Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation

Bing He, Mustaque Ahamad, Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu,mustaq@cc.gatech.edu,srijan@gatech.edu

ABSTRACT

The spread of online misinformation threatens public health, democracy, and the broader society. While professional fact-checkers form the first line of defense by fact-checking popular false claims, they do not engage directly in conversations with misinformation spreaders. On the other hand, non-expert ordinary users act as eyes-on-the-ground who proactively counter misinformation – recent research has shown that 96% counter-misinformation responses are made by ordinary users. However, research also found that 2/3 times, these responses are rude and lack evidence. This work seeks to create a counter-misinformation response generation model to empower users to effectively correct misinformation. This objective is challenging due to the absence of datasets containing ground-truth of ideal counter-misinformation responses, and the lack of models that can generate responses backed by communication theories. In this work, we create two novel datasets of misinformation and counter-misinformation response pairs from in-the-wild social media and crowdsourcing from college-educated students. We annotate the collected data to distinguish poor from ideal responses that are factual, polite, and refute misinformation. We propose MisinfoCorrect, a reinforcement learning-based framework that learns to generate counter-misinformation responses for an input misinformation post. The model rewards the generator to increase the politeness, factuality, and refutation attitude while retaining text fluency and relevancy. Quantitative and qualitative evaluation shows that our model outperforms several baselines by generating high-quality counter-responses. This work illustrates the promise of generative text models for social good – here, to help create a safe and reliable information ecosystem. The code and data is accessible on https://github.com/claws-lab/MisinfoCorrect.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Reinforcement learning**.

KEYWORDS

misinformation, reinforcement learning, text generation

WWW '23, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583388

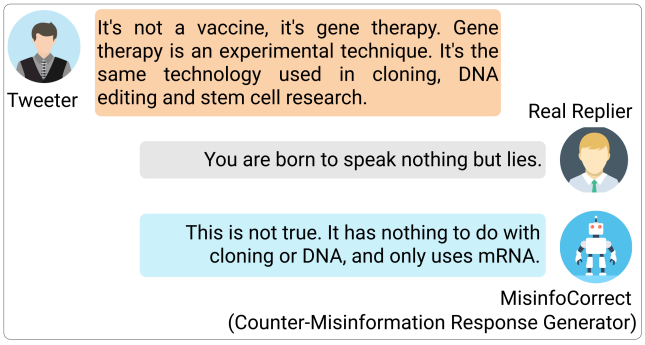

Figure 1: An overview of counter-misinformation response generation task.

ACM Reference Format:
Bing He, Mustaque Ahamad, 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 (WWW '23), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583388

1 INTRODUCTION

Online misinformation reduces trust in vaccines and health policies [7, 46, 68], leads to violence and harassment [6, 84], questions democratic processes and elections [80–82], increases polarization [86], and harms well-being [97]. Most people receive information and news from social media [108], which is often "ground-zero" for health misinformation and where misinformation spreads faster and farther than truth [46, 101]. COVID-19 vaccine misinformation, including false claims that the vaccine causes infertility, contains microchips and even changes DNA and genes has fueled vaccine hesitancy, reduced vaccine uptake, and prolonged the pandemic. Besides, misinformation also causes harms to people directly. For example, misinformation that Bill Gates creates vaccines to depopulate people led to distrust and verbal attacks [25]. Thus, it is critical to curb the spread of online misinformation [13, 28, 47, 50, 58, 112, 120]. In this work, we use a broad definition of misinformation which includes falsehoods, inaccuracies, rumors, decontextualized truths, or misleading leaps of logic [45, 115].

Professional fact-checkers and journalists provide objective fact-checks for viral claims and release their determination on their website, which are incredibly useful to create detection models. However, fact-checkers do not actively engage with misinformation

spreaders on social media platforms [58]. On the other hand, non-expert social media users, i.e., ordinary users or crowd, act as eyes-on-the-ground who proactively question and counter misinformation, including emerging misinformation [9, 58, 60, 84, 94, 98, 119]. They complement fact checkers who can only verify a handful of stories after they have gone viral [3, 42]. Recent evidence shows that 96% of counter-misinformation responses are made by ordinary users, while professionals account for the rest [58].

Recent research on social correction [54], i.e., countering of misinformation claims by other social media users, has proven to be as effective as professional correction [77], curbs misinformation spread [17, 24, 113], and works across topics [8, 9, 11, 103–105], platforms and demographics [102, 105–107]. Corrections work [14, 69, 109–111] without causing an increase in misperception (i.e., the backfire effect has not been replicated) [31, 90, 114]. While corrections are not expected to convince everyone, they are most effective in reducing misinformation consumers' misperceptions [9, 10, 17, 76, 113]. Thus, empowering users to effectively correct misinformation promises a scalable solution towards information integrity. This solution is independent of, but complements, the efforts of social media platforms to detect misinformation via the crowd, e.g., Twitter Birdwatch [70].

Alarmingly, linguistic analyses of in-the-wild crowd-generated counter-responses revealed that 2 out of 3 counter-misinformation posts are rude and do not use fact-checking evidence to support their counter-response [58]. Uncivil counter-responses can lead to reduced trust in the correcting user [23, 93] and result in arguments [15, 44, 57]. This implies an urgent need to empower crowds so that they counter misinformation more effectively.

Thus, in this work, we seek to facilitate healthy misinformation correction by the crowd, which includes being objective, evidenced, and polite – properties that have been shown to be effective [77, 91]. To do so, we propose to create a counter-misinformation response generator, which generates desirable counter-response for a misinformation post (as illustrated in Figure 1). Our study is focused on countering misinformation on Twitter, given its prominence in the spread of online misinformation.

Challenges. Generating effective counter-misinformation responses poses several challenges. First, there is no existing dataset containing pairs of annotated misinformation posts and counter responses. Second, there is no counter-misinformation response generator model. The closest research works in fact-checking generator [99] are non-conversational and related research in counter-hate speech/counter-argument generator [5, 16, 92, 121] do not apply directly since they are not evidence-based or not specific to misinformation. Third, counter-misinformation responses are effective if they have the following desirable properties: objective and evidenced [77, 91], makes rational arguments [65], refutes fallacy in reasoning [87], and polite [55, 85]. Off-the-shelf text generator models do not directly generate counter-responses with this desiderata. Four, bot-generated or template-based responses are not effective since they are non-personalized and non-contextualized with respect to the false claims made in the misinformation post. Thus, the counter-response needs to be relevant to the misinformation post.

Present work. We propose to create two novel datasets containing misinformation and counter-responses (solution to challenge 1) – one collected from in-the-wild social media responses from Twitter and another created by crowd-sourcing from college students. We focus on four popular COVID-19 vaccine misinformation topics on Twitter (e.g., Bill Gates created vaccines to depopulate people [22, 79], and vaccines can cause infertility [1, 39], contain microchip [83], alter DNA [52, 62]). To create the in-the-wild dataset, for each misinformation topic, we collect all the replies to misinformation tweets identified in prior research [34]. We annotate associated replies to identify the responses that counter the tweet along with their textual attributes of refuting, politeness, and factuality. Finally, we have 754 misinformation tweet and countering response pairs. For the crowd-sourced response generation, we recruit and train 17 college students to write counter-misinformation replies when given misinformation posts. In total, we collect 591 crowdsourced replies.

