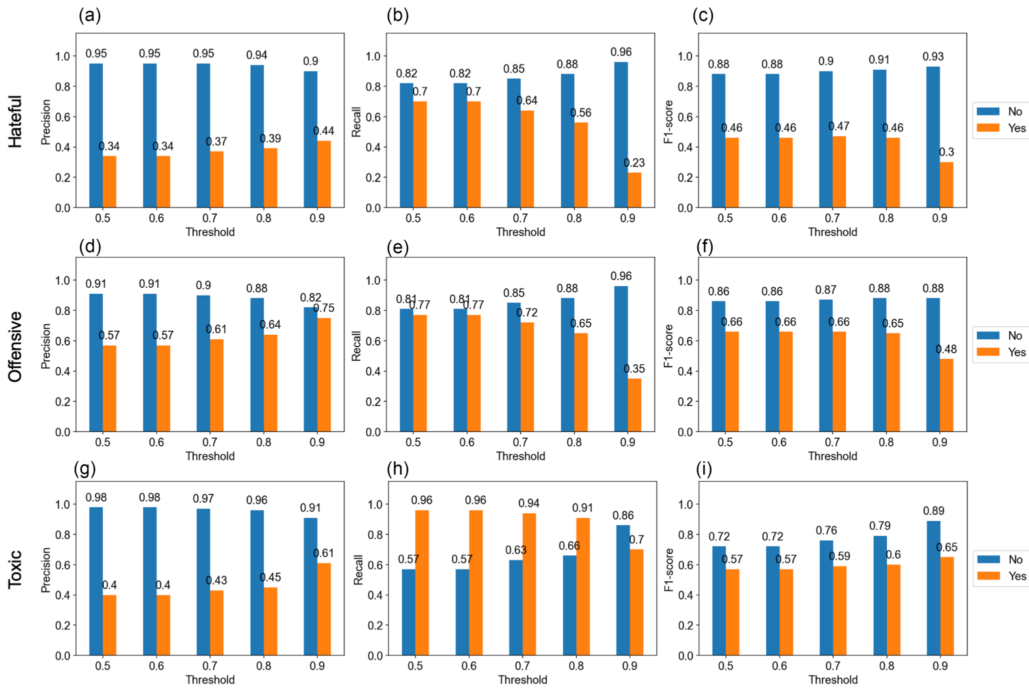


Fig. 5. HOT classification performance based on Prompt 3 compared to MTurker annotations. (a) Precision in the Hateful category. (b) Recall in the Hateful category. (c) F1-score in the Hateful category. (d) Precision in the Offensive category. (e) Recall in the Offensive category. (f) F1-score in the Offensive category. (g) Precision in the Toxic category. (h) Recall in the Toxic category. (i) F1-score in the Toxic category.

Table 9. Consistency (Krippendorff's α) of ChatGPT's Performance in Annotating HOT Comments

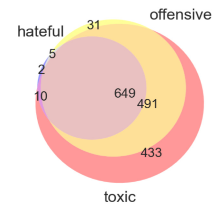
Prompt	Temperature	α of Hateful	α of Offensive	α of Toxic
2	0	0.97	0.98	0.98
3	0	0.95	0.95	0.95
3	1	0.90	0.91	0.92

As Table 9 shows, the results of the experiment demonstrated that ChatGPT generated consistent annotations for the HOT comments – no matter what forms of results were requested, either binary classification or probability, or what temperature parameters, i.e., 0 or 1, is used in the prompt, $\alpha \geq 0.9$. The annotation agreement is also stable regarding the three different HOT concepts, while the agreement for hateful comments is slightly lower in two of the three combinations of Prompt and Temperature parameters. We conducted an additional experiment to test the impact of variations in prompt text on the performance and found no significant differences (see Appendix A.4 for details).

Regarding the influence of different prompts, Prompt 2 was slightly more stable than Prompt 3 in different HOT concepts. This observation suggests that when the same temperature is set, ChatGPT's annotation performance on simpler output requests (binary classification) is probably more stable than on complex requests (probability).

Our experiment also analyzed the influence of the Temperature setting on the model's output. We observed that the model's performance was more consistent at Temperature 0, which agrees

Table 10. Annotation Result of HOT by MTurkers

Hateful	Offensive	Toxic	Count	Venn Diagram
No	No	no	1849	
No	No	yes	433	
No	Yes	No	31	
No	Yes	Yes	491	
Yes	No	No	2	
Yes	No	Yes	10	
yes	Yes	No	5	
yes	Yes	Yes	649	

with Temperature's definition in ChatGPT. In other words, the results are more deterministic when a lower temperature is set.

4.4 Results of Experiment 4 – Annotation Reasoning

Experiment 4 addressed RQ2 regarding how ChatGPT understands and distinguishes different types of HOT content. To answer this question, we first grouped the HOT classifications produced by Prompt 4 and then used the Venn diagram to visualize the classifications. We identified and removed 11 items in which ChatGPT mistakenly treated one comment as multiple comments. We used the remaining 3,470 comments to produce Table 10 and the corresponding Venn diagram, which illustrates the overlap between different HOT concepts.

Out of the 3,470 comments, 1,849 were classified as non-HOT, and 649 were classified as HOT. One notable observation is that comments classified as "hateful" or "offensive" are highly likely to be categorized as "toxic." The Venn diagram in Table 10 shows that the Toxic category largely overlapped with the Offensive and Hateful categories. ChatGPT seemed to have a low threshold for classifying comments as "toxic," while its threshold for "hateful" classification was relatively high.

Table 10 highlights that 491 comments were classified as both "toxic" and "offensive" but not "hateful." When manually reading a sample of these comments, we discovered that the model often categorizes a comment as "hateful" if it targets a particular group of people based on their race, ethnicity, religion, or other characteristics. We found that sometimes, when a comment expresses hatred towards an individual, ChatGPT does not classify it as "hateful." For instance, consider the comment, "*So should you, Mr. Graham! You are a traitor!*" which was labeled as "offensive" and "toxic." The model was classified as "non-hateful" because it did not express hatred towards a targeted group of people despite being insulting or derogatory towards the individual. However, we also found that this rule does not always hold for ChatGPT output (see examples in Appendix A.6).

In addition, we observed that 433 comments were classified as "toxic" but not "hateful" or "offensive." After manually examining a sample of the comments, we discovered that ChatGPT does not categorize a comment as offensive if it does not contain hurtful, derogatory, or obscene language. However, it may classify a comment as "toxic" if it is likely to discourage further discussion (see examples in Appendix A.6). For example, consider the comment *"@CTVNews Well, this is completely false #fakenews."* ChatGPT classified it as "toxic" by explaining that it contains a disrespectful and dismissive tone towards the news source. It then explained that the hashtag #fakenews discredits the information presented, which is likely to discourage further discussion and can be perceived as "toxic." These findings reveal that ChatGPT's classification of "toxic" comments is sensitive to the use of language that may discourage further discussion, even if it does not contain explicit hate speech or offensive language.

We also examined the 31 comments that were classified as "offensive" but not "hateful" or "toxic." However, we are uncertain whether ChatGPT has a clear classification criterion for these

Table 11. Informative Top N-grams among Reasonings across HOT Concepts in Prompts 4 and 5

HOT	Prompt 4	Prompt 5
Hateful	('to', 'be', 'derogatory'), 2,445 ('a', 'targeted', 'group'), 2,388 ('express', 'hatred', 'towards'), 2,304 ('be', 'derogatory', ',', 'humiliate'), 2,223	('hatred', 'towards', 'a', 'targeted', 'group'), 1,331
Offensive	('hurtful', ',', 'derogatory'), 2,153 ('derogatory', ',', 'or', 'obscene', 'language'), 2,121	('derogatory', 'or', 'obscene', 'language'), 2,183 ('be', 'perceived', 'as'), 1,403 ('be', 'interpreted', 'as'), 1,032
Toxic	('make', 'readers', 'want', 'to', 'leave'), 2,022 ('rude', ',', 'disrespectful'), 1,597	('contain', 'any', 'explicit'), 1,708 ('readers', 'want', 'to', 'leave'), 962 ('want', 'to', 'leave', 'the', 'discussion'), 924

comments. For instance, consider the comment, *"@Puffymonsta @CNBCnow You stay home if you are that scared. LMAO. You are vaccinated!!!"* ChatGPT labeled it as "offensive" because it dismisses someone's concerns about COVID-19 and uses derogatory language (i.e., "scared") to belittle them. The model explained that the use of "LMAO" can be perceived as mocking or insensitive. However, if ChatGPT recognizes such language as "offensive," it is also probable that this comment could also be classified as "toxic," according to our definition. This observation also applies to comments identified as "hateful" or "offensive" but not "toxic."

