Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions

Jiawei Zhou
Georgia Institute of Technology
Atlanta, GA, USA
j.zhou@gatech.edu

Yixuan Zhang
Georgia Institute of Technology
Atlanta, GA, USA
yixuan@gatech.edu

Qianni Luo
Ohio University
Athens, OH, USA
ql047311@ohio.edu

Andrea G Parker
Georgia Institute of Technology
Atlanta, GA, USA
andrea@cc.gatech.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, GA, USA
munmund@gatech.edu

ABSTRACT

Large language models have abilities in creating high-volume human-like texts and can be used to generate persuasive misinformation. However, the risks remain under-explored. To address the gap, this work first examined characteristics of AI-generated misinformation (AI-misinfo) compared with human creations, and then evaluated the applicability of existing solutions. We compiled human-created COVID-19 misinformation and abstracted it into narrative prompts for a language model to output AI-misinfo. We found significant linguistic differences within human-AI pairs, and patterns of AI-misinfo in enhancing details, communicating uncertainties, drawing conclusions, and simulating personal tones. While existing models remained capable of classifying AI-misinfo, a significant performance drop compared to human-misinfo was observed. Results suggested that existing information assessment guidelines had questionable applicability, as AI-misinfo tended to meet criteria in evidence credibility, source transparency, and limitation acknowledgment. We discuss implications for practitioners, researchers, and journalists, as AI can create new challenges to the societal problem of misinformation.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

large language model, GPT, misinformation, generative AI, AI-generated misinformation, COVID-19, responsible AI

ACM Reference Format:
Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3544548.3581318

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581318

1 INTRODUCTION

The Coronavirus Disease (COVID-19) pandemic has brought attention to the proliferation of health misinformation[1]. From fake cures to conspiracy theories, misinformation has led to substantial adverse effects at the individual as well as societal levels. Examples of such effects include mortality and hospital admissions [20, 48], public fear and anxiety [78, 106], eroded trust in health institutions [86], and exacerbated racial discrimination and stigma [41, 48]. Finding ways to combat misinformation is therefore of critical importance from the perspectives of both public health and governance. Manual identification of misinformation is, however, extremely laborious and often does not scale: a key issue given the rise of misinformation on social media [70]. As such, artificial intelligence (AI) techniques have been touted as a timely and scalable solution for misinformation detection when compared to manual efforts [3, 25].

Unfortunately, AI techniques are far from being a savior in the battle against misinformation, but instead, can be used to generate misinformation [14]. For example, *Large Language Models (LLMs)* — machine learning algorithms that can recognize, predict, and generate human languages on the basis of large sets of human-written content [13] — are now widely used in producing human-like texts. Leveraging the power of LLMs, AI-generated content is increasingly indistinguishable from human-written information, and in certain cases even perceived to be more credible [59].

Once LLMs are used for generating misinformation, the ease and speed of producing high-volume text can significantly magnify views that are otherwise fringe or outright misleading, by creating an illusion of a majority perspective [26]. The spread of misinformation is already known to precipitate a distrustful environment, but what is new is AI's ability to easily and quickly generate persuasive misinformation. The scalability presents malicious actors with a new tactic to perpetuate false narratives to unsuspecting users, which may create public confusion at a scale not previously possible. In November 2022 – although released with the best of intentions – Meta had to take down Galactica three days after release, an LLM for science that can "summarize academic papers, solve math problems, (and) generate Wiki articles" [43]. During that time, Galactica was found to generate biased and even incorrect results with fake papers and sometimes attribute references

[1]We focus on misinformation as a broad category that concerns "false or partially false information which can be spread both unintentionally and intentionally" [80].

to real researchers [43]. One month later, the question-answering chatbot ChatGPT was released and gained million-plus users in five days [73]. As ChatGPT further brought LLMs into the public eye, it was also criticized for biased or false outputs [8].

Despite the capabilities of AI fabricating fluent and seemly credible misinformation, little work has examined the differences between AI-generated and human-created misinformation or assessed the extent to which pre-existing solutions are applicable to AI-generated misinformation. With off-the-shelf LLMs becoming more accessible to the general public, this work responds to this critical gap. Leveraging a state-of-the-art (SOTA) LLM, we seek to answer the following research questions:

RQ1. What are the characteristics of AI-generated misinformation compared with human-created misinformation?
RQ2. How do existing misinformation detection models perform on AI-generated misinformation?
RQ3. How do existing assessment guidelines for spotting misinformation work on AI-generated misinformation?

We situate our work in a health crisis, a highly polarized and uncertain time during which people are susceptible to worry, vulnerability, and rumors [95, 114]. To answer the research questions, we first compiled a dataset of human-created misinformation from existing work [18, 82, 89] and extracted the most representative documents. Guided by Narrative Theory [17], we abstracted representative documents into what we labeled as *"narrative prompts"* that captured the core narrative elements. Those prompts were then used for a SOTA LLM GPT-3 [13] to output AI-generated misinformation that is paired with human-created content.

We examined the characteristics of AI-generated misinformation through text analysis and rapid qualitative analysis. Our results suggest significant linguistic differences in AI-generated misinformation as it had more emotions and cognitive processing expressions than human creations. We also observed that AI-generated misinformation tended to enhance details, communicate uncertainties, draw conclusions, and simulate personal tones. Next, we evaluated two common and pre-existing misinformation solutions: misinformation detection models and information assessment guidelines developed by journalists. We discovered existing detection models had performance degradation when classifying AI-generated misinformation as opposed to human creations. Similarly, information assessment guidelines had questionable applicability, as AI-generated misinformation was more likely to mimic criteria in credibility, transparency, and comprehensiveness.

Overall, this work makes three contributions. **(1)** We offer a comprehensive understanding of AI-generated misinformation and its risks. To our knowledge, this is the first work to examine the characteristics of AI-generated misinformation and how existing solutions work on AI generations. **(2)** We propose a theory-guided approach to compile AI-generated misinformation comparable with human creation, allowing us to explore relative differences. Accordingly, we also contribute an AI-generated (i.e., GPT-3) dataset to facilitate future research in AI and misinformation[2]. **(3)** We provide empirical evidence on the risks of LLMs and discuss implications for adapting current misinformation solutions in collective efforts from practitioners, researchers, and journalists, as well as developments of moderation strategies.

Content Warning. We caution the readers that some of the misinformation examples included in this paper, for the purposes of better explication of results, can be profusely misleading and/or outright false. Some readers may find certain expressions to be offensive, divisive, violent, or emotionally triggering.

2 BACKGROUND AND RELATED WORK

In this work, misinformation is referred to as the umbrella term that includes "false or partially false information which can be spread both unintentionally and intentionally" [80]. We chose to focus on misinformation as a broader category than disinformation which implies an intention to deceive or mislead people [44]. Below we first give an overview of generative AI and its role in misinformation, as well as pre-existing algorithm- and human-driven solutions to misinformation. Then we provide background on COVID-19-related misinformation that serves as the topical focus of our work.

2.1 Generative AI and Its Role in the Age of Misinformation

This study is motivated by the rapidly improving capabilities and accessibility of generative AI that can use training data to generate content in the forms of text, images, audio, and videos. The past decade has witnessed uplifting progress and wide applications in art [87], journalism [16], and screenplay [21]. Yet, behind the hype and landmark advancement is the concern of misuse and lack of governance. Researchers and journalists have tried to rein back the technology-centric complacency by calling out the issues of biases [11, 99], stereotypes [11, 52, 76], and malicious use [11, 34] in AI-generated content.