Next, we propose a reinforcement learning-based framework, called MisinfoCorrect, that learns to generate counter-misinformation responses that are polite, evidenced, and refute misinformation (solutions to challenges 2 and 3). Specifically, this agent utilizes a policy network on a transformer-based language model adapted from GPT-2 [72]. During training, we reward the generation that increases the politeness and refutation attitude. Additionally, we ensure text fluency and relevancy to the misinformation post by adding fluency and relevance rewards in the reinforcement learning framework (solution to challenge 4).

MisinfoCorrect is evaluated against five representative baselines on the task of counter-misinformation response generation. Quantitative and qualitative experiments show that it can outperform the baselines by generating high-quality counter-responses.

To summarize, our contributions are as follows:

- We create two large novel and annotated datasets containing misinformation and counter-response pairs from social media (in-the-wild) and generated via crowd-sourcing (in-lab). Together, both datasets contain 1,345 counter-misinformation responses.
- We propose a reinforcement learning based counter-response generation framework, where the counter-response is especially rewarded for being polite, evidenced, and refuting misinformation.
- Results on actual COVID-19 vaccine misinformation conversations show that the proposed model outperforms existing representative baselines.

The code and data is accessible on https://github.com/claws-lab/MisinfoCorrect.

2 RELATED WORKS

2.1 Social Correction of Misinformation by Non-Expert Ordinary Users

Recent studies have shown remarkable effectiveness of social correction by non-expert users by conducting experiments via interviews [12, 43, 94], surveys [43, 96], and in-lab experiments [94]. This correction has been shown to be as effective as professional correction [77], curbs misinformation spread [17, 24, 113], and works across topics [8, 9, 11, 103–105], platforms and demographics [102, 105–107]. Notably, users' polite and evidenced responses that refute misinformation are shown to effectively counter misinformation and reduce the belief in misinformation [14, 55, 65, 65, 77, 85, 87, 91]. Users correct others, typically friends [56], owing to a sense of social duty [24, 30, 61, 66, 96], anger, or guilt [88]. These works provide

considerable evidence that correction by ordinary users is effective when countering misinformation and in mitigating the spread of misinformation. On the other hand, considering the limited capability of professional fact-checkers, the large number of ordinary users and their efforts in social correction show great potential for a scalable solution to countering misinformation.

2.2 Analysis of Crowd-Generated Misinformation Flagging and Countering

Crowd-generated counter-misinformation complements fact-checking and correction by professionals – the latter has already been studied extensively [32, 33, 67, 77, 110, 117]. Emerging research has analyzed the role that non-experts play in flagging and countering misinformation. Twitter's Birdwatch [70] is a recently-launched platform that allows users to report and flag misinformation. Studies have analyzed the data from Twitter Birdwatch [4, 21, 60, 70], which have shown how users actively engage to identify tweets that they believe are misleading and provide contextual notes to debunk them. Users have different levels of debunking capability. However, Birdwatch only allows users to flag misinformation and does not allow user-to-user communication and countering of misinformation on Twitter. Thus, user flagging behavior within the Birdwatch ecosystem is not representative of user behavior on the broader Twitter platform or on other social media platforms. Recent works by Micallef et al. [58, 59] have analyzed how users counter misinformation in-the-wild on Twitter, Facebook, and Reddit. They showed that 96% of all counter misinformation posts on Twitter are made by "ordinary citizens" [58] and counter-misinformation behavior happens on multiple platforms [59]. Existing works, however, have not studied how to empower the crowd to counter and correct misinformation by generating effective responses.

2.3 Fact-Check Generation Methods

The goal of fact-check generation methods [99, 100] is to respond to misinformation with a fact-checking URL. However, we consider a broader task of counter-response generation where the response text has to be generated. Existing works [99, 100] consider any post with a fact-checking URL to two websites (Snopes and Politifact) as a fact-checking response, which is an inaccurate assumption – a fact-checking URL can be present to ridicule or oppose the fact-check [59] and can be taken out of context [59]. Importantly, only 1 out of 3 users use URL evidence when correcting misinformation [58] and YouTube is the most frequently used URL, instead of fact-checking URLs [59]; consequently, studies relying only on fact-checking URLs are limited in their scope and do not learn from the majority of user-generated corrective posts. Our work overcomes these shortcomings by creating two novel datasets (Section 4), one using social media, including both URL and non-URL responses, and another using crowdsourced data collection. We further perform several manual annotation steps while creating the data to ensure only exact counter-responses are present in the data.

2.4 Counter-Hate and Counter-Argument Text Generation

Counter-hate [16, 36, 92, 121] and counter-argument [5, 37, 40] text generation tasks are also related to our problem setting, where the generated text is aimed to refute the original post spreading hate and any generic argument, respectively. Some proposed models fine-tune large scale unsupervised language models on the hate-speech or argument text for text generation [75, 92]. Other models first generate a set of candidate counter-hate/counter-argument replies, and then select one based on the relevance to the original post in a generate-then-retrieve or identify-substitute manner [37, 40, 121] . Meanwhile, some related counter-hate/counter-argument datasets have also been released [40, 71, 74]. However, it should be noted that compared to counter-misinformation response generation, the task of counter-hate generation does not necessitate responses to be evidence-based. Similarly, the counter-argument generation is a generic task (e.g., arguing whether immigration is good) and is not specific to misinformation. Additionally, large annotated and curated datasets exist for counter-hate and counter-argument [71, 74], which is not the case for counter-misinformation generation. To fill these gaps, we both curate two novel datasets and propose a counter-misinformation generator which can refute misinformation while being polite and providing evidence.

3 PROBLEM DEFINITION

Given a misinformation post m, we aim to build a text generator g such that it can output counter-response $\hat{c} = g(m)$, which has certain desirable properties $\mathcal{P}$.

The **desirable properties** of $\hat{c}$ are motivated by research works from social scientists, journalists and psychologists regarding misinformation correction, which shows that counter responses are effective if they have the following desirable properties: politeness [55, 85], objective and evidenced [77, 91], make rational arguments [65], convey the competence of the commenter [65], and refute fallacy in reasoning [14, 87]. More elaborately, the desirable properties include:

- *Refuting:* the response explicitly refutes the the misinformation to correct the misinformation spreader. The expressed refutation via explicitly and objectively refuting misinformation in counter response can reduce misinformation's impact [91].
- *Evidence:* the response contains supporting sentences to back up the refutation. Evidenced-based responses can more effectively debunk the misleading claims, and likely reduce the belief of misinformation poster [14]. More importantly, people are more willing to agree with a countering response when it is evidence-based [14].
- *Politeness:* the response is polite to avoid possible backfire. When countering misinformation, uncivil responses can aggravate the misinformation poster, while it is more likely that the misinformation spreader favorably considers the true information when responses are polite [55, 85].

Beyond these specific requirement in misinformation correction domain, other textual properties are also required in generated text:

- *Fluency:* the generated text should be fluent in expression such that it is natural for people to read and understand.
- *Relevance:* the response should be relevant to the misinformation post and ensure coherent expression.

4 COUNTER-RESPONSE DATASETS: IN-THE-WILD AND CROWDSOURCED

We create two novel counter-response datasets, first containing in-the-wild social media counter-responses and second containing crowdsourced in-lab counter-responses.