Through our previous analysis, we identified certain nuances in ChatGPT's ability to understand HOT content. We undertook a deeper investigation to understand how ChatGPT reaches its classifications by analyzing n-grams in ChatGPT's explanations. We extracted and analyzed the informative n-grams (n = 3, 4, or 5) from Prompts 4 and 5 reasonings. As Table 11 shows, the n-grams mainly addressed the HOT definitions, implying that ChatGPT's decisions align with the definitions we provided. For the Hateful category based on Prompt 4, for example, ('to,' 'be,' 'derogatory') is a trigram addressing the provided definition. Another trigram ('a', 'targeted,' 'group') indicates that the GPT's judgment includes checking if the potential hate speech is targeting a specific group. From another perspective, ChatGPT's annotation mechanism varies for different HOT concepts. The reasoning for "hateful" comments needs to target a group, or the comment needs to be derogatory while reasoning for "toxic" comments implies that the conversation makes people want to leave the discussion. In this sense, ChatGPT's reasoning on HOT comments conforms with the definition we provided for decision-making, while the carried-on knowledge in the model is probably used to assist in deciding HOT annotations.

To better understand the distinctions between HOT annotations in a category using binary classification and those using probability, we grouped the results of Prompts 4 and 5, respectively, based on the "yes/no" or "probability." Then, we applied the n-gram analysis to each category. We focused on analyzing the semantic features that expressed the level of certainty in the reasonings. Our analysis of Prompt 4, as presented in Table 12, revealed that the "yes" reasoning tends to use confirming expressions, although words such as "likely" and "can" suggest that the reasoning within this category is conservative. In contrast, the "no" reasonings are typically more definitive, using language such as "does" and "not" to confirm their stance.

We also analyzed the reasoning behind assigning probability in Prompt 5. Our findings are illustrated in Figure 6, where the classifications are binned: we defined the scores of 0.2 and 0.8 as the thresholds among the identifications of HOT concepts being "Unlikely," "Uncertain," and "Likely." Comparing this figure to the one obtained from Prompt 3 (Figure 4), we observed that the model is even more hesitant to assign an extremely unlikely score (i.e., 0) or a highly likely one (i.e., 1) under Prompt 5. Similar to Prompt 3, we noticed a relative scarcity of probability falling between 0.3 and 0.7.

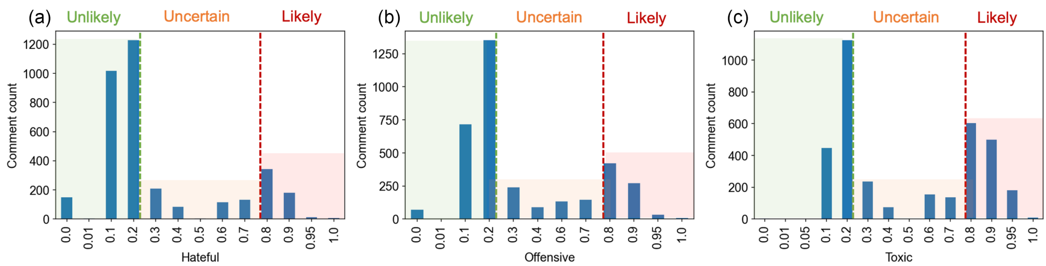


Fig. 6. HOT classification results based on Prompt 5. (a) Hateful classification results. (b) Offensive classification results. (c) Toxic classification results.

Table 12. Informative Top N-grams across the Binary Classification in Prompt 4

Binary class	n-gram and count
HOT - Yes	('is', 'likely', 'to'), 949
	('comment', 'contains', 'derogatory', 'language'), 767
	('which', 'can', 'be'), 684
HOT - No	('comment', 'does', 'not'), 4,790
	('not', 'express', 'hatred'), 2,543
	('comment', 'does', 'not', 'contain', 'any'), 2,415

Table 13. Informative Top N-grams across the Probabilities in Prompt 5

Probability class	n-gram and count
Unlikely [0, 0.2]	('comment', 'does', 'not'), 3,970
	('not', 'contain', 'any', 'explicit'), 2,107
Uncertain (0.2, 0.8)	('not', 'contain', 'any', 'explicit'), 676
	('could', 'be', 'perceived', 'as'), 533
	('could', 'be', 'interpreted', 'as'), 518
Likely [0.8, 1]	('The', 'comment', 'contains'), 1,721
	('the', 'comment', 'does'), 912

As Table 13 indicates, for Prompt 5, the language use in reasoning is more certain, especially compared to the frequent n-gram based on Prompt 4 (see Table 12). When a comment is "unlikely" to be a HOT concept, "does not" or "does not contain" are prevailing, and "contains" and "does" are frequently used when a comment is "likely." This higher level of certainty shows that when we requested probability instead of binary classification and excluded the "uncertain" comments, ChatGPT showed more confidence in providing an answer for HOT annotation. The "uncertain" category features a higher frequency of words such as "could" while also including the same n-gram "not contain any explicit" as the reasonings in the "unlikely" category. This mixed-use of expressions of uncertainty and comparatively more certain expressions implies some randomness in ChatGPT's reasoning process. As such, it may be necessary for its developers to provide further clarification and explanation in this regard.

4.5 The Effects of Prompts

In this section, our goal was to compare the results from all five prompts. For Prompts 3 and 5, which generated a probability, we chose a threshold of 0.7, 0.7, and 0.9 for identifying hateful, offensive, and toxic content, respectively. This threshold was determined based on the results of

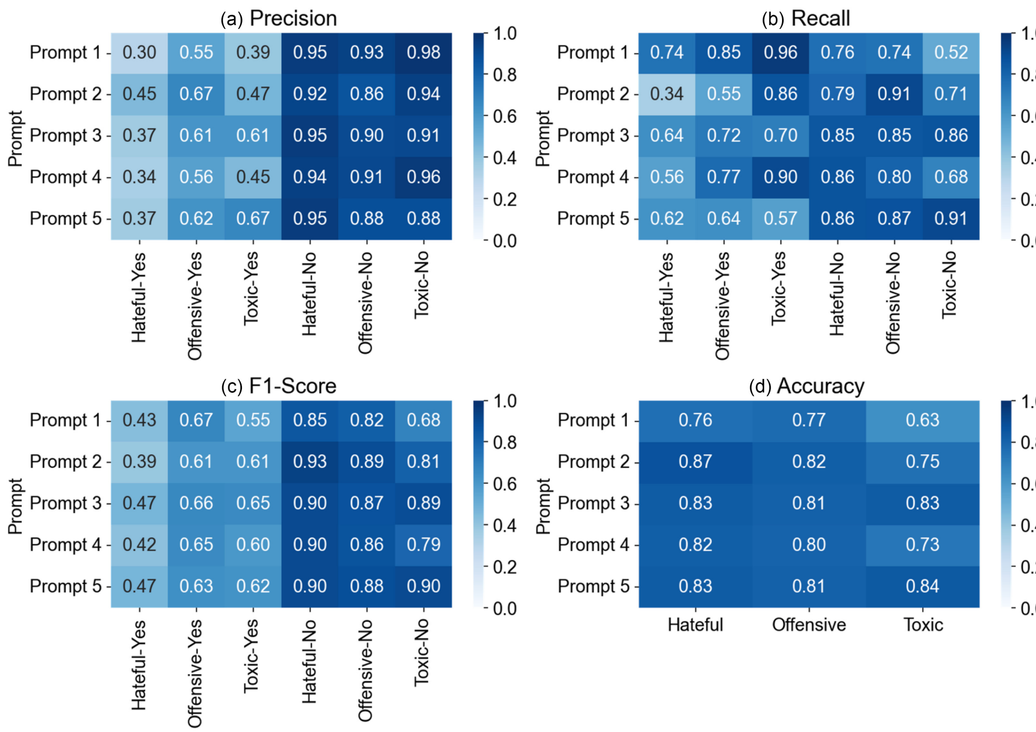


Fig. 7. Classification performance of all five prompts as compared to MTurker annotations. (a) Precision. (b) Recall. (c) F1-score. (d) Accuracy.