Doctoring of content is not new; as part of his Great Purge [35] and to alter history books with revisionist views favorable to his regime and the Soviet Communist Party, Joseph Stalin removed Nikolai Yezhov from still images [98]. Today, creating misinformation does not require the types and extents of power wielded by the likes of Stalin. With the rapid evolution of generative AI [11, 110], it is increasingly easier to doctor content and perpetuate falsehoods at scale, whether by bad actors, discreet amateurs, or others. It has been shown that AI tools are capable of creating deep fakes of political leaders by adapting their actual video, audio, and pictures [105, 110] – such as one in which Barack Obama was seen calling Donald Trump "a total and complete dipshit" [31]. Media outlets have thus highlighted the dangers of deep fake: it has been used to sow the seeds of discord in society and create chaos in public discourse [23, 26, 55]. As such, it is crucial to examine the role of generative AI as misinformation surges as a prominent challenge in and threat to the healthy functioning of the public sphere and democracy [60].

However, little work has investigated the plausibility and risks of AI-generated misinformation. Some exceptional examples include Gamage et al. [32] which examined deepfake-related conversations on Reddit and found that people did not pay attention to the harms of AI-generated misinformation. While people expressed concerns about deep-faked videos, there was significantly less awareness

[2]This dataset can be made available to researchers, subject to adequate data usage agreements.

of the deceiving power of AI-generated text. Kreps et al. [59] conducted experiments on AI-generated and human-written news articles and found that people could not distinguish between AI- and human-generated texts and that AI-generate news was perceived to be equally or more credible than human-written articles. Buchanan et al. [14] demonstrated LLM's ability to create moderate-to-high quality misinformation messages with little human involvement. They also discovered that AI-generated misinformation could customize language for specific groups and sometimes deploy stereotypes and racist language on certain topics. Yet, to fully understand the potential risks of AI-generated misinformation, there is still a gap in the literature regarding how existing misinformation solutions work on AI-generated misinformation. Our study builds on prior scholarship and fills this gap by evaluating the applicability of existing solutions.

2.2 Algorithmic and Human Solutions for Misinformation

Prior work has explored algorithm- and human-centered approaches to address the issues of misinformation. Algorithmic-centered approaches primarily focus on automatic misinformation detection and correction, and characterization of misinformation and its creators [3]. On the other hand, human-centered approaches study how experts or crowds can help combat misinformation, and ways to influence human perceptions and behaviors to misinformation [3].

Algorithm-Centered Solutions. Corrections made by detection models are proven to help reduce people's belief in misinformation [101]. Accordingly, a sizable number of methodological studies in AI and machine learning have explored misinformation detection models through the use of linguistic, syntactical, semantic, and social features and achieved high predictability [37, 71]. For example, in the AAAI 2021 COVID-19 Fake News Detection challenge [82], the winning team achieved a weighted F1-score of 0.987 in classifying fake news from social media post data [37]. While most detection efforts focus on content-level features, some work highlights the importance of contextual factors and proposes frameworks to translate and operationalize publisher-news-user tri-relationship [96] and intrinsic uncertainty of misinformation detection [58]. Other works have focused on engineering novel solutions that consider the holistic multimodal context surrounding information and misinformation [97].

Despite the fruitful outcomes of modeling impact factors of misinformation and enhancing detection accuracy, some scholars have raised concerns regarding the generalizability of detection models [6, 100]. Specifically, little is known about the applicability of pre-existing models to AI-generated misinformation. Therefore, our work seeks to address this gap by evaluating the generalizability of existing misinformation detection models on AI-generated text, a hitherto less explored form and source of misinformation.

Human-Centered Solutions. Research has also investigated how experts and the general public can help combat misinformation. One of the most commonly explored approaches is fact-checking – the process of evaluating information veracity and correctness. Research has demonstrated its efficacy in helping debunk fake news [19, 101] but also acknowledged that the labor-intensive nature made it challenging to scale [70]. However, it remains unclear as to what extent fact-checking, whether by humans or assisted by algorithms, is agile and adaptable to the rapidly evolving misinformation ecosystem [38]. Orthogonally, researchers have questioned the efficacy of fact-checking and the potential backfire effect where corrections inadvertently reinforce misinformation [64]. Studies show being corrected could decrease the quality and increases language toxicity in subsequent retweets [75], even when the corrections came from a bot [5]. While some researchers have explored the psychological foundations of this phenomenon [39], others have advocated unpacking the characteristics of fake news to understand why corrections may not necessarily stick to consumers [61, 84]. Here we study the characteristics of one such type of misinformation, one that is AI-generated.

Another common approach to mitigating the harmful effects of misinformation is improving information literacy. Journalists, scholars, and government agencies have developed information assessment guidelines [56, 112, 113, 113] to help people spot misinformation. Educational approaches have been found to be helpful in increasing the likelihood of identifying misinforming news [53]. However, current educational efforts have failed to address the current media environment [15] and technology capabilities [88]. Relevantly, little is known about the effectiveness of currently-available guidelines on AI-generated misinformation. As such, there is an urgent need for scholars and educators to create an up-to-date media literacy education agenda with consideration of AI capabilities. Our work seeks to address this need by examining misinformation generated by a SOTA language model.

2.3 COVID-19 and Misinformation

We situate our work in the COVID-19 pandemic and highlight this crisis background to help readers interpret our results. COVID-19 misinformation has spread quickly since the start of the pandemic and has covered a variety of content, ranging from prevention, treatment, vaccine, and politics [12]. There are mortality and hospital admissions resulting from the misinformation that drinking methanol or alcohol-based cleaning products can cure the virus [20]. Polling shows that 28% of Americans think that Bill Gates uses vaccines to implant microchips in people [20]. The tremendous amount and the great variety of misinformation were also associated with the nature of this public health crisis – featured by its high level of polarization, constantly changing situations, and high level of uncertainty [91, 114]. When facing a great level of uncertainty, it is challenging to think rationally and people are more vulnerable to worry and conspiracy ideas [95]. Accordingly, individuals turn to unofficial sources for information [45, 102]. This is also when misinformation campaigns effectively generate confusion [102].

Prior work studying COVID-19-related misinformation has primarily focused on curating datasets of COVID-19 rumors and misinformation [18, 82, 89], understanding how people perceives and processes misinformation [62], as well as detection of COVID-19 misinformation [37, 94]. However, little work has examined the potential "darker" side of AI in enabling misinformation in the context of a health crisis. As such, we note the vastly missing consideration of AI risks in existing COVID-19 misinformation datasets [18, 82, 89] and detection models [37, 94]. Our work extends this emergent scholarship by contributing a more comprehensive examination of AI-generated misinformation and its risks.

