NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps

Kristina Gligorić Myra Cheng Lucia Zheng Esin Durmus Dan Jurafsky
Stanford University
gligoric@cs.stanford.edu

Abstract

Warning: content in this paper may be upsetting or offensive.

The use of words to convey speaker's intent is traditionally distinguished from the 'mention' of words for quoting what someone said, or pointing out properties of a word. Here we show that computationally modeling this use-mention distinction is crucial for dealing with counterspeech online. Counterspeech that refutes problematic content often mentions harmful language but is not harmful itself (e.g., calling a vaccine dangerous is not the same as expressing disapproval of someone for calling vaccines dangerous). We show that even recent language models fail at distinguishing use from mention, and that this failure propagates to two key downstream tasks: misinformation and hate speech detection, resulting in censorship of counterspeech. We introduce prompting mitigations that teach the use-mention distinction, and show they reduce these errors. Our work highlights the importance of the use-mention distinction for NLP and CSS and offers ways to address it.

1 Introduction

The **use-mention distinction** is the difference between using words (*Bananas have a peel*) and mentioning them (*"Bananas" has 7 letters*, or *Dan said "Bananas"*). The distinction has long been important in the philosophy of language (Sperber and Wilson, 1981; Saka, 1998), where discussions date back to Tarski (1931) and Quine (1940).

The ability to correctly make this distinction is particularly relevant to dealing with problematic language online. While detecting problematic text has become a standard task in NLP, there has been less consideration of text that might **mention** harmful content without directly **using** it. Mentioning problematic content is critical to language like counterspeech that challenges or opposes hateful or misleading narratives (Table 1) (Wright et al., 2017; Mun et al., 2023; Hangartner et al., 2021; Ecker et al., 2022). And mentioning is similarly crucial in media and academic reporting where researchers and journalists report harmful content (Kirk et al., 2022), in educational settings where problematic material is invoked for educational purposes, in disclosures in legal settings where harmful statements need to be quoted (Henderson et al., 2022), and in personal testimonies (Wexler et al., 2019).

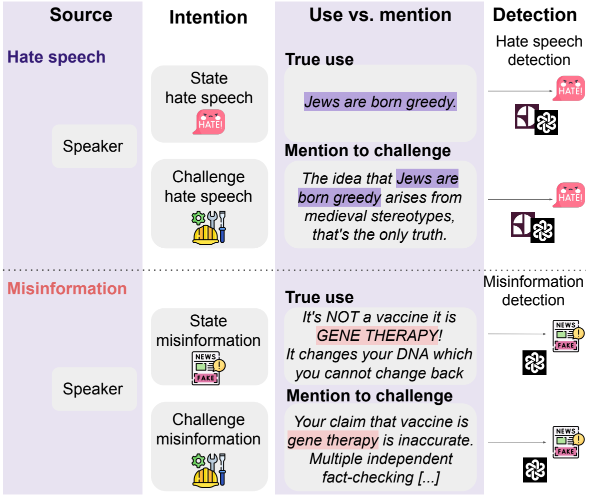

Figure 1: **Illustration of harmful censorship of counterspeech.** Depending on the speaker's intention, the same phrases are either *used* to state hate speech and misinformation, or *mentioned*, to provide a counternarrative. Algorithms that indiscriminately classify counterspeech as abusive censor constructive discourse.

In this work, we focus on the first of these: **online counterspeech**, defined as speech produced by users of online platforms to counteract harmful speech of others. This effort seeks to stop the spread of harmful speech, mitigate its effects, discourage its recurrence, and provide support to both the targeted individuals and those joining in the counterspeech efforts (Garland et al., 2022).

Counterspeech statements often involve refer-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5942–5959
June 16-21, 2024 ©2024 Association for Computational Linguistics

Hate speech statements, use $\{u_t\}$	**Counterspeech, mention $\{m_t\}$**
Our children are **being forced fed** with this 'Islam' nonsense in schools [...]	What do you mean by **being forced fed**? I haven't seen anything like this.
Jews are born greedy	The charge that **Jews are born greedy** arises from medieval stereotypes and has nothing to do with the actual financial practices of Jews.
I'm tired of woman taking part to political debate. **Feminism is a divisive force in society**.	Of course, if you've never read anything about feminism except what its detractors purport that it's about, then you come to believe **feminism is a divisive force for society**. But if you open yourself to different point of view you will change your mind, trust me!
Misinformation statements, use $\{u_t\}$	**Counterspeech, mention $\{m_t\}$**
My new book says, "a vaccine, while not impossible, may be far off." Then people say, "Wait Jim, there is a vaccine." Actually, no. What we have is **an experimental gene therapy**. The mRNA shot is something like chemotherapy for people who don't have cancer.	The claim that the vaccine is **an experimental gene therapy** is inaccurate. The vaccine only uses mRNA for immunization goal. It does not intertact with the gene at any levels. Please do not share wrong information.
Chief Justice Mogoeng of South Africa calls COVID vaccine "satanic ," a **plot to alter DNA** to inscribe mark of Antichrist on everyone. These delusions may start in US, they spread through the world	"The **plot to alter DNA**" is wrong. As mentioned by CDC, the vaccine only uses mRNA, and do not interact with DNA. The Chief is misquoted by something, You should convey true information.

Table 1: **Examples of statements using problematic language, and paired statements mentioning it.** Focal phrases are highlighted (used on the left, and mentioned on the right).

ring to or quoting problematic content (Vidgen et al., 2021). Errors in distinguishing use from mention might therefore lead to failures in downstream classification, making counterspeech statements more likely to be misclassified as harmful by modern NLP systems, as shown in Figure 1. Yet counterspeech helps curb online abuse (Bonaldi et al., 2022) and make online spaces safer (Siegel and Badaan, 2020). Thus, erroneously classifying counterspeech as problematic leads to content removal with significant implications: misclassification erases opportunities to rectify false narratives, and in doing so, risks further censoring those already most affected by harmful language (Sahoo et al., 2022; Rahman, 2012; Park et al., 2018).

But addressing the use-mention distinction and assessing its impact on downstream tasks is challenging. First, reasoning about the distinction itself has been hard, due to the lack of datasets, resources, and quantitative measurement methods. The few prior studies have been small and limited to linguistic features like particular mention verbs (Wilson, 2011b). Second, online counterspeech generally occurs in informal contexts, where markings that formally indicate mention, such as quotation marks or italics, are often missing (Wilson, 2011a). Finally, since mentioned language is less frequent than use (Wilson, 2011b), it is easy for researchers to overlook the downstream performance of NLP systems on mentioned language.