Experiment 2. To compare the outcomes generated from different prompts, we drew heatmaps to visualize Precision, Recall, F1-score, and Accuracy (see Figure 7). Among all the prompts, Prompt 1 generated the worst overall performance. This result highlights the importance of providing ChatGPT with clear instructions, such as specifying a binary classification or a probability, to enhance its performance on annotation tasks.

For the binary output, a comparison between Prompt 2 and Prompt 4 revealed that ChatGPT classified more comments as HOT when asked to provide reasoning for the HOT dataset. This result is supported by the greater number of HOT classifications in Table 6 compared to Figure 3. Moreover, the model's indication for "hateful" and "offensive" classifications shows a slight improvement, while non-HOT classifications show a slight decrease in F1-scores, as indicated in Figure 7.

For the probability output, comparing Prompt 3 and Prompt 5 showed that ChatGPT is less inclined to assign an extremely high or low probability score (e.g., a smaller number of 0 and 1 classifications) when asked to provide reasoning, as evidenced by the score distributions in Figure 4 and Figure 6. Nevertheless, the performance of these two prompts is similar.

Upon comparing the binary and probability outputs, we observed that the classification of "toxicity" displays a clear improvement in the probability prompts, particularly for the non-toxic class, as demonstrated by the higher F1-score for Toxicity-No in Figure 7. This improvement could be attributed to the lower threshold for classifying toxicity, which, upon being raised, led to better performance compared to that of MTurkers. Moreover, we observed that adjusting the threshold for "hateful" can also affect the performance, as evidenced by the higher F1-score for Hateful-Yes in Figure 7. However, we did not observe any noticeable improvement in the other classes.

5 DISCUSSION

Motivated by the work from Davidson et al. (2017) [5] and Wu et al. (2023) [26] emphasizing the importance of distinguishing HOT concepts, we set out to measure how well ChatGPT performs on HOT content annotation tasks and to understand how its reasoning concerning the annotations it provides. We conducted four experiments in which we varied the information we provided the model, the type of response we requested, and whether we requested reasoning for the output. We found that ChatGPT can obtain roughly 80% accuracy compared to human crowd workers in identifying HOT content, and ChatGPT provides consistent results. ChatGPT also parroted our prompts when providing reasoning for its decisions.

1. ChatGPT provides reliable and consistent responses. We found that ChatGPT can obtain an accuracy of approximately 80% when compared to MTurker annotations. In general, ChatGPT exhibited a more consistent classification with MTurkers for non-HOT comments, as evidenced by higher F1-scores but less agreement for HOT comments. Notably, ChatGPT showed significant disagreement with MTurkers when classifying "hateful" comments despite being provided with definitions of hateful content. We observed that ChatGPT is generally consistent. It provided the same response more than 90% of the time, even when changing the temperature setting, particularly for simpler output requests like binary classification, compared to complex requests such as probability. Our observation regarding reliability and consistency is generally consistent with [55].

2. ChatGPT repeats HOT definitions for its reasoning. Our findings suggest that ChatGPT may have a lower threshold to label a comment as "toxic" in the HOT concept. As a result, the "toxic" classification included a greater number of comments classified as "offensive" and "hateful" compared to the MTurker classification, as illustrated by the Venn diagrams in Table 2 and Table 10. Additionally, we observed that ChatGPT repeated our definitions to explain its classification for HOT concepts. For instance, ChatGPT at times categorized a comment as "hateful" if it targets a specific group of people based on their race, ethnicity, religion, or other attributes, as demonstrated by the reasoning analysis in Experiment 3 (see Table 11 and examples in Appendix A.6). Regarding its reasoning format, we found that ChatGPT frequently used certain words, such as "does not" or "not contain," for "unlikely" classifications, and "contains" and "does" for the "likely" classifications. The "uncertain" category included more phrases like "could be perceived."

3. Prompts affect ChatGPT's performance. We found that the choice of prompts to interact with ChatGPT can impact its performance. For example, Experiment 1 illustrates that when no context or instructions are provided for data annotation, the model exhibited low agreement with MTurkers. This was observed when the output of ChatGPT was unrestricted. Additionally, we observed that the model may have varying thresholds for HOT classification. We found that prompts using probability tended to perform slightly better than those using binary classification when we used higher thresholds for "toxic" classifications (e.g., 0.9 instead of 0.7). Specifically, we found higher F1-scores and accuracy for classifying "toxic" comments, as shown in Figure 7.

5.1 Practical Implications for Researchers using ChatGPT to Annotate HOT Content

We noted four practical implications for using ChatGPT to annotate HOT content. First, ChatGPT could apply our definitions of hateful, offensive, and toxic accurately and consistently when we provided the definitions. Therefore, we suggest providing explicit definitions when eliciting classifications from ChatGPT. Similarly, prompts should also indicate whether ChatGPT should provide a binary or probability and an explanation for its response. The next two implications address these prompts.

We observed that ChatGPT hesitated to provide a probability between 0.3 and 0.8, resulting in a much lower number of outputs falling within this range. This result suggests that researchers should exercise caution when using probability, particularly those in the "uncertain" range, as they may not accurately reflect the level of HOT present in comments. We also noted that ChatGPT may generate fewer classifications of extreme probability (e.g., 0 or 1) when asked to provide reasoning. Researchers can request this additional information if they require a more cautious outcome.

Requesting explanations may raise the likelihood that ChatGPT classifies a comment as HOT. Therefore, requesting explanations is a potentially useful way to receive a conservative output (i.e., more HOT classifications) from the model. However, it is worth noting that this result does not necessarily imply a stronger agreement between ChatGPT and human annotators. This is supported by the fact that the F1-scores and accuracy from Prompts 4 and 5 do not show a clear improvement over Prompts 2 and 3, as shown in Figure 7.

Finally, our study has practical implications for researchers who are interested in using ChatGPT to annotate samples for building detection models or classify large-volume data, such as social media data. We demonstrated that ChatGPT can generate reliable results in comparison to MTurkers, particularly for non-HOT classifications. However, for HOT classifications, we noticed that the selected threshold of probability can significantly impact the model's outputs, particularly for "toxic" classifications, and the assignment of "hateful" classifications depends on the definitions provided to the model. We recommend setting a low temperature to achieve more deterministic results. This recommendation is also in line with Gilardi et al. (2023) [55] in that a lower temperature yields more accurate annotations.

5.2 Practical Implications for using ChatGPT for Online Content Moderation

We noted several practical implications for using ChatGPT to annotate HOT content. First, ChatGPT was able to provide consistent outputs even when we changed the temperature parameter. Making consistent decisions is crucial for moderating online content. Users' perception that social media platforms make inconsistent decisions [80] has made them think the decisions are unfair. Our results were consistent when we used the same model version to obtain all of our outputs. It is unclear how the output might change with different model versions (e.g., gpt-4). Therefore, practitioners should be careful about the model versions they use for making decisions and, wherever possible, validate the model's consistency with past decisions before making any changes.

Our results also show that ChatGPT was more precise at classifying not-HOT comments than HOT comments (see Table 7). ChatGPT was more likely to misclassify not-HOT comments as HOT than the other way around. Practically, this implies that using ChatGPT to identify HOT content for taking potential actions may lead to over-moderation, where platforms ultimately take more action than is desired. Over-moderation can hinder the platform's growth and drive members away [81] while also causing additional harm to certain minoritized populations [82]. Through Experiment 2, we show how the accuracy of classifying HOT content can be improved by setting different thresholds. For instance, setting a higher threshold on the probability generated by the ChatGPT model improved the model's precision at classifying HOT comments (see Figure 5). Therefore, asking the model to generate probability outputs (instead of binary outputs) could potentially provide platforms with more control over their decisions. However, as prior research has argued, platforms must be careful with setting the thresholds for taking their actions, as both over (i.e., flagging or removing content that should be harmless) and under-moderation (i.e., overlooking content that should be flagged or removed) can cause different types of harm [83].