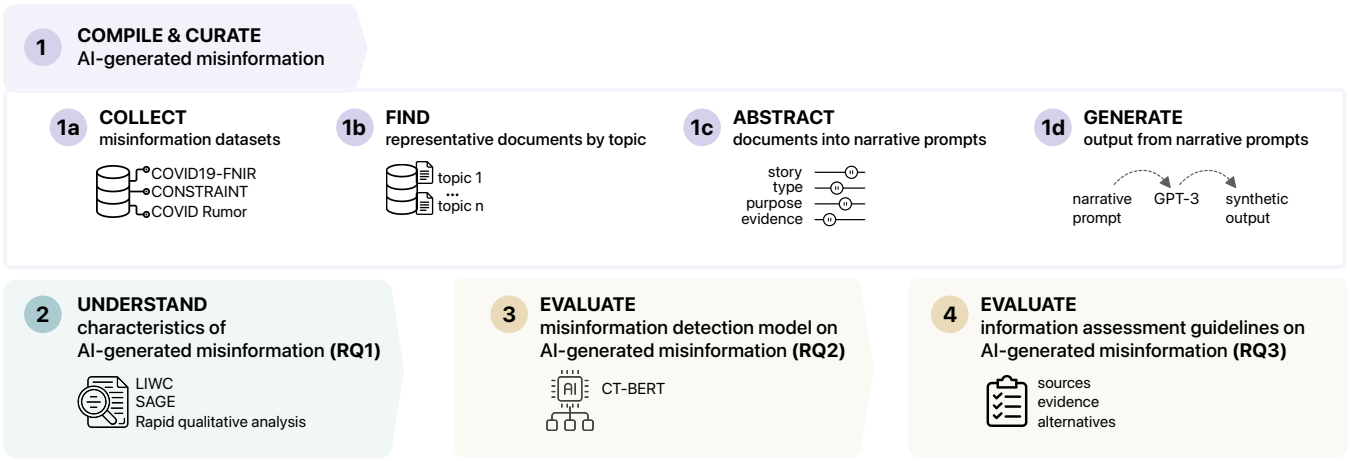


Figure 1: An overview figure summarizes the major steps of our study: 1 compiling and curating AI-generated misinformation that includes four sub-steps 1a 1b 1c 1d (Section 4), 2 understanding characteristics of AI-generated misinformation (Section 5), 3 evaluating misinformation detection models on AI-generated misinformation (Section 6), and 4 evaluating information assessment guidelines on AI-generated misinformation (Section 7).

3 STUDY OVERVIEW

While prior work has acknowledged the plausibility and risks of AI-generated misinformation [14, 59, 77], there is no currently available data that is large enough and comparable with human creation. Therefore, as shown in Figure 1, the first part of our study involved compiling and curating a dataset of AI-generated misinformation (Section 4), allowing us to investigate the differences between AI-generated and human-created misinformation.

LLMs can generate a text completion based on an inputted prompt that specifies and instructs the task, such as "write about why vaccines can protect people from COVID-19". For our work, we utilize this feature of LLMs as a generative approach, a technique that has been adopted in recent research for compiling datasets on news stories [59] and greeting card messages [99]. To create AI-generated misinformation that is comparable to human-created content while leaving linguistic flexibility for AI creation, we propose an approach to collecting paired human- and AI-misinformation. This approach consists of four steps: 1a identifying human-created COVID-19 misinformation that is manually annotated and peer-reviewed (Section 4.1), 1b extracting topic clusters and the most representative documents to typify the human-created misinformation dataset (Section 4.2), 1c abstracting the representative documents into "narrative prompts" that characterize the core elements or salient attributes in documents (Section 4.3), 1d generating synthetic outputs from a state-of-the-art LLM GPT-3 using narrative prompts (Section 4.4).

Thereafter, corresponding to RQ1, we studied the linguistic features and expression patterns to examine the characteristics of the thus compiled AI-generated misinformation compared with the original human creations, as described in Section 5 (labeled as 2 in Figure 1). Next, we assessed how current solutions for misinformation work (or do not work) on AI-generated misinformation. Specifically, for RQ2, we evaluated existing misinformation detection models on classifying AI-generated misinformation in Section 6 (labeled as 3 in Figure 1). For RQ3, we assessed to what extent do information assessment guidelines work on AI-generated misinformation in Section 7 (labeled as 4 in Figure 1).

4 COMPILING AN AI-GENERATED MISINFORMATION DATASET

As introduced in Section 3, the first part of this study is to compile and curate an AI-generated misinformation dataset that is comparable with human-created content. In this task, our goal is to ensure that the compiled AI-generated misinformation reflects the salient narratives in human-created misinformation, while also having linguistic flexibility in the creation process. Specifically, we first summarized human-created misinformation with the most representative documents in various topical clusters and then extracted their core elements (story, type, purpose, evidence) through a content analysis guided by Narrative Theory [17]. Those core elements helped us abstract representative documents into what we label as "narrative prompts", which were then used in a SOTA large language model to output AI-enabled misinformation.

4.1 Collect Human-created Misinformation

As our first step to compile AI-generated misinformation (1a in Figure 1), we utilized peer-reviewed COVID-19 misinformation datasets collected from news websites and social media platforms. Datasets were selected based on five criteria: (1) the content is in English, (2) labels are publicly available, (3) the dataset is about COVID-19 misinformation, (4) the determination of information veracity is annotated manually by referring to fact-checked sources rather than through algorithms, and (5) the dataset or publication attached to it is peer-reviewed. This gave us three datasets:

- COVID19-FNIR [89]: 3,727 fake news scraped from Poynter, collected between February and June 2020 and published at IEEE

Table 1: Human-created COVID-19 misinformation dataset. News includes information and articles from news agencies and media organizations, in formats of newspapers, online platform posts, radio, and cable.

Data	News stories	Social media	Total
COVID19-FNIR	3,727	/	3,727
CONSTRAINT	/	5,100	5,100
COVID-Rumor	3,041	540	3,581
Total	6,768	5,640	12,408

Dataport. This data spans from India, United States, and European regions. For our purpose, we exclude posts that are recorded as mostly true (11), in dispute (1), and no evidence (56).

- CONSTRAINT [82]: 5,100 misinformation posts gathered from social media platforms such as Twitter, Facebook, and Instagram. This dataset was collected before December 2020 and published at AAAI 2021 the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation.
- COVID Rumor [18]: 3,581 misinformation posts from news reports and Twitter (excluding true and unverified posts), among which 3041 are news and 540 are tweets. This dataset was collected between December 2019 and March 2020, and published at Frontiers in Psychology.

Since these datasets do not note the presence of any AI-generated or synthetic text, we assume them as human-created misinformation. Our human-created misinformation datasets were collected mostly before June 2020 and no later than December 2020. At at that time, GPT-3 was the only open-to-public LLM with proven proficiency and its API was released in late June 2020[3]. Although less likely, we acknowledge this assumption might not be true and discuss the potential downstream impacts in the Limitations section (Section 8.3). Annotations of three datasets were all made by researchers and cross-validated to authoritative fact-checking organizations such as Snopes and PolitiFact. The veracity labeling was determined based on the best knowledge at the annotation time. Table 1 summarizes the combined human-created COVID-19 misinformation dataset with a size of 12,408 (dataset addressed as human-misinfo for the rest of this paper).

4.2 Extract Representative Documents

To summarize and present the human-misinfo dataset, we applied topic modeling to gather latent topic clusters and the most representing documents within each topic (1b in Figure 1). We chose to employ Latent Dirichlet Allocation (LDA) [9], an unsupervised machine learning algorithm widely used for analyzing a corpus of documents to reveal latent prevalent topic distributions [115]. We pre-processed and cleaned the data through tokenizing, stop word removal (e.g., 'but', 'that', 'what'), and lemmatization. We included n-grams (n=1,2) with a frequency of appearance greater than 5, and then converted our dataset into a bag-of-words. Since LDA does not determine the optimal number of topics, we used the coherence measure as a metric to find the optimal number of topics for the best model fit. The coherence metric measures how words within a topic tend to co-occur, with evidence showing it correlates with expert opinions of topic quality [72]. We altered the number of topics (k) from 5 to 30 and calculated the coherence score for the pre-processed corpus. Fig. 2 shows the coherence score distribution with the highest coherence score at k=5. Table 2 gives an overview of the five topic clusters with top keywords and example posts.