4.1 Misinformation Topics

We focus on COVID-19 vaccine misinformation due to its impact across the world. We mainly choose four popular misinformation topics to which a large number of users have been exposed and impacted [1, 22, 39, 52, 58, 62, 79, 83]. These misinformation topics gained popularity from December 2020 when the COVID-19 vaccines were approved by the FDA [79], in essence,

- Bill Gates conspiracy theories [22, 79]: This includes conspiracies claiming that Bill Gates created the COVID-19 vaccine to depopulate people or he holds the patents for COVID-19 vaccine to profit from the vaccine sales.
- COVID-19 vaccines contain microchips to track people [83].
- COVID-19 vaccines cause infertility and prevent pregnancy in women [1, 39].
- COVID-19 vaccines alter DNA or the vaccine is gene therapy [52].

4.2 In-the-wild Social Media Counter-Response Dataset

4.2.1 ***Misinformation Tweet and Response Collection.*** Our dataset builds on 14, 123, 473 COVID-19 vaccine-related tweets crawled by Hayawi et al. [34] from Dec 1, 2020 to July 31, 2021. Since we are more focused on responses rather than tweets themselves, we only keep tweets having at least one response, resulting in 1, 609, 069 tweets.

To identify misinformation tweets, we first create a COVID-19 vaccine misinformation tweet classifier using BERT [20] based on tweet annotations provided by Hayawi et al. [34]. This classifier has a performance in precision, recall and F1 scores of 0.972, 0.979 and 0.975, respectively. Then, we use this classifier to classify all remaining non-annotated tweets. Finally, we have 141,766 classified misinformation tweets. We crawl all their direct replies, resulting in 793, 828 replies.

Next, we filter tweets to retain those within the scope of our misinformation topics (Section 4.1), with at least one of the following (non case sensitive) keywords in the tweet textual string: "bill gates", "fertility", "pregnancy", "pregnant", "gene", "dna", "gene therapy", and "microchip", resulting in 1, 655 tweets with 26, 190 responses.

To create a high-quality dataset, we manually annotate all the classified 1, 655 misinformation tweets by the textual content to remove false positives and only focus on original tweets (no retweets) and English-language content, as is common practice [38, 58].

Finally, this dataset contains 798 misinformation tweets and associated 11, 970 responses.

4.2.2 ***Annotating Counter-Misinformation Replies and Training the Classifier.*** Naturally, not all responses to misinformation tweets counter it. Therefore, to develop a counter-response dataset, we create the following procedure.

Training a counter-response classifier: Since annotating all 11, 970 responses manually is labor-intensive, we leverage existing work by Jiang et al. [41] to create a belief versus disbelief classifier in social media responses. Specifically, following their pipeline, we create the classifier using RoBERTa [51] and train it on their annotated responses. Since the topics of the original data and trained classifier in Jiang et al. [41] are different from ours, we annotated additional responses. Specifically, two students annotated 500 randomly-selected responses from the unlabeled 11, 970 responses, resulting in an inter-rater agreement score of 0.7033 measured by percent agreement. This gave 244 responses expressing belief and 118 expressing disbelief, while the remaining were neither expressing belief or disbelief. We used these annotated responses to fine-tune the disbelief classifier to our data and topic. Conducting five-fold cross validation, the classification performances of the classifier per precision, recall and F-1 scores were 0.695, 0.687 and 0.691, respectively. Finally, we use the fine-tuned classifier to identify all potential disbelief replies among all the 11, 970 responses. This resulted in 2, 852 responses classified as disbelief or counter-response. Then, we manually verify all the classified responses through the textual content to remove all false positives. Finally, 754 true counter-responses are identified, which we use in our work.

4.2.3 ***Annotating Linguistic Properties of Counter-Responses.*** Two students annotated 50 counter-responses as per the three desired properties [14, 55, 85]:

- Refuting: is the response explicitly rejecting the false claim or the misinformation spreader?
- Evidence: does the response contain evidence or supporting words or sentences to back up the counter-response?
- Politeness: Is the reply rude, neutral, or polite like having a soft and friendly tone in the expression?

The measured inter-rater agreement score by percent agreement is 78%. Disagreements were discussed and a final label was given. Next, each annotator annotated the remaining counter-responses to assign final labels to them.

Finally, this results in 754 annotated (misinformation tweet, counter-response) pairs from 238 misinformation tweets. The distribution of the linguistic properties of counter-responses is shown in Table 1.

	Politeness		Evidence?		Refutes?
Polite	51	Yes	181	Yes	588
Neutral	415	No	573	No	166
Rude	288				

Table 1: Statistics of 754 social media counter-responses.

As per the statistics, in-the-wild counter-responses are very low quality – 38.19% responses are rude, 75.99% do not have evidence, and 22.02% do not explicitly refute the misinformation. This indicates they may not be effective. This further reinforces the critical and timely need for our research to develop an effective counter-response generator.

4.3 Crowdsourced In-lab Counter-Misinformation Responses

The above statistics show that most in-the-wild responses are rude and lack evidence. As a result, it will be challenging to train an effective counter-response text generator model using this data alone. Thus, we create an alternate dataset via crowdsourcing. Motivated by similar text generation for healthy and social good online communication [71, 78, 92], we recruit users familiar with Twitter to generate counter-misinformation responses that have the desired properties mentioned earlier in Section 3.

Ethics: This protocol was approved by Georgia Tech's IRB.

Procedure: We use the following three-step process:

First, we recruited 20 college undergraduate and graduate students majoring in engineering domains in March 2022. During the screening, subjects provided background information including: (1) Highest education level: high-school, bachelors, masters, or doctorate; (2) Fluency in English: basic, intermediate, advanced (fluent or native speaker); (3) Familiarity with the concept of online misinformation on Twitter: not familiar, somewhat familiar, highly familiar; and (4) Witnessed countering misinformation online: yes or no.

Out of these, 17 participants met the criteria of having least high-school education, being fluent in English, highly familiar with online misinformation, and having seen online debunking.

Second, each subject is provided written guidance about writing an effective counter-misinformation response governed by existing literature [14, 14, 55, 85]. Representative counter-misinformation examples are shown that are manually selected by the authors from the in-the-wild dataset (Section 4.2.3). Each subject is given up to 50 randomly-selected COVID-19 vaccine misinformation tweets (from the in-the-wild social media dataset) identified in Section 4.2.1. These tweets span all four misinformation topics (Section 4.1) to ensure diverse responses from different subjects.

After filtering out 90 written responses that do not satisfy any desirable properties (Section 3), we finally created a high-quality counter-misinformation response dataset containing 591 crowd-generated responses. A representative example is shown below:

> *Misinformation Post:* It's not a vaccine, it's gene therapy. Gene therapy is an experimental technique. It's the same technology used in cloning, DNA editing, and stem cell research.
> *In-the-wild Counter-response:* You are born to speak nothing but lies.
> *Crowdsourced Counter-response:* Sorry to see you think in this way. It is not correct. The vaccine is not gene therapy. Instead, it uses mRNA to generate spike protein to protect people. Please do not say the misinformation again.

5 MISINFOCORRECT: A COUNTER-RESPONSE GENERATION MODEL

Here we describe our proposed counter-response generation model that leverages the two datasets to generate counter-responses for a given misinformation post. The generated counter-responses should have the desirable properties described in Section 3.