Motivated by these technical challenges, failure cases in counterspeech (Figure 1 and Table 1), and literature on lexical and topic biases in harmful content detection (Dixon et al., 2018; Ethayarajh et al., 2022), we make the following hypotheses:

H1 NLP models fail to distinguish use from mention in counterspeech.

H2 Failure in distinguishing use from mention impacts downstream tasks like hate speech and misinformation detection.

H3 What is considered 'permissible' by hate speech or misinformation classifiers is influenced by the presence of identity terms and other targeted entities, as well as the strength of the stance expressed toward mentioned language.

H4 Downstream performance can be improved by teaching the use-mention distinction to explicitly encode the treatment of mentioned language.

Testing these hypotheses, our contributions include:

(1) Use-mention tasks We formalize two tasks with challenging use-mention examples: (1) use versus mention classification and (2) downstream hate speech and misinformation detection (Sec.3).

(2) Failure analyses We identify use-mention distinction failures, show that errors propagate to downstream tasks, and trace failures to specific target entities (misinformation-related and identity terms) and to the strength of the expressed stance (Sec.4–4.3).

(3) Mitigations We investigate prompting mitigations and implications for downstream tasks. We show that our interventions lead to a significant reduction in error (Sec. 4.4).

To support the further evaluation and development of models, our code is available at https://github.com/kristinagligoric/use-mention. Datasets are publicly available and can be used for research purposes.

2 Background

Counterspeech Building upon social science literature showing that counternarratives are effective against hate speech (Andrews, 2002; Benesch, 2014; Schieb and Preuss, 2016; Garland et al., 2022), previous work in NLP and HCI has studied counterspeech from various perspectives. Scholars have curated datasets of counterspeech from different sources, including experts, NGO workers, and social media comments (Mathew et al., 2019; Chung et al., 2019; Qian et al., 2019; Garland et al., 2020; Fanton et al., 2021). Others have investigated different types of counterspeech strategies and performed user studies to evaluate the effectiveness of machine-generated counterspeech (Mun et al., 2023; Fraser et al., 2023). Our work also builds upon existing models for counterspeech detection (Garland et al., 2020).

Content moderation policies on counterspeech mentions Several online platforms acknowledge the importance of counterspeech in their content policies. At the time of writing, TikTok's Community Principles states "*We do not allow language or behavior that harasses, humiliates, threatens, or doxxes anyone. This also includes responding to such acts with retaliatory harassment (but excludes non-harassing counter speech)*" (TikTok, 2023). Similarly, Facebook publisher and creator guidelines state "*We know that many publishers use Facebook to challenge ideas, institutions and practices. Such discussion can promote debate and greater understanding*". The guidelines also provide advice on how to write counterspeech to avoid mislabeling as hate speech (Meta, 2023). Development of models that enable enforcement of such policies is thus pressing.

While previous work has not explored the impact that the use-mention distinction has on online text classification tasks, mentions of toxic phrases were described as one class of false positive errors in toxic comment classification (Van Aken et al., 2018), e.g., "*I deleted the <identity group> are dumb comment*", establishing that errors due to use-mention distinction are prevalent.

The use-mention distinction The use-mention distinction has been studied in philosophy (Sperber and Wilson, 1981), computational linguistics (Wilson, 2010, 2011b; Behzad et al., 2023), and HCI (Anderson et al., 2002). In general, mention is defined as follows:

Definition 1 *"for a token or a set of tokens t in a sentence S, if t refers to a property of the token t, then t is an instance of mention." (Wilson, 2010)*

While many facts about a token can be a mention property, in our domain of counterspeech we focus on two properties of mentioned language (Sandhan et al., 2023; Wilson, 2011b): **attributed language** (e.g., mentions to refer to the quotes of original source or stance towards the source) and **words or phrases as themselves** (e.g., mentions of words to refer to stance towards their use).[1]

The use-mention distinction and related tasks Like the use-mention distinction, other NLP tasks bear on speaker intent, such as those related to factuality (Saurí and Pustejovsky, 2012, 2009; Murzaku et al., 2022) and committed belief (Prabhakaran et al., 2010, 2015; De Marneffe et al., 2019). In the context of mentioned language, such tasks aim to directly take into account the speaker's intention when mentioning specific phrases, and in particular, whether the speech act commits the speaker to the truth or factuality of the expressed proposition. However, in both used and mentioned language, statements do not necessarily commit the speaker to the truth of the source statement or to its factuality. Moreover, harmful language is often implicit, formalized as questions and nuanced statements that need not constitute a committed belief or be factual either. The distinction between using and mentioning is thus related to but distinct from factuality and committed belief.

[1] See the Limitations section on other types of mentioned language which do not lead as clearly to practical harms.

Lastly, mentioned language is also an instance of metalanguage (Behzad et al., 2023; Perlis et al., 1998; Wilson, 2012, 2013), and our task builds upon work on similar tasks like distinguishing whether personal names are being used to mention or address (Prabhakaran et al., 2023).

3 Methods

3.1 Tasks

We focus on language indicative of hateful speech or misinformation, two frequent and connected types of harmful online content (Mosleh et al., 2024), which can be used or mentioned to express disapproving attitude towards it (as illustrated in Fig 1). We hypothesize that use-mention distinction failures cause harmful misclassifications of counterspeech on downstream tasks. To test this hypothesis, we operationalize two tasks: the use-mention distinction and downstream classification.

Task 1: Use-mention classification For a given text, the task is to classify whether hateful/misinformative language is used or whether it is mentioned. True use U are statements which use hate or misinformation, and M are counterspeech mentions. For each text $u \in U$, $m \in M$, we classify the text as either use (positive class) or mention (negative class). Metrics we report are **false positive rate** and **false negative rate** in detecting use, and **average error rate**, capturing the average of the two rates. False positives are mentions misclassified as uses, while false negatives are defined as uses misclassified as mentions. On Task 1, false positives (mistaking mention for use) are the errors of primary interest due to their hypothesized impact on downstream tasks.

Task 2: Downstream classification where use-mention distinction matters We address two important downstream sub-tasks: hate speech detection and misinformation detection, with challenging use-mention examples. Similarly, for each text $u \in U$, $m \in M$, we classify each statement as either the positive ("misinformation" and "hate speech") or negative class ("not misinformation" and "not hate") on the downstream task. Since the standard metric is false positive rate capturing how often non-harmful content is misclassified as harmful (Dixon et al., 2018; Markov and Daelemans, 2021), we report **false positive rate** (mentions misclassified as "misinformation" or "hate"). We also report **false negative rate** (uses misclassified as "not misinformation" or "not hate"), and **average error rate**, capturing the average of the two rates. An ideal model would classify all u_i as "misinformation" or "hate speech" (0% false negative rate), while all counterspeech statements m_i would be classified as "not misinformation" or "not hate" (0% false positive rate). On Task 2, false positive rate on M is the pragmatic concern central to our investigation as it captures the censorship rate.