LDA calculates a document' percentage of contribution in topic clusters, which allows us to find the most representative documents. Specifically, we selected the most representative 50 documents in each topic cluster, giving us a total of 250 documents for annotation. Among the 250 documents, 97 are news stories and 153 are social

[3]https://openai.com/blog/openai-api/

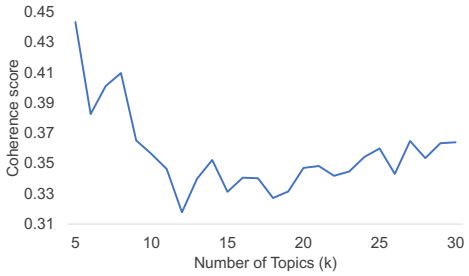


Figure 2: Coherence scores for topic modeling in human-created misinformation (highest value observed at k=5).

Table 2: List of topics in human-misinfo with top keywords and example posts.

Topic#	Top 10 Keywords	Freq
1	people, spread, kill, test, chinese, infect, prevent, quarantine, corona, home	23.52%
"A natural remedy that kills coronavirus. Start pot of boiling water on stove. Cut peels of oranges or lemons or both, your choice. Add sea salt to pot of boiling water. Add orange or lemon peels to pot of boiling hot water. Boil on high for a few minutes. When water and ingredients in pot have been brought to a boil, turn down the heat, put your face down to pot and breathe in steam. Do this for 15 minutes or as much as you can stand."		
2	virus, vaccine, china, cure, au, new_coronavirus, mask, find, call, work	21.52%
"An asymptomatic person is a HEALTHY person. He is someone who has a virus but his body developed antibodies. This is called attenuated virus, which means that he dominated the virus thanks to his healthy lifestyle habits. This person does not spread the virus, but communicates antibodies to the rest of the people and generates herd immunity."		
3	india, lockdown, due, government, state, outbreak, country, infection, report, city	18.58%
"The State of Florida has announced measures all workplaces with 10 employees or more are to have paid mandatory leave to avoid the spread of the COVID-19 coronavirus starting on March 6, 2020. All schools are to close for 2 weeks also from March 6th. Offices will resume after 2 weeks of the mandatory closure. A list of all schools and businesses in your area are shown on the list."		
4	claim, video, hospital, die, patient, show, doctor, italy, man, time	18.96%
"A photo of a man and woman embracing has been shared hundreds of times on Facebook and Twitter alongside a claim that it shows two Italian doctors who died of a novel coronavirus, COVID-19, after contracting the disease from the patients they treated"		
5	https_co, pandemic, case, death, mail, wuhan, make, day, news, trump	17.42%
"More racist shit to distract you. Illegal immigration's happening worldwide, engineered by all UN member nations to distract attention from the real issue. Totalitarianism, the fake pandemic and the removal of human rights in the development of mass democide and a #NewWorldOrder"		

media posts. For the purpose of content analysis in the next step, mega threads and fact-checking corrections were excluded and replaced with the next representative documents. Mega threads refer to documents that contain a collection of information, and fact-checking corrections are posts made by professional fact-checking organizations and contain both misinformation and corrections.

4.3 Abstract Documents into Narrative Prompts

The third step was extracting key elements in the 250 representative human-created misinformation from Section 4.2 while detaching modifiers and secondary details (1c in Figure 1). We conducted a content analysis guided by Narrative Theory [17, 22] to abstract representative documents into what we label as *"narrative prompts"* that summarize the story, type, purpose, and evidence in narratives.

4.3.1 Theoretical Framework: Narrative Theory. The intuition of using Narrative Theory is to help distinguish the story itself from its representation. Previous work has successfully adopted it to extract elements in a narrative [54]. In the eyes of formalist-structuralist narrative theory, each narrative contains two elements – the story and the discourse [17]. ***Stories*** contain the content of narratives through *events* (e.g., actions and happening) and *existents* (e.g., characters and settings) [17]. The story structure (event + existent) is similar to subject-verb-object triplets of key elements in previous work in extracting conspiracy theories [92], e.g., "lab created COVID-19". ***Discourse***, on the other hand, refers to the form and substance of expression [17]. Drawing from previous work in journalism communication [57] and misinformation linguistic signals [51], we specified the expression form as (i) *type* and (ii) *purpose*, and referred expression substance to the (iii) *evidence*. Specifically, type includes news reports and non-news posts, as previous work has demonstrated the impact of article genre and publisher on credibility assessments [7, 30, 33, 49]. Then, we categorized purposes into reports, instructions, and commentary, following the journalism practice of differentiating reporting and opinions [104] – a practice adopted by news agencies and organizations such as the Wall Street Journal, New York Times, Fox News, etc. Lastly, evidence refers to any form of information used to support a point.

4.3.2 Content Analysis and Prompt Creation. This section explains how we conducted a content analysis based on Narrative Theory [17] to create narrative prompts. Narrative prompts focus on four core elements in a message: story, purpose, evidence, and type. Elements of story, purpose, and evidence were captured through a content analysis, while the type element was manipulated as a variable to give us two subsets of news and posts. Purpose and evidence were coded by two researchers separately. Inter-rater reliability was calculated (results below) and any coding mismatch was discussed. Core stories were summarized by one researcher and verified by another researcher, and rejected summarizations were discussed between the two researchers till reaching an agreement. Table 3 overviews all elements in a narrative prompt with definitions and examples of each element. As an example, one narrative prompt is "write a **news *[type]*** to give **instructions on how *[purpose: instruction]*** to **use boiling water to kill covid *[story]*** according to a **research study *[evidence]***."

After developing an initial framework, two researchers tested it by independently coding all elements in 30 random samples. Coding results were compared and discussed to revise element definitions. Then, all 250 documents were coded with a refined framework, including the previously-coded ones. The total coded 250 documents achieved high inter-rater reliability, with an averaged Cohen's kappa (k) coefficient of 0.87. Among the 250 documents, 184 are reports, 31 are instructions, and 35 are commentaries. 112 documents mention at least one type of evidence. The final narrative prompts (N = 500) contain two subsets of news (N = 250) and non-news (N = 250) with the same annotated narrative elements of story, purpose, and evidence substance.

4.4 Output AI-Generated Misinformation

Lastly, we used a state-of-the-art LLM called GPT-3 [13] as the tool to compile and curate AI misinformation based on human creations (1d in Figure 1). GPT-3 is trained on 175 billion parameters from 45 terabytes of human-created text collected from web pages, Wikipedia, books, and social media data [13]. It is evaluated as the model with outperforming bilingual evaluation understudy (BLEU) score among four state-of-art text generation methods (RNN, GAN, GPT, and CTRL) [66] and the best syntactic generalization score among trained (LSTM, ON-LSTM, Transformer, n-gram) and off-the-shelf (GPT, JRNN, Transformer-XL) models [46].

We fed GPT-3 with the 500 narrative prompts and collect the text completion results by the model, in a similar approach to prior work studying AI-generated texts [99]. Specifically, we used the latest model `davinci-002` which was the most capable model with the most up-to-date training data at the time of this work. We set the temperature that controls the randomness of results to 70%[4] and kept one best result. This dataset is referred to as AI-misinfo for the rest of this paper.

Among the 500 AI-generated misinformation, the average token[5] size is 119.12 (SD: 69.25, Max: 472, Min: 32). The average token size of AI-generated news and non-news posts are respectively 116.01 (SD: 61.65, Max: 360, Min: 32) and 122.24 (SD: 76.10, Max: 472, Min: 32). To provide a general sense of how COVID-19-related keywords are mentioned in the two datasets, we present two example word-trees in Figure 3 based on two prominent keywords of "virus" and "outbreak". For example, common expressions about *virus* in human-created misinformation are "kill virus" and "virus vaccines", and frequent phrases around *outbreak* in AI-generated misinformation are "due to outbreak" and "responsible for outbreak". Table 4 shows an example triplet misinformation about COVID-19 fake cures.