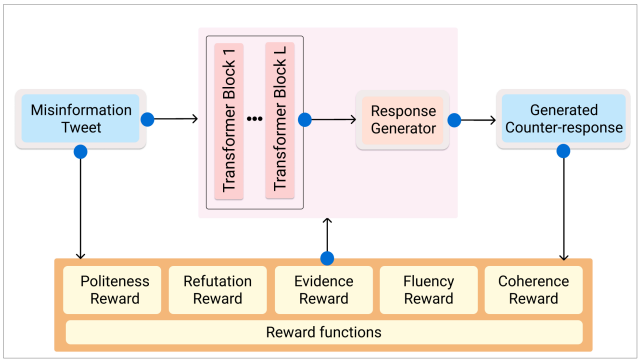

Figure 2: The overview of the MisinfoCorrect framework.

5.1 A Reinforcement Learning Framework

We choose a reinforcement learning-based approach due to its success in a variety of controllable text generation tasks [49, 78]. Moreover, we utilize reinforcement learning (RL) on top of a GPT-2 transformer-based text generation model since it is capable of generating quality example with limited number of examples derived from its strong generation power and is widely-used in text generation task [78]. By this design, we can bias the text generation process such that the generated counter-response is of high quality. Figure 2 presents the overview of MisinfoCorrect. Below we describe the components of the RL agent:

5.1.1 ***State:*** The misinformation post provides the conversational text. The RL agent takes the misinformation post as the input to enhance the quality of counter-response text so that the response is relevant to the misinformation claims. Formally, the state $s \in \mathcal{S}$ is the same as the content of the misinformation post m, i.e., $s = m$. Our policy uses a string containing s for representation, which is also widely used in BERT-like models [20].

5.1.2 ***Action:*** Given state s, the agent generates a candidate counter response $\hat{c}$. This generation action is represented as a lying in the whole action space $\mathcal{A}$, $a \in \mathcal{A}$, which is composed of all arbitrary-length sentences. We represent g as the text generator, and the action is $a = g(s)$.

5.1.3 ***Policy:*** The policy is based on the transformer language model with the task of masked multi-head self-attention layers on GPT-2 [72, 95]. The input is an encoded representation of the state s and output is the action a. The generation task is framed as a language modeling problem where the goal is to generate $\hat{c}$ that maximizes the conditional probability $p(\hat{c}|m)$. When using transformer component of GPT-2, we first encode our input string "m". Then, after transforming the encoded representation as a vocabulary-sized vector using a softmax layer, we have a probability distribution over the entire vocabulary tokens. Next, top-p sampling method is used with the probability distribution to sequentially output a sequence of tokens to form a sentence. When the sampling process selects a special end-of-sequence token, the generation process stops. This generates the candidate counter-response $\hat{c}$.

5.1.4 ***Reward:*** Research has shown that counter-misinformation responses are effective if they are polite, provide evidence, and explicitly refute the misinformation (Section 3). We design multiple

novel reward functions to encourage the generated response to have these properties along with ensuring that the generated text is fluent, coherent, and relevant to the misinformation post. We describe the rewards below.

• **Politeness Reward**: Polite counter-responses are preferred (Section 3). We quantify the preference toward politeness as a politeness reward $r_{politeness}$ and create a politeness classifier $f_{politeness}$ using BERT [20] to measure politeness of text leveraging existing work [19]. The classifier fine-tuned and tested in our data in Section 4 has a classification performance measured via precision, recall and F1 score of 0.8864, 0.9512, 0.8001. The politeness reward is formally computed as $r_{politeness} = f_{politeness}(\hat{c})$.

• **Refutation Reward**: Counter-responses that explicitly refute the misinformation are more effective (Section 3). Thus, we define the refutation reward $r_{refutation}$ to reward the actions that increase refutation of $\hat{c}$ and penalize actions that decrease the refutation of $\hat{c}$. Following similar disbelief and polarity classification research works [2, 41], we build the refutation classifier $f_{refutation}$ using BERT [20] which measures whether the text expresses refutation. However, distinct from Jiang et al. [41], who only use the response text for classification, we use both the tweet and generated response as input. The reason is that the refutation relationship would be better predicted by capturing the relative stance between the tweet and its response. We quantify the refutation reward as $r_{refutation} = f_{refutation}(m, \hat{c})$. In our experiments, the refutation classifier is first trained on the annotated data by Jiang et al. [41]. Then, we fine-tuned and tested it on our data (Section 4), which finally achieves reasonable performance in precision, recall and F1 score with values of 0.7917, 0.8085, 0.7999, respectively.

• **Evidence Reward**: Responses containing evidence are more effective in countering misinformation [14]. Thus, we seek to generate response that provides textual evidence. We do not seek to provide a fact-checking URL as evidence, since readers are unlikely to click and read an external article from social media platforms [26, 27]. To effectively quantify the presence of evidence in responses, we consider the counter-response content where the response counters the misinformation post with supporting and relevant sentences.

We create an evidence classifier $f_{evidence}$ to predict whether the response provides evidence that counters the misinformation post. The classifier is trained by combining two sets of evidence-providing responses – first is the in-the-wild social media counter-responses that contain evidence (Section 4.2.3), and second is the subset of crowdsourced responses (Section 4.3) with evidenced responses. Finally, we create a balanced dataset of 573 evidenced-responses and 573 non-evidenced-responses to train the classifier.

We use BERT [20] as the classifier which takes both the post and response as inputs in a pair-wise setting [73] to measure the post-response pairwise relationship. After five-fold cross validation, the performance score of precision, recall and F1 score is 0.8864, 0.9512, 0.9176. The output of the classifier is the evidence reward, $r_{evidence}$, computed as $r_{evidence} = f_{evidence}(m, \hat{c})$.

• **Fluency Reward**: The agent needs to ensure that the response is fluent and grammatically correct. Thus, we want to reward actions that generate fluent outputs and penalize ones that result in non-fluent responses. To achieve this goal, following the previous work [53], we design the fluency reward $r_{fluency}$ which is the inverse of perplexity of the generated countering reply $\hat{c}$ as $r_{fluency} = p_{GPT-2}(\hat{c})^{\frac{1}{M}}$, where p_{GPT-2} is the GPT-2 language model for English and M is the number of words in $\hat{c}$.

• **Coherence Reward**: Given a misinformation post, the generated response should be relevant to the post. We design a coherence reward $r_{coherence}$ which is computed via semantic similarity between m and $\hat{c}$ as $r_{coherence} = sim(m, \hat{c})$, where sim measures the semantic similarity between two posts. In practice, we utilize the embedding from BERT model of the two text pieces [20] and compute their cosine similarity.

Total reward: Finally, the total reward is as

$$r = \alpha * r_{politeness} + \beta * r_{refutation} + \gamma * r_{evidence} + \theta * r_{fluency} + \lambda * r_{coherence} \quad (1)$$

where $\alpha, \beta, \theta, \gamma, \lambda$ are weights indicating the importance of rewards.

5.2 Optimization and Training

Warm-up start: We first use the pre-trained weights of DialoGPT [118] to initialize the weights in the transformer-based GPT-2 language model. Next, motivated by the warm-up approaches in reinforcement learning for dialogue generation by Li et al. [49], we use the warm-start strategy on the paired data of misinformation posts and countering replies.