3.2 Datasets

Countering hate For this task we rely on two datasets: Knowledge-grounded hate countering (Chung et al., 2021) and Multi-Target Counternarratives (Fanton et al., 2021). The datasets contain pairs of (hateful statement, counterspeech), illustrated in Table 1. We select counterspeech statements written by human experts.

Countering misinformation For this task we leverage misinformation counternarratives (He et al., 2023). The dataset contains pairs of (misinformative statement, counterspeech), illustrated in Table 1. Misinformative statements were posted on social media, while counter-responses are a mix of naturally occurring social media posts and counterspeech statements written by recruited human participants.

Focal tokens To confirm that mentioned language is indeed relevant to the practical case of counterspeech, we verified that counterspeech mentions contain language from the original true use sample it addresses, which we refer to as focal tokens. Across $N = 1826$ pairs $\{(u_t, m_t)\}$, we computed the length of the longest common substring using a dynamic programming algorithm (Suzgun et al., 2023b). We found substantial overlap (as in examples in Table 1, focal tokens), with average $M = 3.44$ words in the longest common substring. Additionally, for both datasets, we manually verified that focal tokens are used in true use statements (and not mentioned), and that counterspeech is not using, but mentioning (see Appendix, Sec. A for details).

3.3 Models

For the use-mention task, we tested the two best performing GPT models at the time of writing (gpt-3.5-turbo, gpt-4), as well as gpt-3.5 instruct, the non-RLHF legacy variant, using zero-shot prompting. For downstream tasks, we tested these three models as well as four widely-used models for

Sub-task	Model	False positive rate ↓	False negative rate ↓	Average error rate ↓
Hate speech	gpt-3.5-instruct-turbo	17.98 ± 7.55	14.77 ± 6.73	16.38 ± 5.34
	gpt-3.5-turbo (ChatGPT 3.5)	6.82 ± 4.37	20.00 ± 6.38	13.48 ± 3.91
	gpt-4	20.00 ± 8.57	4.44 ± 4.34	**12.22** ± 4.74
Misinformation	gpt-3.5-instruct-turbo	34.38 ± 3.76	40.08 ± 3.33	37.22 ± 2.21
	gpt-3.5-turbo (ChatGPT 3.5)	8.05 ± 1.83	49.76 ± 3.17	28.93 ± 1.86
	gpt-4	23.44 ± 2.74	3.89 ± 1.38	**13.64** ± 1.42

Table 2: **Use-mention classification.** False positive, false negative, and average error rates in detecting use (in percentages). Models are sorted by the average error rate. The best performing model has a high false positive rate, mistaking mention for use, which can lead to censorship of useful counterspeech.

Sub-task	Model	False positive rate ↓	False negative rate ↓	Average error rate ↓
Hate speech	toxigen hatebert	24.44 ± 10.0	77.78 ± 7.22	51.11 ± 8.08
	perspective (insult)	4.44 ± 3.89	61.11 ± 9.85	32.78 ± 6.27
	perspective (toxicity)	20.00 ± 7.51	36.67 ± 9.18	28.33 ± 8.49
	perspective (identity attack)	21.11 ± 7.51	33.33 ± 9.74	27.22 ± 9.18
	roberta hate speech	17.78 ± 7.51	26.67 ± 8.65	22.22 ± 8.22
	gpt-3.5-instruct-turbo	25.56 ± 8.62	13.33 ± 7.51	19.44 ± 8.20
	gpt-3.5-turbo (ChatGPT 3.5)	11.11 ± 6.14	22.22 ± 8.07	16.67 ± 7.24
	gpt-4	8.89 ± 5.00	20.00 ± 7.51	**14.44** ± 7.09
Misinformation	roberta fake news	97.93 ± 1.00	5.10 ± 1.52	51.52 ± 1.20
	gpt-3.5-instruct-turbo	26.12 ± 2.40	19.44 ± 2.62	22.78 ± 2.80
	gpt-3.5-turbo (ChatGPT 3.5)	22.11 ± 3.02	13.85 ± 2.74	17.98 ± 3.06
	gpt-4	10.21 ± 2.00	8.02 ± 1.85	**9.11** ± 1.93

Table 3: **Downstream tasks (hate speech and misinformation detection).** False positive, false negative, and average error rates in detecting hate speech and misinformation (in percentages). Models are sorted by the average error rate. Across models, substantial errors persist in false positive rate (classifying counterspeech mentions as harmful).

hate speech and misinformation detection: Perspective (Perspective, 2023), Toxigen (Hartvigsen et al., 2022), RoBERTa (Liu et al., 2019), and RoBERTa fake news (Ahmed et al., 2017).

For prompting, we use default parameters (temperature=1) and max output token length 1 (outputting either A or B). The classification prompt includes the instruction and a definition of the classes (for prompt variants and complete prompt text see Appendix, Section D).

4 Results

4.1 Use-mention classification (H1)

How well do the models distinguish whether problematic language is used or mentioned? Across the models (Table 2), average error rates are high (between 12.22% and 16.38% for hate speech and between 13.64% and 37.22% for misinformation), suggesting that state-of-the-art large language models struggle to distinguish use from mention in domains where the distinction matters. The best-performing model, gpt-4, still has a very high false positive rate, mistaking mention for use; mentions are misclassified as use in **20.00%** of hate speech and **23.44%** of misinformation counternarratives. In these settings, mistaking mention for use leads to the consequential harm of censoring useful counterspeech.

4.2 Downstream content classification (H2)

The prior section showed that state-of-the-art systems often fail at the use-mention distinction. Here we test the impact on hate speech and misinformation detection, two downstream classification tasks where the use-mention distinction matters (Table 3).

Performance on downstream tasks First, on the hate speech detection task, we find that misclassification of counterspeech as hateful is relatively frequent: the popular Toxigen and Perspective API have a false positive rate on counterspeech of **over 20%**, and recent models still have many false positives, although somewhat reduced, e.g., gpt-3.5-turbo has FPR 11.11%. Regarding the average error rate, gpt-4 is the best-performing model (false positive rate on counterspeech **8.89%**).