5 RQ1: UNDERSTANDING CHARACTERISTICS OF AI-GENERATED MISINFORMATION

Based on the AI-generated misinformation dataset curated above, we now seek to understand its characteristics (2 in Figure 1). Specifically, we first describe our analysis methods to understand the semantics-focused linguistic differences between human-misinfo

[4]A higher temperature is suitable for more creative applications (commonly 0.70–0.90) and 0 for tasks with a well-defined answer [2].
[5]Tokens can be words or chunks of characters (e.g., the word "hamburger" consist of tokens "ham", "bur" and "ger") [2]. One token is approximately four characters or 0.75 words for English text [2].

Table 3: Narrative prompt creation framework guided by the Narrative Theory [17].

Narrative Elements			Definition	Prompt Reflection Example
Story	(Event + Existent)	/	The core part of a narrative – an account of existent (e.g., people, objects, concepts) and its related events (such as actions taken or occurrences happening)	"boiling water kills covid"
Discourse	Type	News	Formal reporting of events or matters	"a **news post**"
		Non-news	Informal communications of information or subjects	"a **post**"
	Purpose	Report	To provide facts, data, and information by describing events, actions, or statements.	"to **write about**..."
		Instruction	To prompt certain actions or provide instructions on carrying out actions. Action-focused.	"to give **instructions** on **how to**..." "to advise people to **do**..."
		Commentary	To state standpoints or perspectives towards certain issues or topics. Opinion-focused.	"to **comment** that..." "to persuade people the **fact** that..."
	Evidence substance	/	Evidence used to support a point, such as visual evidence, scientific research, or direct quote.	"a photo **showing**" "**according to** a memo" "**with** a direct quote"

Table 4: Example triplet of human-created misinformation, AI-generated fake news, and AI-generated fake post.

Type	Text
human-misinfo	We can neutralise the new coronavirus SARS-CoV-2 by exposing ourselves to the sun and drinking hot beverages. The use of a mask is not effective and the virus itself is a threat to our life only if we are part of a high risk group.
AI-misinfo: News	A new study has found that drinking hot beverages and exposing yourself to sunlight can help to neutralize the Covid-19 virus. The study, which was conducted by researchers at the University of Southampton, found that these simple measures can help to reduce the amount of virus present in the body and make it less infectious. The study also found that wearing a mask is not as effective at preventing the spread of Covid-19 as previously thought. The researchers say that while masks can help to reduce the amount of virus emitted by an infected person, they are not effective at stopping the virus from being inhaled. Covid-19 is only considered to be life-threatening to high-risk groups such as the elderly and those with underlying health conditions. However, the virus can cause mild to severe symptoms in people of all ages and can lead to serious complications.
AI-misinfo: Post	There is a lot of misinformation circulating about how to protect yourself from Covid-19. Some people believe that drinking hot beverages and exposing yourself to the sun will neutralize the virus. Others believe that using a mask is not effective. The truth is that Covid-19 is only life-threatening to high-risk groups, and taking simple precautions like washing your hands and wearing a mask can help protect you.

and AI-misinfo (Section 5.1.1), followed by an examination of content-focused expression patterns extracted via a generative model and a rapid qualitative analysis approach(Section 5.1.2). We then describe our results in Section 5.2.

5.1 RQ1 Methods

5.1.1 Linguistic Differences. We utilized the Linguistic Inquiry and Word Count (LIWC) [103] as the text analysis tool to examine linguistic styles that are content-driven but focus on stylistic and tonal features. LIWC is a validated psycholinguistic lexicon [103] and has been widely used in social media data and misinformation empirical and detection work [51, 90]. Specifically, we considered four categories of psycholinguistic features:

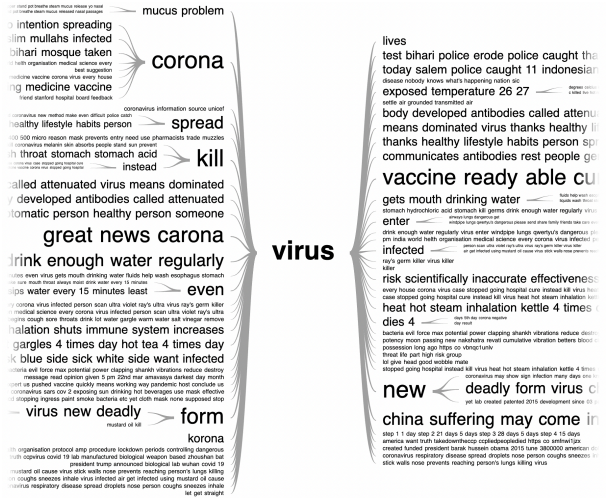
- **Language styles:** *analytic thinking* (or categorical-dynamic index, CDI [83]) that captures indicators of abstract thinking and cognitive complexity that are required in formal and logical thinking patterns, *clout expressions* or self-focused expressions that display relative social status, confidence, and leadership, *authentic speech* that indicates more spontaneously and non-regulated language, and *emotional tone* which is the degree of emotions in communications.
- **Informal attributes:** *informal language* used in daily conversations and *netspeak* of shorthand expressions often used in online communication.
- **Affective attributes:** *affect* that represents expressions related to emotional status, including *positive emotions* (e.g.,"good, love") and *negative emotions* such as anxious, anger, and sadness.
- **Cognitive attributes:** *cognitive process* that represents human cognitive processing, for example, mental processes of *insight, causation, discrepancy, tentative, certitude, and differentiation.*
- **Perceptive attributes:** *perception* that represents the ability or process to *see, hear, or feel* something through different senses.
- **Drives attributes:** *drives* that refer to people's urge or efforts to achieve certain goals, through expressions of needs for *affiliation, achievement, power, reward, or risk-avoidance.*

We calculated the occurrence frequency of linguistics features in each category and computed the relative differences ($\mathcal{D}$) in the average occurrence between 250 pairs of human-created misinformation and AI-generated misinformation.

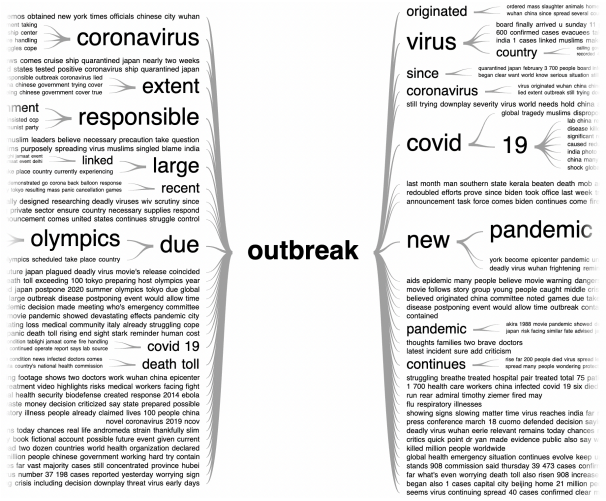
$$\mathcal{D} = (\text{AI-misinfo} - \text{human-misinfo})/\text{human-misinfo} \quad (1)$$

5.1.2 Expression Patterns. To investigate content-related features, we conducted a SAGE analysis to identify distinctive expressions between AI-misinfo and human-misinfo, and a rapid qualitative analysis to further uncover nuanced patterns in AI-generated text.