Reward Increment Training for Reinforcement Learning: To train the agent in the reinforcement learning framework, we take advantage of the existing reward increment training approach where the non-negative factor, offset reinforcement and characteristic eligibility are considered in the standard reinforcement learning setting [89]. In our setting for simplicity, we consider the reward r from the generated post and the probability of generating this post given the misinformation post, $p(\hat{c}|m)$. Finally, the loss function $\mathcal{L}$ is computed as $\mathcal{L}(\theta) = -r * logp(\hat{c}|m)$, where θ is the set of model parameters. We use log to facilitate computation. Meanwhile, we utilize the negative of the reward to deploy the conventional gradient descent approach in experiment. Adam is used as the optimizer for model training [29].

6 EXPERIMENTAL EVALUATION

We examine the performance of the proposed counter-misinformation response generation model. In particular, we focus on answering the following research questions:

• **RQ1**: Can the proposed model generate counter-misinformation responses of high quality with the desirable properties (Section 3)?

• **RQ2**: What is the impact of using in-the-wild data versus crowdsourced data on the generated text output?

• **RQ3**: What is the contribution of each component of the proposed method?

• **RQ4**: Is the text generated good as evaluated by humans?

6.1 Baselines

We compare our model with representative dialog generation baselines and the work in fact-checking text generation:

• Fact-checking Text Generation (**FC-GEN**) [99]: The fact-checking text generation model takes in the tweets and replies for generation using gated recurrent unit.

Method	Polite. ↑	Refut. ↑	Evid. ↑	Perpl. ↓	Rele. ↑
DialoGPT	0.874	0.831	0.693	10.010	0.930
Seq2seq	0.794	0.794	0.621	13.403	0.948
BART	0.824	0.827	0.623	11.909	0.870
Partner	0.892	0.898	0.702	9.781	0.871
FC-GEN	0.815	0.714	0.594	14.971	0.810
MisinfoCorrect	**0.915**	**0.931**	**0.723**	**8.010**	**0.960**

Table 2: Performance comparison of counter-response generators when trained on social media and crowdsourced responses.

• **DialoGPT** [118]: A dialogue generation model built on GPT-2 framework and pre-trained on Reddit conversations.
• Deep latent sequence model (**Seq2Seq**) [122]: A encoder-decoder model for general dialog text generation.
• **BART** [48]: An large pre-trained language model framework for sequence-to-sequence text generation.
• **Partner** [78]: A reinforcement-learning-based text rewriting method to output text.

6.2 Evaluation Metrics

To quantitatively evaluate the performance of the model, we use several metrics to measure both the effectiveness of the counter response and the text quality as follows:
• **Politeness**: We use the politeness classifier $f_{politeness}$ to test the level of politeness expressed in generated responses (Section 5.1.4).
• **Refutation**: We use the trained refutation classifier $f_{refutation}$ to measure refutation score, as defined in Section 5.1.4.
• **Evidence**: We use trained evidence classifier $f_{evidence}$ (Section 5.1.4) to measure how much evidence the reply provides.
• **Perplexity**: Following previous research [18, 53], we use pre-trained GPT-2 language model to quantify perplexity to evaluate the expressed text fluency.
• **Relevance**: Following previous research [116], we compute the semantic similarity using BERT [20] to capture the coherence between posts and generated responses.

6.3 RQ1: Evaluation of the Proposed Model

We train all the models with counter-responses from both *social media dataset* (Section 4.2) and *crowdsourced counter-responses* (Section 4.3). Specifically, we create a "clean" social media dataset by only selecting counter-responses with at least one dimension among politeness, refutation, and evidence labeled as positive. This is because training with low-quality counter-responses will lead to poor generation results. In addition, we use all crowdsourced counter-responses as they are all manually-verified to be polite, refuting, and evidenced.

The results comparing the generation models are shown in Table 2. As can be seen, our proposed model generates the best counter-responses. When compared with baselines, our model has the highest politeness, refutation and evidence scores while still maintaining significantly lower perplexity and comparable relevance scores to ensure text of high quality. Table 3 illustrates responses generated by the proposed model and other baselines. As we can see, compared to other methods, MisinfoCorrect can generate text of desirable properties.

Misinformation Post: It's not a vaccine, it's gene therapy. Gene therapy is an experimental technique. It's the same technology used in cloning, DNA editing, and stem cell research.
MisinfoCorrect (proposed): This is not true. And, the vaccine is not gene therapy. It has nothing to do with cloning or DNA, and only uses mRNA for immunization goal. Please stop this misinformation.
DialoGPT: This is so unbelievably wrong. It is not gene therapy. The vaccine does not change DNA.
FC-GEN: It is misinformation. The vaccine is not gene therapy not gene therapy.

Table 3: Examples of generated counter-responses by the proposed and baseline methods.

Method	Polite. ↑	Refut. ↑	Evid. ↑	Perpl. ↓	Rele. ↑
DialoGPT	0.762	0.726	0.623	12.039	**0.940**
Seq2Seq	0.734	0.641	0.473	14.312	0.820
BART	0.723	0.721	0.607	13.079	0.893
Partner	0.781	0.709	0.632	11.993	0.825
FC-GEN	0.714	0.663	0.515	15.102	0.782
MisinfoCorrect	**0.854**	**0.797**	**0.643**	**10.110**	0.938

Table 4: Performance comparison of counter-response generators when trained on social media responses only.

6.4 RQ2: Impact of Dataset Quality

Here we examine the impact of the dataset quality on the quality of generated response. We train the model using *only a "clean" social* media responses (i.e., responses that are evidenced, refuting, neutral, or polite) and no crowdsourced counter-responses. The performance results are shown in Table 4. First, we observe that compared to Table 2, the quality of responses generated by each model degrades. This highlights the importance of collecting crowdsourced data, which is of higher quality compared to social media data. Second, we note that our proposed model still generates the best counter-responses as per all metrics, except in relevance, in which it performs the second best.

6.5 RQ3: Ablation Study

We examine the contribution of key components for effective counter-response generation (i.e., politeness, refutation and evidence rewards) in MisinfoCorrect on social media and crowdsourced responses data. We compare the model variations when using RL:
• *Base MisinfoCorrect model (Base):* this model is the basic GPT-2 model fine-tuned on our dataset in a dialog manner as DialoGPT [118], but without using any rewards for training.
• *Base + politeness reward:* we only consider the politeness reward
• *Base + refutation reward:* we only consider the refutation reward
• *Base + evidence reward:* we only consider the evidence reward.
• *MisinfoCorrect model:* this is the complete model with all the reward functions.

The results are shown in Table 5. When we only use the politeness, refutation or evidence reward function in the reinforcement learning framework, the corresponding politeness, refutation and evidence score is the highest and shows a significant increase compared to the Base model without any reward. When all the reward functions are combined in the MisinfoCorrect framework, there is a slight drop in each of the individual politeness, refutation, and evidence metrics, but it still has the second highest values along

Method	Polite. ↑	Refut. ↑	Evid. ↑	Perpl. ↓	Rele. ↑
Base	0.874	0.831	0.693	10.010	0.930
+ politeness	**0.953**	0.724	0.627	8.952	0.877
+ refutation	0.794	**0.968**	0.623	9.138	0.856
+ evidence	0.853	0.825	**0.753**	8.912	0.913
MisinfoCorrect	0.914	0.930	0.723	**8.010**	**0.960**

Table 5: Ablation study.

each dimension. This indicates that the MisinfoCorrect model finds a balance between the competing rewards during training.