Model	Use-mention ¬correct	Use-mention correct	χ^2	p
gpt-3.5-instruct-turbo	**32.96%**	19.52%	20.58	5.72×10^{-6}
gpt-3.5-turbo (ChatGPT 3.5)	**28.31%**	14.44%	25.08	5.51×10^{-7}
gpt-4	**15.78%**	4.54%	30.60	3.17×10^{-8}

Table 4: **Error propagation.** False positive rate in downstream classification of counterspeech mentions, stratified by use-mention classification correctness. Error rates on counterspeech are significantly higher when use-mention distinction is incorrect. Statistics for hate speech and misinformation separately are listed in the Appendix, Table 10.

Second, on the misinformation detection task, misclassification of counterspeech as misinformative is relatively frequent: gpt-3.5 models have **over 20%** false positive rate on counterspeech, and substantial errors persist even with the best system, gpt-4 (**10.21%** false positive rate classifying counterspeech mentions as misinformation). We also note that two investigated models, Togixen and RoBERTa fake news, have average error rates above 50%.

Error propagation to downstream tasks As a further test of the hypothesis that errors in distinguishing use from mention cause these failures in downstream tasks, we assess whether examples that cause errors downstream (Task 2) are also misclassified in use-mention classification (Task 1). We do this by testing how the downstream error rate on Task 2 (hate speech detection and misinformation detection) associates with the error rate on Task 1 (use vs. mention). For gpt-4, for example, we contrast the error rate on Task 2 of 15.78% (aggregated over hate speech and misinformation detection) for statements in which the Task 1 classifier failed at distinguishing use from mention, versus the error on Task 2 of 4.54% (aggregated) for statements in which the Task 1 classifier correctly distinguished use from mention, comparing these rates using the chi-squared test.

We find that, among samples where mentioning counterspeech is misclassified as use, downstream misclassification is higher across the three tested models (all $p < 10^{-5}$; Table 4). Results disaggregated by hate speech and misinformation detection are listed in the Appendix (Table 10).

In summary, we find that errors in use-mention distinction do propagate to downstream tasks.

4.3 Why is the distinction hard (H3)?

Counterspeech that opposes harmful narratives is by definition not harmful. What linguistic aspects of this counterspeech cause downstream NLP tools to incorrectly label counterspeech as harmful? Or, viewed from the other perspective, what terms do NLP tools make permissible as mentions?

Target identity	gpt-3.5-turbo	gpt-4
Jewish	14.15%	12.15%
People of color	9.09%	9.09%
Muslims	6.80%	4.83%
LGBT+	6.77%	2.22%
Disabled	4.00%	0.00%
Women	2.41%	1.18%
Other	0.89%	1.77%
Migrants	0.39%	0.76%

Table 5: **Identity.** False positive rate in hate speech classification of counterspeech, stratified by target identity.

Hate speech: Target identity terms We know that toxicity detection algorithms rely heavily on the presence of identity mentions to make their predictions (Zhou et al., 2021). For the hate speech task, we therefore hypothesize that identity terms will impact when mentioned language is permissible. To test this hypothesis, we stratify error rate in classifying counterspeech as hate speech by target identity (as labeled in the metadata of the dataset (Chung et al., 2021)).

We find that errors vary widely depending on the targeted identity (e.g., gpt-3.5-turbo: Jewish 14.15%, people of color 9.09%, Muslims 6.80%, LGBT+ 6.77%, while for the other groups it is less than 5%; Table 5). Even the most recent large language models' treatment of counterspeech similarly varies depending on the mentioned identity. Our results suggest that systems treat the mere mention of certain identity terms as impermissible (see Sec.5).

Misinformation: COVID-19 terms and strength in stance towards the embedded language To understand why counterspeech mentioning misinformation is detected as misinformation, we use the Fightin' Words method (Monroe et al., 2008) to measure statistically significant differences in tokens between two sets—counterspeech mentions classified as misinformation vs. not classified as

Model	Top terms for $D_{\times}$	Top terms for $D_{\checkmark}$
gpt-4	please (-3.67), vaccine (-3.61), therapy (-3.53), gene (-3.32), misinformation (-3.23), stop (-2.82), mrna (-2.52), dna (-2.43), uses (-2.41), wrong (-2.27), change (-2.10), correct (-2.03), safe (-2.01), cdc (-1.99), well (-1.93)	fake (5.03), news (4.89), lying (4.27), lies (4.16), one (3.94), misleading (3.84), lie (3.76), put (3.43), thing (3.35), always (3.13), false (3.10), new (2.98), everything (2.83), anyway (2.83), country (2.83)
gpt-3.5	mrna (-4.24), vaccine (-3.40), dna (-2.82), uses (-2.82), change (-2.77), please (-2.17), genes (-2.15), genetic (-2.12), actually (-2.09), gene (-2.06), therapy (-2.06), understand (-2.02), correct (-1.86), immunization (-1.86), nothing (-1.79)	fake (5.39), misleading (5.08), lying (4.31), lies (4.20), media (4.09), news (4.03), another (3.60), report (3.39), journalism (3.37), blood (2.94), false (2.92), crazy (2.65), justice (2.64), remove (2.48), reporting (2.45)

Table 6: **Top 15 terms for $D_{\times}$ and $D_{\checkmark}$.** Z-scores (in parentheses) are computed using the Fightin' Words method. Words with |z-score| > 1.96 are statistically significant in frequency difference. Misclassification of counterspeech mentions as harmful is influenced by specific COVID-19-related terms ("mRNA", "vaccine"), and strength in stance towards mentioned language ("fake", "misleading").

misinformation—after controlling for variance in words' frequencies.

We compute the top differentiating words for $D_{\checkmark}$ versus $D_{\times}$, where

$$D_{\checkmark} = \{m_i \in M | \text{misinformation}_c(m_i) = \text{True}\}$$

$$D_{\times} = \{m_i \in M | \text{misinformation}_c(m_i) = \text{False}\}$$

respectively, where M are all counterspeech statements and misinformation$_c$ is misinformation classification by classifier c.

We find that terms relating to controversial topics are often misclassified; top terms for $D_{\times}$ include "gene therapy", "mRNA", "vaccine", "DNA", and "CDC", suggesting that counterspeech mentioning COVID vaccination is often misclassified. Certain terms associated with COVID-19 misinformation are not permissible even when mentioned.

We hypothesize that due to interactions with safety features that prevent large language models from generating health disinformation (Menz et al., 2024), the mere presence of specific terms related to COVID vaccination is linked with misclassification of counterspeech as misinformative. This finding that terms related to COVID are treated as 'impermissible' when mentioned parallels our finding that mentions of certain demographic identities are impermissible to hate speech detectors.