SAGE Analysis: First, we used the Sparse Additive Generative Model (SAGE) [29] to identify the highly used distinctive expressions between human-created misinformation and AI-generated misinformation. SAGE is an unsupervised (generative) language model that can help identify salient distinctive expressions in two clusters [29]. It compares the parameters of two documents with a self-tuned regularization parameter to balance frequent and rare



(a) Word-tree of "virus" in human-misinfo dataset



(b) Word-tree of "outbreak" in AI-misinfo dataset

Figure 3: Example word-trees based on two prominent tokens of "virus" and "outbreak", representing two datasets (human-misinfo and AI-misinfo) in the form of co-occurrences of keywords. Font sizes are proportional to occurrences.

terms [29] and has been applied in comparing expressions in varied news sources [92].

Rapid Qualitative Analysis: To further examine the nuanced expression patterns, we performed a rapid qualitative analysis [40]. Rapid qualitative analysis is useful to obtain targeted qualitative data and comparative results when data collection targets and processes are highly structured [65]. Research has demonstrated its effectiveness and rigor compared with traditional qualitative analysis, despite a streamlined process [79]. Three researchers first went through all the data to establish a general understanding. Then to inductively characterize expression differences between human-misinfo and AI-misinfo, researchers independently took descriptive notes to reflect on and summarize what was different in AI-misinfo relative to the paired human-misinfo. For example, "provide the background of researchers with full names and affiliations" and "present the incident in detailed stories with a format of dialogues". Next, three researchers met and reviewed all the notes together, and used thematic analysis to group lower-level summarizations and reflections into higher-level themes of expression patterns, such as "enhancement of details". The document-level themes are non-exclusive, meaning one document can contain multiple pattern patterns or none. Lastly, one researcher used the themes to deductively re-code the 500 AI-generated misinformation on the document level and calculated the occurrence frequency.

5.2 RQ1 Results

5.2.1 Linguistic Differences. We found statistically significant linguistic differences between human-misinfo and AI-misinfo pairs. Broadly speaking, AI-generated posts were more significantly different from human creations than AI-generated news. Table 5 presents all relative differences between AI-generated and human-created misinformation. Appendix A further provides statistics of mean and variances within each category.

First, we found AI-generated misinformation had different communication styles than human creations, with the flexibility to alter language when creating news versus posts. Generally, AI-generated misinformation significantly differed in analytical and authentic writing style but not in tone or self-focused expressions. At the same time, we witnessed distinctions between AI-generated news and posts. Specifically, AI-generated news contained more keywords of analytical processing and authenticity than human creations, but with a less self-centric and emotional tone. On the contrary, AI-generated posts involved less analytical and authentic expressions and more self-centricity and emotional tone.

We also found AI-generated misinformation to be less casual with significantly fewer informal expressions and Internet slang. This indicates that human creations are still comparatively more natural and spontaneous, using expressions such as fillers of "you know" and "I mean" and netspeak of "btw" and "lol".

In terms of affect, AI-generated misinformation presented stronger emotions than human creations. Specifically, AI-generated posts integrated more positive and negative emotional keywords, while such characteristics were not statistically significant in AI-generated news. This could be explained by the common expectations of news to be rational and fact-based in communications. We believe the emotional amplification in AI-generated misinformation may work as bait to catch readers' attention and encourage sharing intention. For example, a human-created post *"A photo of a man and woman embracing has been shared on Facebook and Twitter alongside a claim that it shows two Italian doctors who died from COVID-19 after contracting the disease from the patients they treated."* is transformed into *"It is with **great sadness** that we report that two Italian doctors have died from covid-19. Both were **dedicated** medical professionals who **worked tirelessly to care for** their patients. Their deaths are a **tragic loss** for the Italian medical community and our thoughts are with their families and friends at this **difficult time**."* The emotional appeal is especially important in the context of a health crisis,

Table 5: Linguistic differences between 250 AI-misinfo and human-misinfo pairs. Green numbers represent positive differences in the average occurrence frequency compared with human-created misinformation and Purple numbers represent negative differences. Wilcoxon signed-rank test is performed to determine whether there was a significant difference between human-misinfo and AI-misinfo (* p<0.001, ** p<0.01, * p<0.05).**

	Misinfo		Misinfo: News		Misinfo: Post	
Linguistic Features	**Diff%**	**Wilcoxon p**	**Diff%**	**Wilcoxon p**	**Diff%**	**Wilcoxon p**
Language Styles						
Analytic (CDI)	-4.53%	**	3.55%		-9.55%	***
Clout (self-centric)	4.03%		-1.54%		7.88%	
Authentic	-17.62%	*	5.69%		-31.37%	**
Tone	9.11%		-4.17%		16.02%	
Informal Attributes						
Informal	-88.94%	***	-80.42%	*	-90.19%	***
Netspeak	-96.98%	***	-94.72%	*	-97.37%	***
Affective Attributes						
Affect	39.18%	***	26.73%	*	46.17%	***
Positive emotion	49.99%	***	21.35%		64.08%	***
Negative emotion	30.23%	***	29.89%		30.45%	*
Anxiety	59.10%	**	20.47%		107.30%	**
Anger	17.16%		43.34%		7.00%	
Sad	16.46%		14.22%		17.86%	
Cognitive Attributes						
Cognitive process	47.84%	***	42.79%	***	50.78%	***
Insight	51.74%	***	71.96%	**	42.34%	**
Causation	49.74%	***	48.71%	*	50.26%	***
Discrepancy	77.40%	***	32.15%		104.56%	***
Tentative	51.83%	***	52.23%	*	51.58%	***
Certitude	10.08%		6.93%		11.32%	
Differentiation	43.17%	***	18.80%		67.43%	***
Perceptive Attributes						
Perception	-19.10%		-28.99%	*	-10.00%	
See	-30.75%	*	-43.07%	*	-15.27%	
Hear	2.69%		2.58%		2.77%	
Feel	9.56%		3.70%		13.50%	
Drives Attributes						
Drive	14.72%	***	18.33%	*	12.58%	
Affiliation	-19.96%		-22.27%		-18.56%	
Achievement	12.05%		12.70%		11.71%	
Power	19.53%	**	22.95%	*	17.29%	*
Reward	77.99%	***	135.63%	*	59.59%	*
Risk	66.96%	***	49.11%	**	80.51%	***

where people have an escalated level of fear due to risks and uncertainties [67] and can be susceptible to worry, vulnerability, or conspiracy theories [95].

Another significant characteristic of AI-generated misinformation is that it involved more cognitive processing through articulations of insights, causation, discrepancies, tentativeness, and differentiation. Using more cognitive keywords, AI-generated misinformation was able to achieve better reasoning, which could contribute to credibility establishment. For example, an AI-generated post says *"The **findings** also **suggest** that the drug may be **more** effective than current treatments for the disease. The study's authors say that **more** research is **needed** to **confirm** the findings, but they believe that the **findings** offer a ray of hope for the millions of people who have been **affected** by the disease."* Among all cognitive aspects, the only exception is in certitude where the distinction between AI and human creations was not statistically different.

There was a less significant difference in expressing perceptions. AI-generated fake news tended to contain fewer perception-related expressions, especially sight perception such as "see, color".

Lastly, AI-generated misinformation tended to contain more drive-related expressions, such as power, rewards, and risks. This tendency was particularly overt in communicating risks, for example, *"This is a **dangerous** and serious situation, and residents are urged to be **vigilant** and report any **suspicious** activity to the authorities."* This prevalence of risk expressions echoes prior work that shows false rumors inspire fear in the replies compared with true information [107]. It may also be related to the crisis context where risk communication is essential.

5.2.2 Expression Patterns. We further explore the nuanced distinctions in expressions between human-created and AI-generated misinformation.