6.6 RQ4: Qualitative Evaluation

Experimental Setup: In addition to the quantitative evaluation of response generation, we follow previous research works [35] and also conducted human evaluation experiments to qualitatively examine the model performance. In particular, we recruited 10 subjects following the same procedure described in the counter-response annotation process (Section 4.3). Each subject is presented 30 data points, where each data point consists of one misinformation post and two counter-responses, and then asked "which response is better when countering the misinformation post: the first, the second, or are they equally effective?". We test three settings: (1) the real counter-response versus the generated response by MisinfoCorrect; (2) the generated response by MisinfoCorrect versus the closest method, i.e., fact-checking generator (FC-GEN) [99]; (3) the generated response by MisinfoCorrect versus the most methodologically comparable baseline, i.e., DialoGPT [118]. We do not inform the subjects which response is generated by which method. Within each setting, we randomly pick 50 data points for comparison, and each data point is annotated by two users. In the analysis of the results, we only summarize the data points on which the two users provide the same label, i.e., disagreement cases are discarded. In total, we received 300 data points in human evaluation for the three settings.

Ethics: This protocol was approved by Georgia Tech's IRB.

Results: We get the following result:

(1) *Real response versus MisinfoCorrect:* In 46 out of 50 cases, both annotators provided the same answers. Among these, response generated by MisinfoCorrect were preferred in 76% cases, while in 6.5% cases, both responses were rated as equal. Real responses were preferred in the remaining cases.

(2) *FC-GEN versus MisinfoCorrect:* Annotators agreed in 44 out of 50 cases. Among these, MisinfoCorrect was preferred in 61.36% cases, 18.2% cases were equal, while 20.5% responses by FC-generator were better.

(3) *DialoGPT versus MisinfoCorrect:* Annotators agreed in 41 out of 50 cases. Among these, 36.6% cases prefer MisinfoCorrect, 36.6% cases are equal, and 26.8% cases prefer DialoGPT.

From all three comparison results, we can see that responses generated by MisinfoCorrect are preferred over the responses generated by the competing methods and the real responses. One representative example in Table 3 also illustrates the difference between these models and real responses. Altogether, the qualitative results show the potential for MisinfoCorrect in a real application to empower users to counter misinformation.

7 DISCUSSION AND LIMITATIONS

Generalization across topics, languages, and entire conversations: While MisinfoCorrect only studied one topic (COVID-19 vaccine misinformation) on one platform (Twitter) and in one language (English), the proposed model is general. It can be adopted for other topics easily by providing topic-specific data and content from other platforms. Non-English or multi-lingual language models can be used to develop response generator beyond English. Additionally, our method only generates one direct response and does not generate entire conversation (which can be the future work).

Intended use of the model: The model can be made available via a web portal or an API, where a user can input a misinformation post and our model will generate one or more counter-responses.

Backfire effect: While the backfire effect, i.e., potential increase in misperception due to observing the correction, has been debated for a long time [47, 63, 64], many large follow-up studies have failed to replicate backfire effect [31, 90, 114]. corrections already has proved to be effective by existing research works [14, 24, 69, 77, 109–111]

Counter-response may lead to online arguments: One may wonder whether using the generated counter-responses can lead to online arguments. Our model is intended to encourage users who voluntarily and proactively already counter misinformation to do so in a polite and respectful manner – recall that 96% of all counter-misinformation responses are already generated by ordinary users, even though 2 out of 3 times their responses are rude and abusive. Since our model generates polite responses, it has lower chance of leading to online fights.

Limitation of evaluation based on machine evaluation: The evaluation replying on classifiers can have limits and are faulty. This may lead to the inaccurate comparison results between models. More human evaluations are needed for a comprehensive comparison.

8 CONCLUSION

Overall, this work shows the potential to build on the recent advancements in generative text models to use them for social good applications. In this work, we extended these models for counter-misinformation response generation. Our proposed model showed promise by generating responses that were qualitatively and quantitatively better than real responses and other generated responses.

The future work lies in three directions: (i) deploying and evaluating the model in practice, (ii) collecting data from professional fact-checkers as expert-generated counter-responses and compare the model performance against the current setup, and (iii) developing multi-lingual and multi-modal model to generate visual counter-responses.

ACKNOWLEDGMENTS This research/material is based upon work supported in part by NSF grants CNS-2154118, IIS-2027689, ITE-2137724, ITE-2230692, CNS-2239879, and funding from Microsoft, Google, and Adobe Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the position or policy of NSF and no official endorsement should be inferred. We thank the CLAWS research group members for their help on the project.

REFERENCES

[1] Jennifer Abbasi. 2022. Widespread misinformation about infertility continues to create COVID-19 vaccine hesitancy. *JAMA* 327, 11 (2022), 1013–1015.

[2] Vibhor Agarwal, Sagar Joglekar, Anthony P Young, and Nishanth Sastry. 2022. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In *Proceedings of the ACM Web Conference 2022*. 2729–2737.

[3] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7 (9 2021). Issue 36. https://doi.org/10.1126/sciadv.abf4393

[4] 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 *CHI Conference on Human Factors in Computing Systems*. 1–19.

[5] Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-Argument Generation by Attacking Weak Premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1816–1827.

[6] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 20 (Nov. 2018), 27 pages. https://doi.org/10.1145/3274289

[7] Philip Ball and Amy Maxmen. 2020. The epic battle against coronavirus misinformation and conspiracy theories. https://www.nature.com/articles/d41586-020-01452-z.

[8] Leticia Bode and Emily K. Vraga. 2015. In Related News, That was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication* 65, 4 (06 2015), 619–638. https://doi.org/10.1111/jcom.12166 arXiv:https://academic.oup.com/joc/article-pdf/65/4/619/22320531/jjnlcom0619.pdf

[9] Leticia Bode and Emily K. Vraga. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication* 33 (9 2018), 1131–1140. Issue 9. https://doi.org/10.1080/10410236.2017.1331312

[10] Leticia Bode and Emily K. Vraga. 2021. Correction Experiences on Social Media During COVID-19. *Social Media + Society* 7 (4 2021), 205630512110088. Issue 2. https://doi.org/10.1177/20563051211008829

[11] Leticia Bode, Emily K Vraga, and Melissa Tully. 2020. Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020).

[12] Porismita Borah, Bimbisar Irom, and Ying Chia Hsu. 2021. 'It infuriates me': examining young adults' reactions to and recommendations to fight misinformation about COVID-19. *Journal of Youth Studies* (8 2021), 1–21. https://doi.org/10.1080/13676261.2021.1965108

[13] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. *Proceedings of the 20th international conference on World wide web - WWW '11*, 665. https://doi.org/10.1145/1963405.1963499

[14] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

[15] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

[16] Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783* (2021).

[17] Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior* 97 (2019), 202–215.

[18] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621* (2019).

[19] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Chiara Drolsbach and Nicolas Pröllochs. 2022. Diffusion of Community Fact-Checked Misinformation on Twitter. *arXiv preprint arXiv:2205.13673* (2022).

[22] Sarah Evanega, Mark Lynas, Jordan Adams, Karinne Smolenyak, and Cision Global Insights. 2020. Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic'. *JMIR Preprints* 19, 10 (2020), 2020.

[23] Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 313–319.