We also find that top terms for $D_{\checkmark}$ are related to expressing a strong stance against misinformation in the surrounding context and the strength of meta language (as indicated by terms such as "fake news", "lying", "lies", "misleading"). Downstream classification has fewer errors when the disagreement in mentioning statements is not subtle.

Distancing by using quotation marks Counterspeech that uses verbatim quotes typically involves more severe language, as quotes enable distancing (Wilson, 2011a). We hypothesize that such texts are more censored. Indeed, we find for both hate speech and misinformation that the counterspeech containing quotation marks is more frequently misclassified as harmful (Table 7).

Summary of Error Analysis In summary, misclassification of counterspeech mentions as harmful is influenced by over-reliance on (a) **surface terms** such as identity words and specific COVID-19-related terms, and (b) notions of **strength in stance towards mentioned language**.

4.4 How can we teach the distinction (H4)?

Methods Informed by the analyses revealing that errors propagate from the inability to distinguish use from mention into misclassification on the downstream tasks, we explore a set of prompting mitigations to reduce downstream mistakes. In particular, we explore ways to teach the use-mention distinction through controlled prompting.

To that end, we designed and tested CoT prompting mitigation. In `CoT + mitigation`, we (1) embed the definition of use-mention distinction and an instruction specifying that mention of hateful or misinformative language does not imply that the text is hateful or misinformative. We use prompt formats inspired by BigBench CoT (Suzgun et al., 2023a) and (2) follow the process prompting the LLM to "*think step-by-step*" (Wei et al., 2022) and use the answer extraction prompt "*so the answer is*" to the generated rationale to extract the final answer. We also (3) include few-shot examples of mentioned and used language where the first step considers whether potentially problematic language is used or mentioned, before making the classification in the second step (for complete prompt

Sub-task	Model	Mention quotations	¬Mention quotations	χ^2	p
Hate speech	gpt-3.5-instruct-turbo	**57.14%**	22.89%	3.98	0.046
	gpt-3.5-turbo (ChatGPT 3.5)	28.57%	9.88%	2.24	0.13
	gpt-4	28.57%	7.23%	3.63	0.056
Misinformation	gpt-3.5-instruct-turbo	30.00%	26.12%	0.15	0.70
	gpt-3.5-turbo (ChatGPT 3.5)	**45.00%**	21.79%	6.05	0.014
	gpt-4	**25.00%**	9.84%	4.89	0.027

Table 7: **The impact of quotation marks.** False positive rate in downstream classification of counterspeech mentions, stratified by the presence or absence of quotation marks.

Sub-task	Mitigation	False positive rate (counterspeech) ↓	False positive rate Δ (counterspeech) ↓	True positive rate (true use) ↑	True positive rate Δ (true use) ↑
Hate speech	No mit.	8.89%	—	80.00%	—
	Few shot	5.02%	-43.48%	79.81%	-1.49%
	Mitigation	5.41%	-39.13%	79.81%	-1.49%
	CoT+mit.	**1.55%**	**-82.61%**	77.61%	-2.99%
Mis-information	No mit.	10.21%	—	91.98%	—
	Few shot	5.28%	-48.32%	86.09%	-6.40%
	Mitigation	7.33%	-28.19%	89.57%	-2.62%
	CoT+mit.	**4.18%**	**-59.06%**	89.57%	-2.62%

Table 8: **Mitigations.** For each prompting mitigation, false positive rate, true positive rate, and relative change in the rates calculated as $\Delta = (\text{rate} - \text{rate no mitigation})/\text{rate}$. No mitigation rates are listed for reference. In Figure 2, we illustrate the tradeoff between true positive rate (on true use) and false positive rate (on counterspeech mentions) across the tested models. Mitigation reduces false positive rate on counterspeech, with a small decrease in true positive rate for true use statements. Statistics are reported for gpt-4. For gpt-3.5-turbo, see Appendix B.

text, see Appendix, Table 13).

Furthermore, we perform an ablation study isolating the impact of use-mention examples alone (`Few shot`) and embedding the instruction to make the use-mention distinction alone (`Mitigation`).

For the best-performing model (gpt-4), we tested the prompting mitigation on counterspeech mentions and true uses. Intuitively, the mitigation should reduce the misclassification of counterspeech mentions (false positive rate on counterspeech) while not reducing correct positive classification of true uses (true positive rate on uses).

Results We find that the CoT mitigation reduces false positive rate among counterspeech mentions by 82.61% for hate speech and 59.06% for misinformation (Table 8). Among true use statements, CoT mitigation reduces true positive rates only marginally (2.99% for hate speech and 2.62% for misinformation). Ablated mitigations reduce false positive rate among counterspeech mentions independently, but `CoT + mitigation` is the best performing condition. With gpt-3.5-turbo, similar patterns were observed (see Appendix B).

In summary, encoding the use-mention distinction reduces downstream misclassification with otherwise minimal reductions in performance on true use of hate speech and misinformation.

5 Discussion

Implications for content moderation Existing datasets of harmful content are typically collected by sampling keywords that co-occur with harmful content and performing annotation. For example, during hate speech annotation, a typical question might ask: “Does the above text contain rude, hateful, aggressive, disrespectful, or unreasonable language?” (Rae et al., 2021). Previous work has documented that existing datasets gathered through such a process contain mentioned language misclassified as harmful (Van Aken et al., 2018). This implies that researchers may not typically consider the special case of mentioning statements (including counterspeech) when collecting annotations, or that annotators with varying backgrounds and positionality may not agree on how to treat mentioning statements (Santy et al., 2023).

By classifying hate speech and misinformation, NLP tools make *implicit* judgments regarding when mentioned language is permissible. However, counterspeech should be permissible by definition, as it challenges and opposes harmful narratives (Mun et al., 2023; Hangartner et al., 2021). We suggest that in downstream content classification, treatment of mentioned language should instead be *explicitly* encoded. More broadly, efforts to use LLMs for the design of conversational socio-technical systems should take into account the use-mention distinction and strive to embed values regarding how content should be treated. To that end, our prompting mitigation serves as an example of how one societal value can be encoded by leveraging a linguistic construct to specify the treatment of counterspeech.

Beyond surface features A substantial body of literature has examined shortcomings of online content classification (Garg et al., 2023; Van Aken et al., 2018). Our results suggest that many such error cases are linked to the inability to sufficiently distinguish between use and mention.