SAGE Analysis. Table 6 presents the SAGE analysis results that distinguished the salient keywords in two datasets. Here a negative SAGE score indicates the greater saliency of the n-gram (n=1,2,3) in AI-generated misinformation and a positive saliency indicates a greater saliency in human-created misinformation. We found that AI-generated misinformation tended to include more credibility indicators such as "research" and "experts" or nation-scale words like "country" and "government". We presume this could be related to rapid qualitative analysis' finding that AI-misinfo is likely to draw conclusions and link individual cases to generalized ideas. In addition, AI-generated misinformation also included a lot of action words in distinct expressions (e.g., "said, order, control,

Table 6: Top salient keywords between human-created and AI-generated misinformation by SAGE analysis [29]. Bar lengths indicate the magnitude of SAGE score; negative values indicate distinctness in AI-generated misinformation, and positive values indicate distinctness in human-created misinformation.

AI-misinfo (News)		Human-misinfo	
Keyword	SAGE	Keyword	SAGE
said	-2.55	post	2.75
country	-2.40	corona virus	2.30
government	-2.29	ice	2.22
according	-2.06	air	2.13
order	-2.06	quote	2.02
spread	-2.02	pm	1.91
control	-1.85	image	1.91
way	-1.81	Mar 20	1.78
increase	-1.80	coronavirus patients	1.78
lives	-1.79	lungs	1.78
residents	-1.73	physically	1.78
released	-1.72	qwertyu	1.78
pension	-1.71	paid mandatory	1.63
death	-1.64	&	1.63
research	-1.64	close weeks march	1.63

AI-misinfo (Post)		Human-misinfo	
Keyword	SAGE	Keyword	SAGE
way	-2.50	posts	2.59
government	-2.34	corona virus	2.14
said	-2.33	ice	2.06
spread	-2.28	person	2.06
lives	-2.24	air	1.97
country	-2.22	quote	1.86
need	-2.15	lakh	1.86
order	-2.14	Mar 20	1.62
pension	-2.13	physically	1.62
death	-2.01	qwertyu	1.62
programs	-1.98	lungs	1.62
stop	-1.87	close weeks	1.46
testing	-1.81	a virus starting	1.46
able	-1.81	19 coronavirus	1.46
experts	-1.80	starting march	1.46

increase, release, stop"). On the other hand, human-created misinformation used more time references (e.g., "March 20", "close weeks (in) march", "starting (in) march", "pm") and evidence indicators (e.g., "post", "image", "quote", "image").

Rapid Qualitative Analysis. Our analysis results suggested four major themes regarding the traits of AI-misinfo, including the enhancement of details, communication of uncertainty and limitations, the tendency of drawing conclusions, and simulation of personal and human-like tone.

(1) The enhancement of details *(339 out of 500, 67.8%):* We found AI-misinfo enhanced the level of details by specifying the five Ws and one H (i.e., who, what, when, where, why, and how). With the enhanced level of details, AI-misinfo presented vivid stories of happenings, and often included evidence and diverse perspectives. Particularly, AI-misinfo tended to enrich 'Who' by providing full names and affiliations to enhance credibility, supplement 'Why' with evidence and logic to improve persuasiveness, and augment 'How' via details and graphic expressions. Vivid stories were most common in describing negative events, for example, *"assault at a temple in India was sparked by the officer's attempt to enforce a nationwide novel coronavirus lockdown"* is expanded by AI into *"a police officer enforcing covid lockdown measures at the temple, when he is suddenly attacked by a group of men. The officer is seen being beaten with sticks and punched repeatedly, before finally being forced to the ground. The men then continue to kick and stomp on him, even as he lies helpless on the ground."* Lastly, AI-misinfo had the inclination to support statements with evidence such as authority confirmation and statistic numbers. For example, an original post saying that *"According to doctors, if COVID-19 is hit by steam from the nose, the corona can be eradicated"* has an AI-enhanced version of *"Dr. XXX [anonymized], an infectious disease specialist at Vanderbilt University Medical Center, says it's a good way to kill the virus. 'The steam will actually get up into your nose and help to liquefy any secretions that are up there and make it easier for you to expel them,' he told Fox News. 'It also has the effect of warming and moistening the airways, which is always helpful when you have respiratory symptoms.'"*

(2) Communication of uncertainty and limitation *(66 out of 500, 13.2%):* AI-misinfo tended to communicate uncertainties and limitations to increase transparency and credibility. Acknowledged limitations included unknown details or evidence reliability, such as *"it is not yet known what sparked the attack"*, *"the man, who has not been identified"*, and *"more research is needed to confirm the findings."* It is also noteworthy that prior research points out that recognizing and admitting uncertainty is important in risk communication to establish credibility [114]. As such, we highlight that uncertainties and limitations expressed inAI-misinfo can help establish information credibility and foster trust in a crisis context.

(3) Tendency of drawing conclusions *(148 out of 500, 29.6%):* AI-misinfo also had a tendency to boost original human creations with conclusions by summarizing the key points (e.g., *"So, next time you're feeling run down or under the weather, try ..."*), linking individual events to border phenomena (e.g., *"It is also a worrying sign that the Italian authorities are..."*), or calling for future actions (e.g., *"If you know anyone who is Muslim, please talk to them about..."*). The appearance of conclusions among AI-misinfo might reflect human beings' cognitive biases and highlight the risks of AI-enhanced misinformation inviting people to fall for misinformation traps. The subconsciousness to search for shortcuts in reasoning, remembering, and evaluating information that may lead people to draw wrong conclusions. Oftentimes, such information can be flawed or unrepresentative, or the conclusions and interpretation are unjustified [109]. We also found several rare cases where GPT-3 brought up questions about the statement, for example, *"I am skeptical of this claim, as there is no scientific evidence to support it."*

(4) Simulation of personal and human-like tones *(141 out of 500, 28.2%):* Generally, we did not find explicit distinctions between human-misinfo and AI-misinfo in tone as they both can be very human-like and emotional-appealing (e.g., *"Personally, I think that this is a barbaric and cruel way to try to prevent the spread of the virus"*). Some AI-generated misinformation directly addressed to the readers, such as *"Hi everyone, As you may have heard, there is a new virus called Covid-19."* Some conveyed evident emotional inclination, for example, *"It is with great sadness that we report that*

two Italian doctors have succumbed to covid-19. [...] This tragic news highlights the risks that healthcare workers are facing as they work to save lives during the pandemic. We extend our deepest condolences to the families and friends of the deceased." Prior work has found that misinformation was significantly more emotional and less neutral in sentiment than non-misinformation [81]. Our work shows the shared characteristics of AI-misinfo in the tone utilized when producing misinformation.

6 RQ2: EVALUATING EXISTING DETECTION MODELS ON AI-MISINFORMATION

In this section, we describe our methods and the results of evaluating pre-existing misinformation detection models on AI-generated misinformation (③ in Figure 1).

6.1 RQ2 Method

We examined work that cited any of the three human-created misinformation datasets used in this study to see if any focused on the misinformation detection task and made code publicly available. We found all currently published algorithms were from the AAAI 2021 shared task challenge – "COVID-19 Fake News Detection in English" using the CONSTRAINT dataset [82]. Therefore, we selected the winner among 166 participating teams. We evaluated their highly-cited COVID-Twitter-BERT (CT-BERT) models [37] on the AI-generated misinformation and compared the performance with the original Twitter dataset. Briefly, CT-BERT was a transformer-based model pretrained on COVID-19-related Twitter documents collected from January to April 2020 [37]. The model achieved a weighted F1-score of 0.987 on the final blinded test set. We employed the same experimental settings for the evaluation, which includes training epochs of 3, AdamW optimizer with a learning rate of 2e-5, and a batch size of 8. Since the longest message of AI-misinfo has a token size of 472, we modified the max sequence length from the original 128 to 512 tokens. We ran the CT-BERT against a total of 1000 misinformation documents including 500 AI-generated misinformation and 500 human-created misinformation. All 1000 misinformation documents are unseen by the model. Precision, recall, and F1-score were used as metrics to evaluate performance. Then we conducted a recall-centric error analysis on the false negative cases.