[24] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *proceedings of the international AAAI conference on web and social media*, Vol. 8. 101–110.

[25] Christian Fuchs and Christian Fuchs. 2021. Bill Gates Conspiracy Theories as ideology in the context of the COVID-19 crisis. (2021).

[26] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206.

[27] Maria Glenski, Svitlana Volkova, and Srijan Kumar. 2020. User engagement with digital deception. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 39–61.

[28] Mahak Goindani and Jennifer Neville. 2020. Social Reinforcement Learning to Combat Fake News Spread, Ryan P Adams and Vibhav Gogate (Eds.). *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference* 115, 1006–1016. https://proceedings.mlr.press/v115/goindani20a.html

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[30] Sukeshini Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. 2021. By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5 (7 2021), 1–24. Issue GROUP. https://doi.org/10.1145/3463930

[31] Andrew Guess and Alexander Coppock. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science* 50, 4 (2020), 1497–1515.

[32] Michael Hameleers. 2022. Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society* 25, 1 (2022), 110–126.

[33] Michael Hameleers and Toni GLA Van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research* 47, 2 (2020), 227–250.

[34] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public health* 203 (2022), 23–30.

[35] Bing He, Mustaque Ahamad, and Srijan Kumar. 2021. Petgen: Personalized text generation attack on deep sequence embedding-based classification models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 575–584.

[36] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 90–94.

[37] Christopher Hidey and Kathleen McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1756–1767.

[38] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sameer Singh, and Sean Young. 2020. Detecting covid-19 misinformation on social media. (2020).

[39] Albert L Hsu, Traci Johnson, Lynelle Phillips, and Taylor B Nelson. 2022. Sources of vaccine hesitancy: pregnancy, infertility, minority concerns, and general skepticism. In *Open forum infectious diseases*, Vol. 9. Oxford University Press US, ofab433.

[40] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. *arXiv preprint arXiv:1906.03717* (2019).

[41] Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. 2020. Modeling and measuring expressed (dis) belief in (mis) information. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 315–326.

[42] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 324–332.

[43] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4 (10 2020), 1–27. Issue CSCW2. https://doi.org/10.1145/3415211

[44] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*. 933–943.

[45] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

[46] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[47] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[49] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541* (2016).

[50] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media* 2 (8 2017), 19–31. https://doi.org/10.1016/j.osnem.2017.07.001

[51] 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* (2019).

[52] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.

[53] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816* (2020).

[54] Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and Predicting Social Correction on Twitter. In *15th ACM Web Science Conference 2023*.

[55] Pranav Malhotra, Kristina Scharp, and Lindsey Thomas. 2022. The meaning of misinformation and those who correct it: An extension of relational dialectics theory. *Journal of Social and Personal Relationships* 39, 5 (2022), 1256–1276.

[56] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* 35, 2 (2018), 196–219.

[57] Gina M. Masullo and Jiwon Kim. 2021. Exploring "Angry" and "Like" Reactions on Uncivil Facebook Comments That Correct Misinformation in the News. *Digital Journalism* 9 (9 2021), 1103–1122. Issue 8. https://doi.org/10.1080/21670811.2020.1835512

[58] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic. In *2020 IEEE international Conference on big data (big data)*. IEEE, 748–757.

[59] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 651–662.

[60] Kunihiro Miyazaki, Takayuki Uchiba, Kenji Tanaka, Jisun An, Haewoon Kwak, and Kazutoshi Sasahara. 2022. " This is Fake News": Characterizing the Spontaneous Debunking from Twitter Users to COVID-19 False Information. *arXiv preprint arXiv:2203.14242* (2022).

[61] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[62] Tasmiah Nuzhath, Samia Tasnim, Rahul Kumar Sanjwal, Nusrat Fahmida Trisha, Mariya Rahman, SM Farabi Mahmud, Arif Arman, Susmita Chakraborty, and Md Mahbub Hossain. 2020. COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data. (2020).

[63] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[64] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L Freed. 2014. Effective messages in vaccine promotion: a randomized trial. *Pediatrics* 133, 4 (2014), e835–e842.

[65] Gábor Orosz, Péter Krekó, Benedek Paskuj, István Tóth-Király, Beáta Bőthe, and Christine Roland-Lévy. 2016. Changing Conspiracy Beliefs through Rationality and Ridiculing. *Frontiers in Psychology* 7 (10 2016). https://doi.org/10.3389/fpsyg.2016.01525

[66] Anjan Pal, Y Alton, et al. 2019. Rumor analysis visualization system. *Proceedings of the international multi conference of engineers and computer scientists*.

[67] Tanja Pavleska, Andrej Školkay, Bissera Zankova, Nelson Ribeiro, and Anja Bechmann. 2018. Performance analysis of fact-checking organizations and initiatives in Europe: a critical overview of online platforms fighting fake news. *Social media and convergence* 29 (2018), 1–28.

[68] Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific reports* 12, 1 (2022), 1–7.

[69] Ethan Porter and Thomas J Wood. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences* 118, 37 (2021), e2104235118.

[70] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 794–805.

[71] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251* (2019).

[72] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[73] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[74] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. *arXiv preprint arXiv:2106.03794* (2021).

[75] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Aspect-controlled neural argument generation. *arXiv preprint arXiv:2005.00084* (2020).

[76] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2021. (In) effectiveness of Accumulated Correction on COVID-19 Misinformation. (2021).

[77] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2022. If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 896–907.

[78] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*. 194–205.

[79] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. COVID-19 Vaccine Misinformation Campaigns and Social Media Narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 920–931.

[80] Jieun Shin and Kjerstin Thorson. 2017. Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. *Journal of Communication* 67, 2 (02 2017), 233–255. https://doi.org/10.1111/jcom.12284 arXiv:https://academic.oup.com/joc/article-pdf/67/2/233/22321279/jjnlcom0233.pdf

[81] Craig Silverman. 2015. Lies, damn lies, and viral content: How news websites spread (and Debunk) online rumors, unverified claims and misinformation. *Tow Center for Digital Journalism* 168, 4 (2015), 134–140.

[82] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed news* 16 (2016).

[83] Ingjerd Skafle, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, Elia Gabarron, et al. 2022. Misinformation about COVID-19 vaccines on social media: rapid review. *Journal of medical Internet research* 24, 8 (2022), e37367.

[84] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *IConference 2014 Proceedings* (2014).

[85] Maryke S Steffens, Adam G Dunn, Kerrie E Wiley, and Julie Leask. 2019. How organisations promoting vaccination respond to misinformation on social media: a qualitative investigation. *BMC public health* 19, 1 (2019), 1–12.

[86] Leo G Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, Vol. 70.

[87] Ana Stojanov. 2015. Reducing conspiracy theory beliefs. *Psihologija* 48 (2015), 251–266. Issue 3. https://doi.org/10.2298/PSI1503251S

[88] Yanqing Sun, Jeffry Oktavianus, Sai Wang, and Fangcao Lu. 2021. The Role of Influence of Presumed Influence and Anticipated Guilt in Evoking Social Correction of COVID-19 Misinformation. *Health Communication* (2 2021), 1–10. https://doi.org/10.1080/10410236.2021.1888452

[89] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).

[90] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of applied research in memory and cognition* 9, 3 (2020), 286–299.