For instance, previous work has suggested that the mere fact that a text contains **swear words** (Ethayarajh et al., 2022), **identity terms** (Dixon et al., 2018) related to ethnicity (Ghosh et al., 2021), religion (Sheth et al., 2022), gender, race, and disability (Dias Oliva et al., 2021; Díaz and Hecht-Felella, 2021), or **dialect markers** (Sap et al., 2019; Halevy et al., 2021) increases toxic classifications (Zhou et al., 2021). Such lexical biases where surface features spuriously influence classification are indicative of inadequate ability to make a use-mention distinction.

Surprisingly, we find that even the latest models (i.e., gpt-4) are very sensitive to shallow lexical biases. Although non-literal language understanding beyond surface features is essential to human communication, large language models tend to struggle with interpretations of subtle utterances (Hu et al., 2023; Ocampo et al., 2023; Yu et al., 2022). Our work thus extends prior work aiming to perform more nuanced online content classification (Pamungkas et al., 2020; Goyal et al., 2022; Sap et al., 2020). Future work needs to attend to these subtle and implicit meanings to keep misinformation and hate speech classifiers from being mere surface topic classifiers.

Beyond counterspeech While we focus on online speech, the use-mention distinction might be important in other contexts like education or tutoring (where mention language occurs frequently to specify spelling or translation), law (Henderson et al., 2022), and human-AI interaction (Shaikh et al., 2023). In general, our work offers promising directions for embedding meta-linguistic reasoning to improve performance on challenging downstream tasks, such as eliciting metaphorical meanings, or dog whistle classifications, where existing methods are not reliable (Wachowiak and Gromann, 2023; Mendelsohn et al., 2023) and teaching meta-linguistic skills might offer a possible way forward.

6 Conclusion

This work highlights the theoretical and practical importance of the use-mention distinction in NLP and CSS. We also provide guidelines and directions for future research in modeling mentioned language and mitigating the impact on downstream tasks where failure to distinguish mention leads to harmful misclassifications.

Ethical implications

The central implication of our work is that downstream tasks in applications that involve occurrences of mentioned language should be handled with caution, especially when misclassification of mention could be harmful. However, content categorized as counterspeech might not be universally beneficial. For example, humans can demonstrate bias in use of counterspeech to preferentially challenge content from those with whom they disagree politically (Allen et al., 2022). In that way, counterspeech can be leveraged as a tool to harass others with opposing views. Consequently, our mitigation to prevent censorship of counterspeech might allow harmful content to proliferate due to taking a form of counterspeech. We note that incorporating specific guidelines into platforms requires further testing.

We also note that strategies involving mentioned language can equally be used to veil the speaker's true intention. Using specific terms might be socially acceptable for a person of a given identity, while unacceptable for someone else. For instance, acceptability of mentions of certain slurs is debated, especially if a non-derogatory version is readily available (Green, 2023). Addressing such complexities in the mentioned language largely remains an open question.

Lastly, despite known limitations, natural language systems are widely used for online content moderation (Welbl et al., 2021; Gehman et al., 2020). It remains unclear how well they are expected to perform and whether it is appropriate to deploy models that might produce harmful errors (Fortuna et al., 2022). Our work echoes the need for more cautious development and deployment of such systems that accounts for the conversational context, identities, and intentions. Our work also builds on existing literature that challenges the very concepts of binary classification (Pachinger et al., 2023; Davani et al., 2022) into constructs such as hate speech.

Limitations

We limit the scope of our work to testing publicly available out-of the box classifiers. Similarly, we do not investigate all the possible mitigation strategies. For example, fine-tuning with more examples could help further decrease the error rates. We also note that we do not take into account how humans would rate studied texts. Previous work studying online discourse has found that people too struggle in disentangling mentioning a fact from stating an opinion, which makes the subsequent conversation more likely to derail into uncivil behavior (Chang et al., 2020). Nonetheless, counterspeech examples examined in this work were written by social media users and expert humans who deemed them appropriate for responding to harmful narratives.

Finally, the present study is limited to specific types of mentioned language related to counterspeech. We do not solve all the problems of use-mention in the broader linguistics literature (Sandhan et al., 2023; Wilson, 2011b), including attributed language, words or phrases as themselves, proper names, and translations and transliterations. In this work, we are primarily interested in attributed language and words or phrases as themselves, as they closely relate to counterspeech which needs to mention phrases stated by others. Future work should consider other forms of mentioned language, and their impact on NLP systems more broadly. For instance, grammatical error correction might be impacted by failures in use-mention distinction (Arand, 2022). Our work is a first step toward analyzing the language of mention in different tasks, contexts, languages, cultures, and times.

Acknowledgments

Kristina Gligorić is supported by Swiss National Science Foundation (Grant P500PT-211127). Myra Cheng is supported by an NSF Graduate Research Fellowship (Grant DGE2146755) and Stanford Knight-Hennessy Scholars graduate fellowship. This work is also funded by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Michael L Anderson, Yoshi Okamoto, Darsana Josyula, and Don Perlis. 2002. The use-mention distinction and its importance to HCI. In *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialog*, pages 21–28.

Molly Andrews. 2002. Introduction: Counter-narratives and the power to oppose. *Narrative Inquiry.*

Dustin Arand. 2022. Grammarly Doesn't Understand The Use/Mention Distinction So be sure that you do.

Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider, and Amir Zeldes. 2023. Elqa: A corpus of metalinguistic questions and answers about english. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2031–2047.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876.*

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan P. Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020. Don't Let Me Be Misunderstood:Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2066–2077, New York, NY, USA. Association for Computing Machinery.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25:700–732.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Ángel Díaz and Laura Hecht-Felella. 2021. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law*.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5988–6008.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP Practices Applied to Online Hate Speech Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38, Toronto, Canada. Association for Computational Linguistics.

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

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1):3.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Caitlin Green. 2023. Beyond "Mention vs. Use": The Linguistics of Slurs.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, USA. Association for Computing Machinery.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph

Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, and Matthias Jochum. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Bradley D Menz, Nicole M Kuderer, Stephen Bacchi, Natansh D Modi, Benjamin Chin-Yee, Tiancheng Hu, Ceara Rickard, Mark Haseloff, Agnes Vitry, Ross A McKinnon, et al. 2024. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *bmj*, 384.

Meta. 2023. Hate speech: Publisher and Creator Guidelines.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Mohsen Mosleh, Rocky Cole, and David G Rand. 2024. Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS nexus*, page pgae111.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777.