From a broader viewpoint, especially considering a recent news report that a federal judge in the United States blocked administration agencies and officials from communicating with social media companies about "protected speech" [84], we would like to highlight several potential concerns tied to the use of AI tools like ChatGPT for content moderation. Despite the potential of over and under-moderation, it is important to acknowledge these AI models might encounter difficulties in accurately comprehending context. "Context" encompasses components like sarcasm, humor, cultural subtleties, and regional colloquialisms, areas where AI may misconstrue the context, potentially leading to improper content moderation recommendations. AI models' effectiveness is intrinsically linked to the quality and comprehensiveness of their training data. If there are gaps or inaccuracies in the training data, the resulting AI classifications will reflect these limitations [85]. Moreover, depending on how these AI models are implemented, there could be concerns about invasions of privacy if content and conversations are being scanned by AI systems [86].

5.3 Limitations and Future Work

We addressed three types of limitations. First, we compared the ChatGPT annotations with MTurker annotations by calculating several performance metrics. However, MTurker annotations may not necessarily represent appropriate or accurate "ground-truth" classifications. Our findings show only the degree of agreement between ChatGPT and MTurkers; we did not evaluate the accuracy of ChatGPT's HOT classifications relative to any other ground truth. Future research could involve the use of expert knowledge to annotate our HOT dataset or the use of other datasets, such as the DATASET [5], OffensEval [33], Wikipedia Toxicity Corpus [87], or the Gab Hate Corpus [88]. Annotators vary in their sensitivity and interpretation of these concepts, and ChatGPT may more closely mirror another group of annotators' classifications.

Next, we acknowledge two limitations associated with the prompts used to interact with ChatGPT and its outputs. First, we applied only one set of definitions for the HOT concepts. Using different definitions for the concepts would likely produce different results. The second limitation relates to how ChatGPT interprets prompts. We identified instances where ChatGPT produced unexpected outputs. For example, the model occasionally returned multiple classifications for a single comment. We also found some of the explanations provided by ChatGPT merely repeated the definitions that we provided, raising questions about the reasonableness and representativeness of the model's explanations. To address this limitation, future work could investigate the impact of using different definitions, such as definitions used by the Moderation API [93], to interact with ChatGPT.

The third type of limitation in our work relates to outputs and their meanings. While probability can provide some indication of the likelihood that a comment is classified as HOT, not all probabilities are informative. For instance, probabilities of 0.01, 0.05, or 0.95 only reflect the ChatGPT's assessment that a comment is highly likely or unlikely to be classified as HOT, but they do not provide much granularity about the model's degree of certainty. Furthermore, we have doubts about whether certain probabilities can accurately reflect the degree of HOT, particularly those in the "uncertain" range. For example, it remains unclear whether a comment with a probability of 0.6 is more likely to be classified as HOT than one with a probability of 0.4, as ChatGPT could interpret both probabilities as indicating that the comment may be perceived as HOT by certain individuals. Therefore, it may be necessary to develop additional methods to complement the use of probability to provide more nuanced insights into the degree of HOT. Possible avenues for future research could include exploring more contextual information to improve the accuracy and interpretability of the model's outputs.

Lastly, we note that several LLMs (e.g., PaLM, LLaMA2, Falcon) incorporate safety filters that limit their utility for labeling HOT content. When passed content the models considered inappropriate, it failed to respond to our prompts. We used gpt-3.5-turbo in part because we could adjust filters to ensure the model completed our tasks. Off-the-shelf and commercial LLM APIs will not be useful for content moderation if their safety guardrails prevent them from classifying content, but finding the right balance between labeling content and preventing abuse of the model will require additional research.

Several avenues of future work deserve further investigation. We plan to investigate the impact of MTurker annotators' demographics on the performance of ChatGPT. Specifically, we propose to test whether ChatGPT shows better agreement with annotations from certain demographics than others. It is also worth investigating whether ChatGPT can provide demographic-dependent answers, as prior work has shown that certain demographic users may systematically differ in their annotations of HOT concepts. Overall, this line of research can help us to identify any biases or limitations in the model's ability to classify HOT comments across different demographic groups.

In relation to the first avenue, future work could also consider integrating measures of an individual's exposure to HOT content and evaluate its impact. Personal experiences of hatred and harassment could significantly shape the evaluations of HOT content. This could manifest in the individual's sensitivity to, or perception of, what is considered HOT content. Therefore, it could be worthwhile to examine the interaction between an individual's exposure and the identification of HOT content, for example, designing experiments to identify factors (e.g., frequency of exposure) and assess their impact.

Next, we can build on prior studies [94, 95], which indicate that annotators' demographic characteristics could influence their evaluations. Another future work could address the persona and context elements in annotation tasks. Such research could include demographics like gender and ethnicity in the design of the prompt and further evaluate their potential impact on the model's outputs.

Last, given the rapid development and shifting landscapes of generative AI models, we plan to test the performance of ChatGPT model's performance through time [96], as well as other models in comparison to ChatGPT, such as the LLaMa model [97], using broader datasets related to harmful textual content. This will allow us to compare the strengths and limitations of different generative AI models and potentially identify promising approaches to improving the accuracy and interpretability of HOT classification.

6 CONCLUSION

This study investigated the potential of using generative AI models for annotating HOT comments and compares its results with those from MTurkers. Our findings show that ChatGPT exhibits approximately 80% accuracy and provides the same response more than 90% of the time in terms of reliability and consistency. It displays a more consistent classification with MTurkers for non-HOT comments but less agreement for HOT comments. It shows significant disagreement with MTurkers when classifying "hateful" comments. When requesting probability outputs, we found that ChatGPT provides a greater number of results with extreme probabilities for HOT comments, and probabilities between 0.3 and 0.7 are less common. Next, our n-gram analysis shows that ChatGPT conforms to some of the provided definitions, but its generalization ability for reasoning needs further study. We found that different prompts can affect ChatGPT's performance. Requesting an explanation could probably receive more conservative outputs. In summary, we suggest that ChatGPT can be a useful tool for annotating large samples of content quickly and cheaply. However, to receive high-quality and useful annotations, researchers need to take care in defining the classes of HOT content and in designing the prompt they provide ChatGPT.

A APPENDIX

A.1 Inter-coder Agreement between MTurkers

Table A.1 shows the Krippendorff's α to show the inter-coder agreement between MTurkers for labeling HOT content. The table is obtained from Schöpke-Gonzalez et al. (2023) [74]. This table serves to understand the level of consistency of the HOT dataset published by Wu et al. (2023) [26].

Table A.1. Krippendorff's α of Inter-coder Agreement (Schöpke-Gonzalez et al. 2023) [74]

Category	Krippendorff's α (with bootstrap)			Krippendorff's α (without bootstrap)
	25%	50%	75%	
Hateful	0.40	0.43	0.46	0.43
Offensive	0.47	0.50	0.53	0.50
Toxic	0.41	0.44	0.48	0.44

A.2 Prompt Descriptions

Below are the details of prompts we provided to ChatGPT to implement our experiments.

Prompt 1

Provide the HOT concept and ask: "Do you think this comment is hateful? (1) Yes, (2) No." This is the same as how we asked MTurkers.

(1) Do you think this comment is hateful? (1) Yes, (2) No. A comment is "hateful" if you perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."
(2) Do you think this comment is offensive? (1) Yes, (2) No. A comment is "offensive" if you perceive that it is "a hurtful, derogatory or obscene comment."
(3) Do you think this comment is toxic? (1) Yes, (2) No. A comment is "toxic" if you perceive that it is "a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion."

Prompt 2

Provide the HOT concept to ChatGPT and only ask for a binary classification without further explanations.

(1) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is hateful. A comment is "hateful" if you perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." I want you to only respond with yes or no. Do not provide any other outputs or any explanation for your output.
(2) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is offensive. A comment is "offensive" if you perceive that it is "a hurtful, derogatory or obscene comment." I want you to only respond with yes or no. Do not provide any other outputs or any explanation for your output.
(3) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the comment is toxic. A comment is "toxic" if you perceive that it is "a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion." I want you to only respond with yes or no. Do not provide any other outputs or any explanation for your output.

Prompt 3

Provide the HOT concept to ChatGPT and ask for a probability representing HOT without further explanations.

(1) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being hateful. A comment is "hateful" if you

perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." A probability of 1 means that the comment is highly likely to be hateful. A probability of 0 means that the comment is highly unlikely to be hateful. Do not provide any other outputs or any explanation for your output.

(2) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being offensive. A comment is "offensive" if you perceive that it is "a hurtful, derogatory or obscene comment." A probability of 1 means that the comment is highly likely to be offensive. A probability of 0 means that the comment is highly unlikely to be offensive. Do not provide any other outputs or any explanation for your output.