[91] Yuko Tanaka and Rumi Hirayama. 2019. Exposure to Countering Messages Online: Alleviating or Strengthening False Belief? *Cyberpsychology, Behavior, and Social Networking* 22 (11 2019), 742–746. Issue 11. https://doi.org/10.1089/cyber.2019.0227

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

[93] Kjerstin Thorson, Emily Vraga, and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society* 13, 3 (2010), 289–313.
[94] Melissa Tully, Leticia Bode, and Emily K. Vraga. 2020. Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users' Responses to a Health Misinformation Post? *Social Media + Society* 6 (10 2020), 205630512097837. Issue 4. https://doi.org/10.1177/2056305120978377
[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[96] Jeyasushma Veeriah. 2021. YOUNG ADULTS'ABILITY TO DETECT FAKE NEWS AND THEIR NEW MEDIA LITERACY LEVEL IN THE WAKE OF THE COVID-19 PANDEMIC. *Journal of Content, Community and Communication* 13 (2021), 372–383. Issue 7.
[97] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (2022), 1–9.
[98] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 275–284.
[99] Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–344.
[100] Nguyen Vo and Kyumin Lee. 2020. Standing on the shoulders of guardians: Novel methodologies to combat fake news. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 183–210.
[101] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[102] Emily Vraga, Melissa Tully, and Leticia Bode. 2021. Assessing the relative merits of news literacy and corrections in responding to misinformation on Twitter. *New Media & Society* (2021), 1461444821998691.
[103] Emily K Vraga and Leticia Bode. 2018. I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society* 21, 10 (2018), 1337–1353.
[104] Emily K Vraga and Leticia Bode. 2020. Correction as a solution for health misinformation on social media. , S278–S280 pages.
[105] Emily K Vraga and Leticia Bode. 2021. Addressing COVID-19 misinformation on social media preemptively and responsively. *Emerging infectious diseases* 27, 2 (2021), 396.
[106] Emily K Vraga, Leticia Bode, and Melissa Tully. 2021. The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health communication* (2021), 1–9.
[107] Emily K Vraga, Sojung Claire Kim, John Cook, and Leticia Bode. 2020. Testing the effectiveness of correction placement and type on Instagram. *The International Journal of Press/Politics* 25, 4 (2020), 632–652.
[108] Mason Walker and Katerina Eva Matsa. 2021. News consumption across social media in 2021. (2021).
[109] Nathan Walter, John J Brooks, Camille J Saucier, and Sapna Suresh. 2021. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication* 36, 13 (2021), 1776–1784.
[110] Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37, 3 (2020), 350–375.
[111] Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs* 85, 3 (2018), 423–441.
[112] Zhihong Wang and Yi Guo. 2020. Rumor events detection enhanced by encoding sentimental information into time series division and word representations. *Neurocomputing* 397 (7 2020), 224–243. https://doi.org/10.1016/j.neucom.2020.01.095
[113] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2021. Effect of Conformity on Perceived Trustworthiness of News in Social Media. *IEEE Internet Computing* 25 (1 2021), 12–19. Issue 1. https://doi.org/10.1109/MIC.2020.3032410
[114] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41, 1 (2019), 135–163.
[115] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.
[116] Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *arXiv preprint arXiv:1809.06873* (2018).
[117] Jingwen Zhang, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine* 145 (2021), 106408.
[118] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).
[119] Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. 2022. "This is Fake! Shared it by Mistake": Assessing the Intent of Fake News Spreaders. In *Proceedings of the ACM Web Conference 2022*. 3685–3694.
[120] Jianming Zhu, Smita Ghosh, and Weili Wu. 2021. Robust rumor blocking problem with uncertain rumor sources in social networks. *World Wide Web* 24 (1 2021), 229–247. Issue 1. https://doi.org/10.1007/s11280-020-00841-8
[121] Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625* (2021).
[122] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1097–1100.

A APPENDIX

A.1 Data Annotation and Collection

The brief guide to writing counter-misinformation responses is:

Application Se ing:
On Twitter, when someone writes a misinformation tweet, we would like to write a reply to counter the misinformation such that we can mitigate the spread of misinformation.
Guidance:
Please write a response like you would try to engage or counter the misinformation. When writing replies, you may want to consider the following:
i. You may want to refute the tweet or express disagreement towards the tweet;
ii. You may want to include supporting sentences, reasons, or evidence to make the reply reliable;
iii. You may want to be polite, avoid confrontation, or avoid any impolite or rude expressions in the response.
One Example:
Tweet: "The Biden "vaccine passport" is here. You either get a non-FDA approved experimental gene modification therapy (euphemistically called a "vaccine") or you'll be denied access to public transportation, sports venues, air travel and more. Obey or stay home.";
Reply: "To correct you: it is not a gene modification therapy, there is no proof of this nor a scientific rationale, mRNA does not integrate in the human genome. I am for freedom of choice and against vaccine passports but let's stick to the facts"
One Tweet You Write A Response To:
Tweet: "It's not a vaccine, it's gene therapy. Gene therapy is an experimental technique. It's the same technology used in cloning, DNA editing, and stem cell research."
Your Response: "Sorry to see you think in this way. It is not correct. The vaccine is not gene therapy. Instead, it uses mRNA to generate spike protein to protect people. Please do not say the misinformation again."

A.2 Experiment Details

Some experiment details are included here:

- During our experiments, all methods are fine-tuned or trained from scratch on the annotated tweet-response pairs from the in-the-wild and crowdsourced datasets from Section 4.
- We cut off responses beyond 280 characters per Twitter's rule. Note that from Feb 2023, Twitter extends the character limit

to 4,000. But, when we did this project, the 280-character limit still held.

- For training, MisinfoCorrect takes $7 \sim 10$ minutes while compared methods use $3 \sim 10$ minutes. The inference time is comparable across all methods using $0.3 \sim 0.5$ seconds per example. The one-time longer training ensures higher-quality of generated text.
- In experiments, we set the batch size, training steps, learning rate at 8, 10,000 and 1e-5, respectively. For the the hyper-parameter $\alpha, \beta, \theta, \gamma, \lambda$, we used a grid-search-based method with three values (0.1, 1, 10), and we selected $\alpha = 1, \beta = 1, \theta = 10, \gamma = 1, \lambda = 0.1$. When comparing different methods and reporting results, we take the consistent sampling across all methods for evaluation.

A.3 Limitations

- *False positives or negatives in misinformation detection classifier may lead to misinformation spread*: We will clearly inform users that the classifier may make mistakes and also provide them links to professional fact checking websites so they can check the content themselves.
- *Bias and partisanship in counter-response behavior by users*: Users tend to follow partisan lines in crowdsourced fact-checking. While the model is not designed to reduce partisan-bias among users, we expect that the model will generate counter-responses only for misinformation posts, regardless of the partisan leaning.
- *Erroneous output and potential for our model's misuse*: To prevent generating unreasonable responses for unknown topics of misinformation or for non-misinformation tweets, we can create a filtering step so that the model will only output a counter-response if the tweet is a misinformation post on the topic(s) it is trained on.
- *Harms due to exposure to misinformation in the annotation process*: We did not expose ordinary social media users to misinformation. Misinformation was shown to crowdworkers (college students) to write counter-responses. We informed the crowd-workers up-front that the content is verified misinformation and they are supposed to write counter-responses. We also provided them with fact-checking resources. This protocol was approved by Georgia Tech's IRB.