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining factbank: Predicting the author's presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward Disambiguating the Definitions of Abusive, Offensive, Toxic, and Uncivil Comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France. European Language Resources Association.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Donald Perlis, Khemdut Purang, and Carl Andersen. 1998. Conversational adequacy: mistakes are the essence. *International Journal of Human-Computer Studies*, 48(5):553–575.

Perspective. 2023. Using machine learning to reduce toxicity online.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, Melissa Ferguson, and Stav Atir. 2023. Distinguishing address vs. reference mentions of personal names in text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6801–6809.

Vinodkumar Prabhakaran, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, et al. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Willard VO Quine. 1940. Use versus mention. *Mathematical Logic*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Jacquelyn Rahman. 2012. The N word: Its history and use in the African American community. *Journal of English Linguistics*, 40(2):137–171.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting Unintended Social Bias in Toxic Language Datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Paul Saka. 1998. Quotation and the use-mention distinction. *Mind*, 107(425):113–135.

Jivnesh Sandhan, Om Adideva Paranjay, Komal Digumarthi, Laxmidhar Behra, and Pawan Goyal. 2023. Evaluating neural word embeddings for Sanskrit. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 21–37, Canberra, Australia (Online mode). Association for Computational Linguistics.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. *arXiv preprint arXiv:2306.01943*.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational linguistics*, 38(2):261–299.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding or Guesswork? Large Language Models are Presumptive Grounders. *arXiv preprint arXiv:2311.09144*.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Alexandra Siegel and Vivienne Badaan. 2020. #No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online. *American Political Science Review*, 114(3):837–855.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023a. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can

Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Stuart M Shieber, and Dan Jurafsky. 2023b. string2string: A Modern Python Library for String-to-String Algorithms. *arXiv preprint arXiv:2304.14395*.

Alfred Tarski. 1931. The concept of truth in formalized languages. *Studia Philosophica*.

TikTok. 2023. Safety and Civility.

Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Lesley Wexler, Jennifer K Robbennolt, and Colleen Murphy. 2019. # MeToo, Time's up, and Theories of Justice. *U. Ill. L. Rev.*, page 45.

Shomir Wilson. 2010. Distinguishing use and mention in natural language. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 29–33.

Shomir Wilson. 2011a. *A computational theory of the use-mention distinction in natural language*. University of Maryland, College Park.

Shomir Wilson. 2011b. In Search of the Use-Mention Distinction and its Impact on Language Processing Tasks. *Int. J. Comput. Linguistics Appl.*, 2(1-2):139–154.

Shomir Wilson. 2012. The creation of a corpus of english metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 638–646.

Shomir Wilson. 2013. Toward automatic processing of english metalanguage. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 760–766.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for Counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155.

A Dataset statistics

Two annotators annotated a random sample of statements (160), uniformly distributed across hate speech and misinformation. Half of the instances were uses, and half counterspeech mentions. Each statement was annotated to indicate if the focal tokens are used or mentioned (following Def. 1). Given that we leveraged datasets with curated counterspeech statements (Fanton et al., 2021; Chung et al., 2021; He et al., 2023), as expected, within the selected sample, true use original posts contained no mentions, while counterspeech contained no uses of hate speech and misinformation. Similarly, counterspeech mentions the same focal tokens from the original statement (as opposed to mentioning harmful language not in the original use statement).

B Mitigations

In Figure 2, we illustrate true positive rate on true use and false positive rate for counterspeech statements across the tested models. Mitigation reduces false positive rate on counterspeech, with a small decrease in true positive rate for true use statements. In Table 9, for completeness, we list the results from mitigation study with gpt 3.5-turbo (ChatGPT 3.5).

C Additional statistics

Table 10 lists propagation statistics separately by task. Table 11 lists recall in hate speech classification stratified by target identity. We note that groups that have higher counterspeech false positive rates have a higher recall on true use, while groups with lower counterspeech false positive rates have lower recall too. Some discrepancies exist, as, for instance, false positive rate in mentioning counterspeech for Jewish is higher than for disabled identities, despite similar recall on true use (96% and 98.11%). These patterns are likely associated with biases in training data which implicitly encode the treatment of different groups, and the respective counterspeech.

D Prompt text

For reproducibility, in Tables 12 and 13, we list the complete prompts tested in our studies.

Sub-task	**Mitigation**	**False positive rate $\Delta\downarrow$**	**True positive rate $\Delta\uparrow$**
Hate speech	Few shot use-mention examples	-41.67%	-2.94%
	Use mention mitigation	-41.67%	-5.88%
	Few shot CoT use mention mitigation	**-78.33%**	-4.41%
Misinformation	Few shot use-mention examples	-58.02%	-7.74%
	Use mention mitigation	-46.76%	-6.50%
	Few shot CoT use mention mitigation	**-73.04%**	-27.24%

Table 9: **Mitigation statistics for gpt-3.5-turbo (ChatGPT 3.5)**.

Downstream task	**Model**	**Use-mention ¬correct**	**Use-mention correct**	χ^2	**p**
Hate speech	gpt-3.5-instruct-turbo	**53.12%**	10.34%	19.84	8.42×10^{-6}
	gpt-3.5-turbo (ChatGPT 3.5)	19.35%	7.02%	3.03	0.08
	gpt-4	12.5%	6.9%	0.8	0.37
Misinformation	gpt-3.5-instruct-turbo	**31.41%**	20.84%	11.84	5.80×10^{-4}
	gpt-3.5-turbo (ChatGPT 3.5)	**28.99%**	15.5%	20.54	5.84×10^{-6}
	gpt-4	**16.03%**	4.2%	30.14	4.03×10^{-8}

Table 10: **Error propagation by task.** False positive rate in downstream classification of counterspeech mentions, stratified by use-mention classification correctness. Statistics are reported separately by tasks.