(3) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being toxic. A comment is "toxic" if you perceive that it is "a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion." A probability of 1 means that the comment is highly likely to be toxic. A probability of 0 means that the comment is highly unlikely to be toxic. Do not provide any other outputs or any explanation for your output.

Prompt 4

Provide the HOT concept to ChatGPT and only ask for a binary classification with further explanations.

(1) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is hateful. A comment is "hateful" if you perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." I want you to respond with yes or no in the first line and provide an explanation for your output in the second line.

(2) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the following comment is offensive. A comment is "offensive" if you perceive that it is "a hurtful, derogatory or obscene comment." I want you to respond with yes or no in the first line and provide an explanation for your output in the second line.

(3) I want you to perform a data annotation task. In your output, I want you to return either yes or no, depending on whether you think the comment is toxic. A comment is "toxic" if you perceive that it is "a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion." I want you to respond with yes or no in the first line and provide an explanation for your output in the second line.

Prompt 5

Provide the HOT concept to ChatGPT and ask for a probability representing HOT with further explanations.

(1) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being hateful. A comment is "hateful" if you perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." A probability of 1 means that the comment is highly likely to be hateful. A probability of 0 means that the comment is highly unlikely to be hateful. I want you to respond with a probabilistic score in the first line and provide an explanation for your score in the second line.

(2) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being offensive. A comment is "offensive" if you perceive that it is "a hurtful, derogatory or obscene comment." A probability of 1 means that the comment is highly likely to be offensive. A probability of 0 means that the comment is highly unlikely to be offensive. I want you to respond with a probabilistic score in the first line and provide an explanation for your score in the second line.

(3) I want you to provide a probability score between 0 to 1 where the score represents the probability of the following comment being toxic. A comment is "toxic" if you perceive that it is "a rude, disrespectful, or unreasonable comment that is likely to make readers want to leave a discussion." A probability of 1 means that the comment is highly likely to be toxic. A probability of 0 means that the comment is highly unlikely to be toxic. I want you to respond with a probabilistic score in the first line and provide an explanation for your score in the second line.

A.3 Comparison Results based on Prompt 2

Table A.2 displays the comparison results where the annotations from MTurkers were treated as a score. To ensure a fair assessment between ChatGPT's probability output and the MTurkers' score, we initially applied a specific rule to transform ChatGPT's probability output. This approach aimed to establish a standardized comparison between the two scoring methods, ensuring an "apple-to-apple" comparison. The probability output is based on Prompt 2.

- $0.0 \leq prob < 0.1 \Leftrightarrow 5$ "False" and 0 "True" from MTurkers (score = 0.0)
- $0.1 \leq prob < 0.3 \Leftrightarrow 4$ "False" and 1 "True" from MTurkers (score = 0.2)
- $0.3 \leq prob < 0.5 \Leftrightarrow 3$ "False" and 2 "True" from MTurkers (score = 0.4)
- $0.5 \leq prob < 0.7 \Leftrightarrow 2$ "False" and 3 "True" from MTurkers (score = 0.6)
- $0.7 \leq prob < 0.9 \Leftrightarrow 1$ "False" and 4 "True" from MTurkers (score = 0.8)
- $0.9 \leq prob \leq 1.0 \Leftrightarrow 0$ "False" and 5 "True" from MTurkers (score = 1.0)

Table A.2. Comparison between MTurker Annotation and ChatGPT Output on Prompt 2

Category	Score	Support	Precision	Recall	F1-score	Accuracy
Hateful	0.0	2219	0.94	0.36	0.52	0.36
	0.2	567	0.19	0.53	0.28	
	0.4	276	0.17	0.15	0.16	
	0.6	182	0.11	0.07	0.09	
	0.8	142	0.12	0.42	0.18	
	1.0	77	0.14	0.35	0.20	
Offensive	0.0	1599	0.89	0.35	0.50	0.35
	0.2	607	0.23	0.57	0.33	
	0.4	399	0.13	0.06	0.08	
	0.6	319	0.16	0.07	0.10	
	0.8	325	0.20	0.38	0.26	
	1.0	214	0.31	0.56	0.39	
Toxic	0.0	1542	0.92	0.13	0.22	0.19
	0.2	695	0.18	0.29	0.23	
	0.4	426	0.12	0.06	0.08	
	0.6	359	0.07	0.03	0.05	
	0.8	267	0.06	0.19	0.09	
	1.0	171	0.16	0.87	0.28	

A.4 The Impact of Text Variation based on Prompt 2

In addition to Experiment 3, we conducted an additional experiment to measure consistency. This experiment aimed to examine consistency across variations of Prompt 2, achieved through subtle rephrasing of the same prompt to preserve meaning while varying the expression. For example, we replaced "yes or no" with "true or false" in the instructions, shown below.

> I want you to perform a data annotation task. In your output, I want you to return either True or False, depending on whether you think the following comment is hateful. A comment is "hateful" if you perceive that it "expresses hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." I want you to only respond with True or False. Do not provide any other outputs or any explanation for your output.

Noting that the version of GPT model could affect the output [96], we ran the experiment again using Prompt 2 to eliminate the effect of the version on the model performance. Still, we analyzed the variation of the model's outputs using Krippendorff's α. The results are presented in Table A.3.

Table A.3. Consistency (Krippendorff's α) of ChatGPT's Performance in Experiment 3 for Prompt 2 with Text Variations

Category	Support	Agreement %	Krippendorff's α
Hateful	3480	94.9% (3402 of 3480)	0.78
Offensive	3474	95.1% (3305 of 3474)	0.86
Toxic	3473	95.8% (3327 of 3473)	0.92

A.5 Classification Performance of Different Models based on Prompt 2

To validate our choice of using ChatGPT for annotating HOT content, we also tested two other LLMs, namely Falcon and PaLM, and another traditional toxicity detection tool called Perspective API. Falcon is an open-source LLM developed by the Technology Innovation Institute based in Abu Dhabi [98], while PaLM is a recently released LLM that builds on Google's research in ML and responsible AI [99]. In our study, we used the Falcon-40B model [98] and PaLM 2 [100] model to run Prompt 2. Perspective API was developed by Jigsaw and Google's Counter Abuse Technology team. Perspective API uses ML models to estimate the probability that individuals would perceive a given comment toxic [101].

First, we found that Falcon-40B was frequently incapable of delivering a desired output as "yes" or "no". A common output of Falcon is, "*As an AI language model, I cannot perceive the context of the comment and cannot determine whether it is offensive or not. Therefore, I cannot provide a response to this task.*" This limited our ability to interpret its results or evaluate the model's performance in comparison to the annotations provided by MTurkers.

Second, the PaLM 2 model did have the capability to produce desired responses such as "yes" or "no." Nevertheless, due to PaLM's API having inherent protective measures against core harm content [100], such as the content threatening child safety, these types of harm are always blocked and cannot be adjusted from the user's side. For multiple instances in our dataset, PaLM 2 outputs, "*I'm not able to help with that, as I'm only a language model. If you believe this is an error, please send us your feedback.*" Consequently, there exists a portion of comments that the PaLM 2 model was not able to identify. Based on the outputs of PaLM 2, we followed the same method shown in Section 4.2 to calculate the Recall, Precision, and F1-score.

Third, the Perspective API returns a probability score that represents the share of a sample who would find the comment toxic. For instance, a score of 0.8 implies that 8 out of 10 people perceive that comment as toxic. The Perspective API documentation suggests researchers experimenting with a threshold between 0.7 and 0.9 to classify toxicity [101]. We experimented with the threshold equals to 0.7, 0.8, and 0.9, respectively. The results are presented in Table A.4.