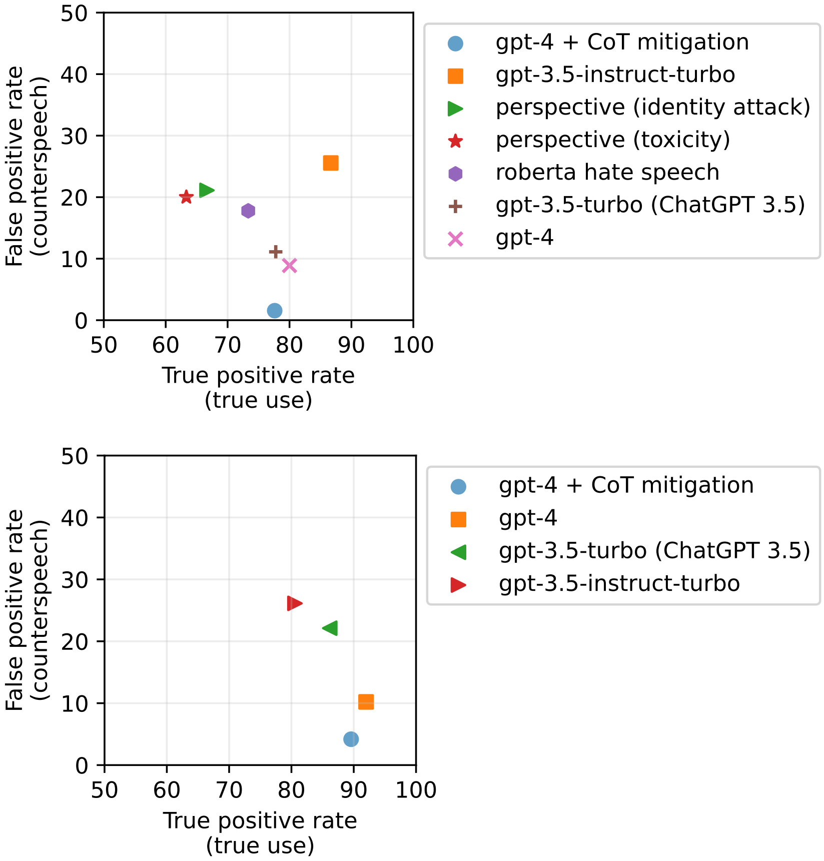

Figure 2: **Mitigation illustration.** Metrics are visualized separately for hate speech detection (top) and misinformation detection (bottom). It is desirable to have a low counterspeech false positive rate (y-axis) and a high use true positive rate (x-axis). For both hate speech and misinformation, CoT mitigation reduces counterspeech false positive rate with a marginal decrease in true positive rate on use.

Target identity	False positive rate	Recall
Jewish	14.15%	98.11%
People of color	9.09%	72.73%
Muslims	6.80%	85.51%
LGBT+	6.77%	80.45%
Disabled	4.00%	96.00%
Women	2.41%	85.71%
Other	0.89%	56.25%
Migrants	0.39%	53.64%

Table 11: **Identity: False positive rate and recall.** False positive rate and recall in hate speech classification of counterspeech, stratified by target identity (gpt-3.5-turbo).

Prompt	Prompt text
Use-mention (hate speech)	*The following text contains language that might be considered hateful. Classify the text into one of the following two categories:* *A: use* *B: mention* *Use means that the hateful language is actively used, and not referred to. Examples of used:* *1. Using hateful language to express, incite, offend, or promote hate.* *2. Using hateful language to imply hate towards an identity group.* *Mention means that hateful language is stated to illustrate properties of the hateful language itself. Examples of mentioned:* *1. Quoting, referring to, or paraphrasing hateful language to explain attitude towards it.* *2. Addressing hateful language when providing a counternarrative.* *Output only A or B.* *Text: <text>* *Category:*
Use-mention (misinformation)	*The following text contains language that might be considered misinformation. Classify the text into one of the following two categories:* *A: use* *B: mention* *Use means that misinformation is actively used, and not referred to. Examples of used:* *1. Using misinformation to deliberately deceive.* *2. Using misinformative language to imply the truthfulness of misinformation.* *Mention means that misinformation is stated to illustrate properties of the misinformative language itself. Examples of mentioned:* *1. Quoting, referring to, or paraphrasing misinformation to express a lack of support for the misinformative language.* *2. Addressing misinformation when providing a counternarrative.* *Output only A or B.* *Text: <text>* *Category:*
Downstream (hate speech)	*Classify the following text into one of the following two categories: [hateful, not hateful]. Hateful content expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.* *Text: <text>* *Category:*
Downstream (misinformation)	*Classify the following text into one of the following two categories: [misinformation, not misinformation]. Misinformation means that the text is either inaccurate or contains an element of truth but ignores critical facts that would give a different impression.* *Text: <text>* *Category:*

Table 12: **Prompt text (use-mention and downstream tasks).**

Prompt	Prompt text
Few shot	*Classify the following text into one of the following two categories: [hateful, not hateful]. Hateful content expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.* *Text: gay people are a burden to our society* *Category: hateful* *Text: I always feel unsafe when surrounded by Arabs* *Category: hateful* *Text: decent people are allowed here, so queers keep out* *Category: hateful* *Text: It is incredibly ignorant to claim that gay people are a burden to our society. This kind of discrimination has no place in our world.* *Category: not hateful* *Text: Feeling unsafe just because someone is Arab is a form of prejudice and discrimination. I completely disapprove of that kind of mindset.* *Category: not hateful* *Text: <>* *Category:*
Mitigation	*Classify the following text into one of the following two categories: [hateful, not hateful]. Hateful content expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.* *Hateful is when hateful language is actively used, and not referred to. When language or arguments related to hate are just mentioned, the text is not hateful when debunking.* *For example, text quoting, referring to, or paraphrasing hate speech to provide a counternarrative is not hateful.* *Text: <text>* *Category:*
CoT + mitigation	*Classify the following text into one of the following two categories: [hateful, not hateful]. Hateful content expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.* *Use means that the hateful language is actively used, and not referred to. Mention means that hateful language is stated to illustrate properties of the hateful language itself.* *Hateful is when hateful language is actively used, and not mentioned. When language or arguments related to hate are just mentioned, the text is not hateful when debunking.* *For example, text quoting, referring to, or paraphrasing hate speech to provide a counternarrative is not hateful.* *Text: gay people are a burden to our society* *Answer: Let's think step by step. Hateful language is actively used in order to express hate. So the answer is: hateful* *Text: I always feel unsafe when surrounded by Arabs* *Answer: Let's think step by step. Hateful language is actively used in order to express hate. So the answer is: hateful* *Text: decent people are allowed here, so queers keep out* *Answer: Let's think step by step. Hateful language is actively used in order to express hate. So the answer is: hateful* *Text: It is incredibly ignorant to claim that gay people are a burden to our society. This kind of discrimination has no place in our world.* *Answer: Let's think step by step. Hateful language is mentioned. Mentioning is to provide a counternarrative. So the answer is: not hateful* *Text: Feeling unsafe just because someone is Arab is a form of prejudice and discrimination. I completely disapprove of that kind of mindset.* *Answer: Let's think step by step. Hateful language is mentioned. Mentioning is to provide a counternarrative. So the answer is: not hateful* *Text: <text>* *Answer: Let's think step by step.*

Table 13: **Prompt text (hate speech mitigations).**