Table A.4. Classification Performance of HOT based on Prompt 2 of Different Models

Model	Category	Class	Support	Precision	Recall	Accuracy
ChatGPT (GPT3.5)	Hateful	Yes	404	0.45	0.34	0.87
		No	3077	0.92	0.79	
	Offensive	Yes	862	0.67	0.55	0.82
		No	2616	0.86	0.91	
	Toxic	Yes	801	0.47	0.86	0.75
		No	2675	0.94	0.71	
PaLM 2	Hateful	Yes	268	0.42	0.53	0.89
		No	2672	0.95	0.93	
	Offensive	Yes	572	0.51	0.71	0.82
		No	2450	0.92	0.84	
	Toxic	Yes	554	0.39	0.92	0.72
		No	2468	0.97	0.68	
Perspective API	Toxic	Yes	803	0.67	0.26	0.80
(threshold = 0.7)		No	2675	0.81	0.96	
Perspective API	Toxic	Yes	803	0.79	0.14	0.79
(threshold = 0.8)		No	2675	0.79	0.99	
Perspective API	Toxic	Yes	803	0.93	0.03	0.78
(threshold = 0.9)		No	2675	0.78	1.00	

A.6 Annotated Samples based on Prompt 4

Table A.5 displays samples that provide insight into ChatGPT's classification of HOT content, to support the discussion in Section 4.4. In particular, Table A.5 presents specific examples that were classified as "toxic" but not "hateful" or "offensive," "toxic" and "offensive" but not "hateful," and "offensive" but not "hateful" or "toxic."

Table A.5. Examples of ChatGPT's Outputs based on Prompt 4

Comment	ChatGPT's Response by Concept		
	Hateful	Offensive	Toxic
So should you Mr. Graham!! You are a traitor!!	No. The comment does not express hatred towards a targeted group. It is directed towards an individual and accuses them of being a traitor. While it may be insulting or derogatory towards the individual, it does not target a specific group of people based on their race, ethnicity, religion, or other characteristic.	Yes. The comment contains derogatory language ("traitor") directed towards Mr. Graham, which can be perceived as hurtful and offensive.	Yes. The comment contains disrespectful language and personal attacks towards Mr. Graham, which can make readers feel uncomfortable and want to leave the discussion. Therefore, it can be considered toxic.

(Continued)

Table A.5. Continued

Comment	ChatGPT's Response by Concept		
	Hateful	Offensive	Toxic
Dumb governor	No. The comment is not directed towards a specific group and does not express hatred towards any particular group. It is a general insult towards a governor, but it does not target any specific group or community.	Yes. The comment is derogatory as it uses the word "dumb" to insult the governor. It is a negative and hurtful comment that is intended to belittle and demean the person being referred to. Therefore, it can be considered offensive.	Yes. The comment is disrespectful and uses a derogatory term to refer to the governor. It is likely to offend and provoke negative reactions from readers, making them want to leave the discussion. Therefore, it meets the criteria for a toxic comment.
Blatant lies show they have zero shame.	No. The comment does not express hatred towards a targeted group or intend to be derogatory, humiliate, or insult any members of a group. It is a statement criticizing someone's behavior or actions.	No. The comment is critical and accusatory, but it does not contain any explicit derogatory or obscene language. It is a statement of opinion rather than a personal attack.	Yes. The comment contains disrespectful language and implies that the person being referred to has no shame. This type of language is likely to make readers feel uncomfortable and want to leave the discussion.
I am going to stop paying my vehicle and mortgage loan. I think the taxpayer should pay it.	No. The comment does not express hatred towards a targeted group or intend to be derogatory, humiliate, or insult any members of a group. It is simply a statement about the author's personal financial decisions and opinion on who should pay for their loans.	No. The comment is not offensive as it does not contain any hurtful, derogatory, or obscene language. It may be considered irresponsible or unethical, but it does not meet the criteria for being offensive.	Yes. This comment is toxic because it is disrespectful and unreasonable. The commenter is suggesting that they will stop paying their own debts and instead expect taxpayers to foot the bill. This kind of attitude is likely to provoke negative reactions from other readers and could lead to a breakdown in civil discussion.
This is the biggest clown show I've ever seen. The world is laughing at us.	No. The comment does not express hatred towards a targeted group. It is a criticism of a situation or event, and does not target any specific group of people. While it may be insulting or derogatory towards those responsible for the situation, it does not aim to humiliate or insult any particular group.	No. The comment is critical and negative, but it does not contain any hurtful, derogatory, or obscene language. It expresses the author's opinion about a situation, but it does not attack or insult any individual or group.	Yes. The comment contains disrespectful language and is likely to make readers want to leave a discussion. The use of the word "clown" is derogatory and the statement "the world is laughing at us" is inflammatory.
Look at his face though, he's so big and strong that he'll DEFINITELY beat it. /s	No. The comment does not express hatred towards a targeted group or intend to be derogatory, humiliate, or insult any members of a group. The comment is a sarcastic remark about someone's physical appearance and their ability to beat something. The "/s" at the end of the comment indicates that it is meant to be taken as sarcasm or a joke.	Yes. The comment is offensive because it is making a derogatory remark about someone's physical appearance and implies that their size and strength are the only factors that matter in a situation. Additionally, the use of the word "beat" suggests violence, which is also inappropriate.	No. The comment does not contain any rude, disrespectful, or unreasonable language that is likely to make readers want to leave a discussion. The use of sarcasm at the end of the comment indicates that it is not meant to be taken seriously.

(Continued)

Table A.5. Continued

Comment	ChatGPT's Response by Concept		
	Hateful	Offensive	Toxic
Just ban her she's the dumbest person in America.	Yes. The comment is derogatory and insulting towards a specific person, which is a form of hate speech. The use of the word "dumbest" is intended to humiliate and insult the person, which is not acceptable behavior.	Yes. This comment is offensive because it is derogatory and hurtful towards the person mentioned. The use of the word "dumbest" is insulting and implies that the person is unintelligent. Additionally, the suggestion to ban someone based on their perceived intelligence is discriminatory and exclusionary. Overall, this comment is not constructive and serves only to attack and belittle the individual.	Yes. This comment is toxic because it is rude, disrespectful, and unreasonable. It attacks a specific person and uses derogatory language, which is likely to make readers feel uncomfortable and want to leave the discussion.
Can't fix stupid, well, I guess covid can.	Yes. The comment is derogatory and insulting towards individuals who may have contracted COVID-19. The phrase "Can't fix stupid" implies that those who have contracted the virus are somehow stupid or deserving of their illness. The second part of the comment, "well I guess covid can," suggests that the speaker believes that COVID-19 is a punishment for stupidity. This type of language is harmful and contributes to the stigmatization of individuals who have contracted the virus.	Yes. The comment is offensive because it uses derogatory language to insult someone's intelligence and makes light of a serious global pandemic that has caused immense suffering and loss of life. It is insensitive and lacks empathy towards those who have been affected by COVID-19.	Yes. The comment is disrespectful and uses a derogatory term to refer to people. It also makes light of a serious situation (COVID-19 pandemic), which can be seen as insensitive and inappropriate. Overall, the comment is likely to offend and discourage readers from engaging in a discussion.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

We want to express our gratitude to Angela Gonzalez and Siqi Wu for their valuable viewpoints on this project. Their expertise and knowledge were instrumental in our analysis. Additionally, we appreciate their provision of the HOT dataset, which enabled us to conduct our experiments.

REFERENCES

[1] L. Fan, H. Yu, and Z. Yin. 2020. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. *Proc. Assoc. Inf. Sci. Technol.* 57, 1 (2020). DOI: https://doi.org/10.1002/pra2.313

[2] E. Whittaker and R. M. Kowalski. 2015. Cyberbullying via social media. *Journal of School Violence* 14, 1 (2015), 11–29. DOI: https://doi.org/10.1080/15388220.2014.949377

[3] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams. 2019. "The enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web* 13, 3 (2019), 1–26. DOI: https://doi.org/10.1145/3324997

[4] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, Florence, Italy, 29–30. DOI: https://doi.org/10.1145/2740908.2742760

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. *ICWSM* 11, 1 (2017), 512–515. DOI: https://doi.org/10.1609/icwsm.v11i1.14955

[6] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *AAAI* 35, 17 (2021), 14867–14875. DOI: https://doi.org/10.1609/aaai.v35i17.17745

[7] Z. Talat, J. Thorne, and J. Bingel. 2021. Correction to: Bridging the gaps: Multi task learning for domain transfer of hate speech entection. In *Online Harassment*, J. Golbeck, Ed., In *Human–Computer Interaction Series*. Springer International Publishing, Cham, C1–C1. DOI: https://doi.org/10.1007/978-3-319-78583-7_12

[8] M. Ibrahim, M. Torki, and N. El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Orlando, FL, 875–878. DOI: https://doi.org/10.1109/ICMLA.2018.00141

[9] M. Diaz, I. Kivlichan, R. Rosen, D. Baker, R. Amironesei, V. Prabhakaran, and E. Denton. 2022. CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency.*ACM, Seoul, Republic of Korea, 2342–2351. DOI: https://doi.org/10.1145/3531146.3534647

[10] F. Huang, H. Kwak, and J. An. 2023. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. (2023). DOI: https://doi.org/10.48550/ARXIV.2302.07736

[11] T. Sorensen, J. Robinson, C. Rytting, A. Shaw, K. Rogers, A. Delorey, M. Khalil, N. Fulda, and D. Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 819–862. DOI: https://doi.org/10.18653/v1/2022.acl-long.60

[12] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans. Visual. Comput. Graphics* (2022), 1–11. DOI: https://doi.org/10.1109/TVCG.2022.3209479

[13] Y. Zhou, A. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. 2023. Large language models are human-level prompt engineers. arXiv, Mar. 10, 2023. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2211.01910

[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, 145–153. DOI: https://doi.org/10.1145/2872427.2883062

[15] M. Das, B. Mathew, P. Saha, P. Goyal, and A. Mukherjee. 2020. Hate speech in online social media. *SIGWEB Newsl.* Autumn, 1–8 (2020). DOI: https://doi.org/10.1145/3427478.3427482

[16] Md. A. H. Wadud, M. M. Kabir, M. F. Mridha, M. A. Ali, Md. A. Hamid, and M. M. Monowar. 2022. How can we manage offensive text in social media - a text classification approach using LSTM-BOOST. *International Journal of Information Management Data Insights* 2, 2 (2022), 100095. DOI: https://doi.org/10.1016/j.jjimei.2022.100095

[17] B. Gambäck and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 85–90. DOI: https://doi.org/10.18653/v1/W17-3013

[18] A. Matamoros-Fernández and J. Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media* 22, 2 (2021), 205–224. DOI: https://doi.org/10.1177/1527476420982230

[19] Z. Mossie and J.-H. Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management* 57, 3 (2020), 102087. DOI: https://doi.org/10.1016/j.ipm.2019.102087

[20] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. 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*. Association for Computational Linguistics, Florence, Italy, 1668–1678. DOI: https://doi.org/10.18653/v1/P19-1163

[21] N. Andalibi, O. L. Haimson, M. D. Choudhury, and A. Forte. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Trans. Comput.-Hum. Interact.* 25, 5 (2018), 1–35. DOI: https://doi.org/10.1145/3234942

[22] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, 11–20. DOI: https://doi.org/10.18653/v1/W18-5102

[23] P. Fortuna and S. Nunes. 2019. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51, 4 (2019), 1–30. DOI: https://doi.org/10.1145/3232676

[24] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2 (2018), 1–33. DOI: https://doi.org/10.1145/3185593

[25] D. Paschalides, D. Stephanidis, A. Andreou, K. Orphanou, G. Pallis, M. Dikaiakos, and E. Markatos. 2020. MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.* 20, 2 (2020), 1–21. DOI: https://doi.org/10.1145/3371276

[26] S. Wu, A. Schöpke-Gonzalez, S. Kumar, L. Hemphill, and P. Resnick. 2023. HOT speech: Comments from political news posts and videos that were annotated for hateful, offensive, and toxic content. *Inter-university Consortium for Political and Social Research [distributor]*, (2023). [Online]. Available: https://socialmediaarchive.org/record/19
[27] S. Duguay, J. Burgess, and N. Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252. DOI: https://doi.org/10.1177/1354856518781530
[28] J. T. Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution.* 2000.
[29] Meta. 2023. Hate Speech. [Online]. Available: https://transparency.fb.com/policies/community-standards/hate-speech/
[30] Twitter. 2023. Hateful Conduct. [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[31] J. Salminen, M. Hopf, S. A. Chowdhury, S. Jung, H. Almerekhi, and B. J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Hum. Cent. Comput. Inf. Sci.* 10, 1 (2020). DOI: https://doi.org/10.1186/s13673-019-0205-6
[32] M. Wiegand, M. Siegel, and J. Ruppendorfer. 2018. *Konvens 2018 - GermEval Proceedings.* Verlag der Österreichischen Akademie der Wissenschaften, 2018, p. 0xc1aa5576_0x003a105d. DOI: https://doi.org/10.1553/0x003a105d
[33] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation.* Association for Computational Linguistics, Minneapolis, Minnesota, USA, 75–86. DOI: https://doi.org/10.18653/v1/S19-2010
[34] T. Jay and K. Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture* 4, 2 (2008). DOI: https://doi.org/10.1515/JPLR.2008.013
[35] Google Jigsaw. 2017. Perspective API. 2017. [Online]. Available: https://perspectiveapi.com
[36] V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics* 4, 2 (2020), 155–190. DOI: https://doi.org/10.1007/s41701-019-00065-w
[37] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing.* IEEE, Amsterdam, Netherlands, 71–80. DOI: https://doi.org/10.1109/SocialCom-PASSAT.2012.55
[38] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. 2019. Detection of abusive language: The problem of biased datasets. In *Proceedings of the 2019 Conference of the North.* Association for Computational Linguistics, Minneapolis, Minnesota, 602–608. DOI: https://doi.org/10.18653/v1/N19-1060
[39] N. D. Gitari, Z. Zhang, H. Damien, and J. Long. 2015. A lexicon-based approach for hate speech detection. *IJMUE* 10, 4 (2015), 215–230. DOI: https://doi.org/10.14257/ijmue.2015.10.4.21
[40] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* Association for Computational Linguistics, New Orleans, Louisiana, 1046–1056. DOI: https://doi.org/10.18653/v1/N18-1095
[41] A. Tontodimamma, L. Fontanella, S. Anzani, and V. Basile. 2022. An Italian lexical resource for incivility detection in online discourses. *Qual. Quant.* 2022. DOI: https://doi.org/10.1007/s11135-022-01494-7
[42] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* ACM, New Orleans, LA USA, 67–73. DOI: https://doi.org/10.1145/3278721.3278729
[43] T. Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press, New Haven, Connecticut.
[44] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion.* : ACM, Perth, Australia, 759–760. DOI: https://doi.org/10.1145/3041021.3054223
[45] D. Chatzakou, I. Leontiadis, J. Blackburn, E. Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis. 2019. Detecting cyberbullying and cyberaggression in social media. *ACM Trans. Web* 13, 3 (2019), 1–51. DOI: https://doi.org/10.1145/3343484
[46] C. Chelmis and D.-S. Zois. 2021. Dynamic, incremental, and continuous detection of cyberbullying in online social media. *ACM Trans. Web* 15, 3 (2021), 1–33. DOI: https://doi.org/10.1145/3448014
[47] P. Malik, A. Aggrawal, and D. K. Vishwakarma. 2021. Toxic speech detection using traditional machine learning models and BERT and FastText embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC).* IEEE, Erode, India, 1254–1259. DOI: https://doi.org/10.1109/ICCMC51019.2021.9418395

[48] Z. Yin, L. Fan, H. Yu, and A. J. Gilliland. 2020. Using a three-step social media similarity (TSMS) mapping method to analyze controversial speech relating to COVID-19 in Twitter collections. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, Atlanta, GA, USA, 1949–1953. DOI: https://doi.org/10.1109/BigData50022.2020.9377930

[49] N. Badri, F. Kboubi, and A. H. Chaibi. 2022. Combining FastText and GloVe word embedding for offensive and hate speech text detection. *Procedia Computer Science* 207 (2022), 769–778. DOI: https://doi.org/10.1016/j.procs.2022.09.132

[50] A. C. Mazari, N. Boudoukhani, and A. Djeffal. 2023. BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Comput.* (2023). DOI: https://doi.org/10.1007/s10586-022-03956-x

[51] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi. 2022. L3Cube-MahaHate: A tweet-based Marathi hate speech detection dataset and BERT models. (2022). DOI: https://doi.org/10.48550/ARXIV.2203.13778

[52] S. Khan, M. Fazil, V. Sejwal, M. Alshara, R. Alotaibi, A. Kamal, and A. Baig. 2022. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University - Computer and Information Sciences* 34, 7 (2022), 4335–4344. DOI: https://doi.org/10.1016/j.jksuci.2022.05.006

[53] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE* 14, 8 (2019), e0221152. DOI: https://doi.org/10.1371/journal.pone.0221152

[54] M. J. Riedl, G. M. Masullo, and K. N. Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior* 107 (2020), 106262. DOI: https://doi.org/10.1016/j.chb.2020.106262

[55] F. Gilardi, M. Alizadeh, and M. Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv, Mar. 27, 2023. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2303.15056

[56] B. Kasthuriarachchy, M. Chetty, A. Shatte, and D. Walls. 2021. Cost effective annotation framework using zero-shot text classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Shenzhen, China, 1–8. DOI: https://doi.org/10.1109/IJCNN52387.2021.9534335

[57] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, (Eds.), Curran Associates, Inc., Dec. 2020, 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[58] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners, 2019. [Online]. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[59] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill. 2023. A bibliometric review of large language models research from 2017 to 2023, 2023. DOI: https://doi.org/10.48550/ARXIV.2304.02020

[60] Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI, (2023). [Online Video]. Available: https://www.youtube.com/watch?v=L_Guz73e6fw

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Presented at the 31st Conference on Neural Information Processing Systems.* Long Beach, CA, USA: arXiv, 2017. DOI: https://doi.org/10.48550/ARXIV.1706.03762

[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[63] A. Feder, N. Oved, U. Shalit, and R. Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics.* 1–54. DOI: https://doi.org/10.1162/coli_a_00404

[64] OpenAI. 2023. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses Quick links Try on ChatGPT Plus. [Online]. Available: https://openai.com/product/gpt-4

[65] S. Pichai. 2023. An important next step on our AI journey. [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/

[66] Y.-S. Wang and Y. Chang. 2022. Toxicity detection with generative prompt-based inference. arXiv, 2022. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2205.12390

[67] I. Pettersson. 2022. Keeping tabs on GPT-SWE: Classifying toxic output from generative language models for Swedish text generation, diva-portal.org, 2022. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1704893&dswid=391

[68] A. Kucharavy, Z. Schillaci, L. Maréchal, M. Würsch, L. Dolamic, R. Sabonnadiere, D. David, A. Mermoud, and V. Lenders. 2023. Fundamentals of generative large language models and perspectives in cyber-defense. arXiv, Mar. 21, 2023. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2303.12132

[69] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T.

Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. 2023. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv, Nov. 22, 2022. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2209.07858

[70] Y. Ji, Y. Gong, Y. Peng, C. Ni, P. Sun, D. Pan, B. Ma, and X. Li. 2023. Exploring ChatGPT's ability to rank content: A preliminary study on consistency with human preferences. arXiv, Mar. 13, 2023. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2303.07610

[71] R. Sridhar and D. Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, Seattle WA, United States, 811–826. DOI: https://doi.org/10.18653/v1/2022.naacl-main.59

[72] OpenAI, "Models." [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[73] K. Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology.* California SAGE Publications, Inc., Thousand Oaks. DOI: https://doi.org/10.4135/9781071878781

[74] A. Schöpke-Gonzalez, S. Wu, S. Kumar, P. J. Resnick, and L. Hemphill. 2023. How we define harm impacts data annotations: Explaining how annotators distinguish hateful, offensive, and toxic comments. (2023). DOI: https://doi.org/10.48550/ARXIV.2309.15827

[75] E. M. Humphries, C. Wright, A. M. Hoffman, C. Savonen, and J. T. Leek. 2023. What's the best chatbot for me? Researchers put LLMs through their paces. *Nature.* DOI: https://doi.org/10.1038/d41586-023-03023-4

[76] T. Kuzman, I. Mozetič, and N. Ljubešić. 2023. ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. (2023). DOI: https://doi.org/10.48550/ARXIV.2303.03953

[77] M. V. Reiss. 2023. Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. (2023). DOI: https://doi.org/10.48550/ARXIV.2304.11085

[78] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. (2023). DOI: https://doi.org/10.48550/ARXIV.2302.11382

[79] J. Lever, M. Krzywinski, and N. Altman. 2016. Model selection and overfitting. *Nat. Methods* 13, 9 (2016), 703–704. DOI: https://doi.org/10.1038/nmeth.3968

[80] S. Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383. DOI: https://doi.org/10.1177/1461444818773059

[81] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (2019), 1–35. DOI: https://doi.org/10.1145/3338243

[82] T. Dias Oliva, D. M. Antonialli, and A. 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. DOI: https://doi.org/10.1007/s12119-020-09790-w

[83] S. Atreja, L. Hemphill, and P. Resnick. 2022. Remove, reduce, inform: What actions do people want social media platforms to take on potentially misleading content?. (2022). DOI: https://doi.org/10.48550/ARXIV.2202.00799

[84] K. Singh. 2023. US judge restricts Biden officials from contact with social media firms. *Reuters.* Accessed: Sep. 29, 2023. [Online]. Available: https://www.reuters.com/legal/judge-blocks-us-officials-communicating-with-social-media-companies-newspaper-2023-07-04/

[85] Q. V. Liao, D. Gruen, and S. Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu, HI, USA, 1–15. DOI: https://doi.org/10.1145/3313831.3376590

[86] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj. 2023. From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, 11 (2023), 80218–80245. DOI: https://doi.org/10.1109/ACCESS.2023.3300381

[87] E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web, Perth Australia: International World Wide Web Conferences Steering Committee* (2017), 1391–1399. DOI: https://doi.org/10.1145/3038912.3052591

[88] B. Kennedy, M. Atari, A. Davani, L. Yeh, A. Omrani, Y. Kim, K. Jr., S. Havaldar, G. Portillo-Wightman, E. Gonzalez, J. Hoover, A. Azatian, A. Hussain, A. Lara, G. Cardenas, A. Omary, C. Park, X. Wang, C. Wijaya, Y. Zhang, B. Meyerowitz, and M. Dehghani. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resources & Evaluation* 56, 1 (2022), 79–108. DOI: https://doi.org/10.1007/s10579-021-09569-x

[89] E. Denton, M. Díaz, I. Kivlichan, V. Prabhakaran, and R. Rosen. 2021. Whose ground truth? Accounting for individual and collective identities underlying dataset annotation. (2021). DOI: https://doi.org/10.48550/ARXIV.2112.04554

[90] P. S. Sachdeva, R. Barreto, C. von Vacano, and C. J. Kennedy. 2022. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 1585–1603. DOI: https://doi.org/10.1145/3531146.3533216

[91] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. 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*. Association for Computational Linguistics, Seattle, WA, United States, 5884–5906. DOI: https://doi.org/10.18653/v1/2022.naacl-main.431

[92] Z. Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. DOI: https://doi.org/10.18653/v1/W16-5618

[93] Moderation API. 2023. [Online]. Available: https://moderationapi.com/content-moderation

[94] J. Pei and D. Jurgens. 2023. When do annotator demographics matter? Measuring the influence of annotator demographics with the POPQUORN Dataset. arXiv, Aug. 28, 2023. Accessed: Sep. 29, 2023. [Online]. Available: http://arxiv.org/abs/2306.06826

[95] E. Denton, M. Díaz, I. Kivlichan, V. Prabhakaran, and R. Rosen. 2023. Whose ground truth? Accounting for individual and collective identities underlying dataset annotation. arXiv, Dec. 08, 2021. Accessed: Sep. 29, 2023. [Online]. Available: http://arxiv.org/abs/2112.04554

[96] L. Chen, M. Zaharia, and J. Zou. 2023. How is ChatGPT's behavior changing over time? arXiv, Aug. 01, 2023. Accessed: Sep. 29, 2023. [Online]. Available: http://arxiv.org/abs/2307.09009

[97] Meta AI, Introducing LLaMA: A foundational, 65-billion-parameter large language model. [Online]. Available: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

[98] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, and G. Penedo. 2023. Falcon-40B: An open large language model with state-of-the-art performance. Jun. 20, 2023. Accessed: Sep. 29, 2023. [Online]. Available: https://huggingface.co/tiiuae/falcon-40b

[99] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. Dai, T. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. 2022. PaLM: Scaling language modeling with pathways. 2022. DOI: https://doi.org/10.48550/ARXIV.2204.02311

[100] Google developers. 2023. Google AI PaLM 2, Google AI. Accessed: Sep. 29, 2023. [Online]. Available: https://ai.google/discover/palm2/

[101] Perspective developers. 2023. Perspective API. 2023. [Online]. Available: https://perspectiveapi.com/research/

Received 3 May 2023; revised 30 September 2023; accepted 12 December 2023