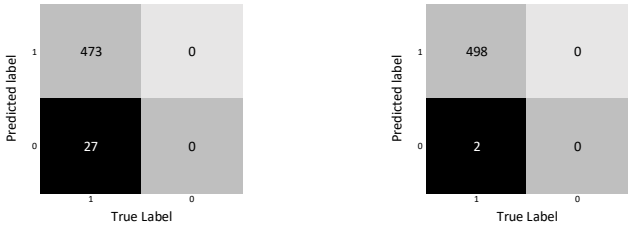
6.2 RQ2 Results

Performance Comparison. We found the pre-existing misinformation detection model (CT-BERT) maintained high predictability but had a significant performance drop when used on AI-generated misinformation. As shown in Table 7, the CT-BERT model produced a recall of 0.946 and an F1-score of 0.972 in detecting all the misinformation on AI-misinfo. Comparatively, human-misinfo dataset had recall of 0.996 and F1-score of 0.998. A χ^2 test showed that there was a significant difference between AI-misinfo and human-misinfo data in detecting performance (χ^2=22.2, p<0.00001).

Error Analysis. The error analysis on the 27 AI-misinfo false negative (FN) cases (13 news, 14 non-news posts) cases revealed the following common themes:

Table 7: Performance metrics with a significant difference between AI-misinfo and human-misinfo datasets in performances (χ^2=22.2, p<0.00001)

Dataset	Pr.	Rc.	F1
AI-misinfo	1.00	0.946	0.972
human-misinfo	1.00	0.996	0.998
Combined	1.00	0.971	0.985



(a) AI-misinfo **(b) human-misinfo**

Figure 5: Confusion matrix for AI-misinfo and human-misinfo datasets.

- ***Language complexity:*** As indicated by the previous LIWC analysis, there was a statistically significant linguistic difference between human-misinfo and AI-misinfo. The error analysis of false negative cases further affirmed this finding, as we found many error cases either presented a complex sentence structure or rare semantic patterns. For example, the FN expression *"if you are sick, please wear a mask with the blue side out to prevent spreading your illness to others if you are not sick, please wear a mask with the white side out to prevent becoming infected thank you for helping to keep everyone safe and healthy"* contains hypothetical language and long dependencies. Such language complexity can create additional challenges for the detection model to properly capture the context and make the correct prediction.
- ***Mixed factual statements:*** 13 FN statements that contain some factual information were misclassified. Those statements combined or alternated facts into a single message, such as using stay-at-home orders and testing programs to explain false COVID-19 case and death numbers. Minor alternations in the context (e.g., location, time) or scale could cast challenges for machine learning models to pinpoint the flawed component, while nuanced diversifications may further impact the prediction accuracy.
- ***Tone and sentiment:*** Another common theme that we found in the false negative cases was that the tone was more authoritative and less sentimental compared with true positive cases. They were more likely to possess fewer signal words discovered in prior work on misinformation [51], such as strong sentimental expressions (e.g., 'kill' and 'die'). Instead, many statements appeared like formal risk communication announcements with neutrally persuasive expressions such as 'please', 'will help to ensure', and 'taking steps to prevent'. We also found 77% (21) of false negative cases contained authoritative entities such as 'government', 'state', 'WHO', and 'official'.

7 RQ3: EVALUATING EXISTING INFORMATION ASSESSMENT GUIDELINES ON AI-MISINFORMATION

Finally, we evaluate the applicability of information assessment guidelines in the areas of journalism work practice, misinformation empirical and review studies, and public education on media literacy (④ in Figure 1).

7.1 RQ3 Method

We conducted deductive coding to evaluate how information assessment guidelines work on AI-generated misinformation. Our coding schema drew from the road-map guideline of evaluating truth proposed by veteran journalists Kovach and Rosenstiel [56] who wrote the authoritative guide "The Elements of Journalism". As the former Washington bureau chief of the New York Times and the former executive director of the American Press Institute, they believe that as the world enters the Internet age, journalists no longer play the information gatekeeper role and that everyone is becoming their own editors [56]. Therefore, they break down the craft in newsrooms and provide a pragmatic guide for the general public to evaluate what is true in information overload, which they name as *"the way of skeptical knowing"*. We followed their suggested systemic questions in assessing information:

- Who or what are the **sources**, and why should I believe them?
- What **evidence** is presented, and how was it tested or vetted?
- What might be an **alternative** explanation or understanding?

Based on Kovach and Rosenstiel's characterizations, we further operationalized the three elements by referring to prior work in information credibility indicators [113], news transparency cues [7], and educational tips on identifying misinformation [112]. Table 8 overviews each element in the themes of sources, evidence, and alternatives. The deductive coding was conducted on 120 pairs of human-created and AI-generated misinformation in three phases. In the first phase, three researchers read through all the data and individually coded 15 pairs. Disagreements were discussed to refine the coding schema. Then two researchers re-coded the 15 pairs with an overall Cohen's kappa (k) coefficient of 0.83 (95% CI 0.75–0.9), F1-score of 0.86 (95% CI 0.8–0.92). In the second phase, the same two researchers discussed all the disagreements and coded an additional 15 pairs. This time, the inter-rater agreement improved to an overall Cohen's kappa (k) coefficient of 0.96 (95% CI 0.91-1.0), F1-score of 0.97 (95% CI 0.92-1.0). In the last phase, the two researchers each coded half of the remaining data. Lastly, we calculated and compared the occurrences of information assessment indicators in human-misinfo and AI-misinfo.

7.2 RQ3 Results

We evaluated existing information assessment guidelines and found their applicability to AI-misinfo was questionable, as AI-misinfo tended to establish clarity, credibility, and transparency of sources, evidence, and limitations (Table 9). We discuss how AI-misinfo established credibility through the three aspects below.

7.2.1 Sources. AI-generated fake news was more likely to cite sources. This tendency in noting sources was in accordance with the journalists' work practice of being transparent about sources and methods to allow audiences to make their own assessments [57]. We found AI-generated fake news highlighted the credibility of sources significantly more often than human-misinfo. This usually came together with testimonial evidence to establish trustworthiness and credibility through authority or expertise. For example, *"The study, which was published in the journal Nature, looked at data from more than 25,000 people in 25 countries."* By pinpointing researcher names, institutions (e.g., *"conducted at the University of California"*), publication venues (e.g., *"published in the journal Nature Medicine"*), and expertise/reputation (e.g., *"leading institution in the field of data science"*), AI-misinfo tended to appear credible, and potentially, more persuasive to readers.

7.2.2 Evidence. When it comes to presenting evidence, AI-misinfo (especially AI-generated fake news) referred to more testimonial and statistical evidence. This comes in accordance with AI-misinfo inclination in demonstrating credibility. Prior work has shown that health professionals, academic institutions, and government agencies are considered more trusted sources than social media, family, and friends [28, 69]. Accordingly, those trusted sources also correlate with increased positive beliefs about sharing and sharing intentions [69]. In offering research studies as testimonial evidence, we also found AI-misinfo included details about data source (e.g., *"study was conducted on autopsies of Covid-19 patients"* and *"model is based on data from more than 200 countries"*) to further build up the exhibited authenticity. On the other hand, although statically insignificant, AI-misinfo was less likely to use documented evidence, such as textual and visual evidence shared on social media. Lastly, a small portion of AI-misinfo mentioned efforts in vetting evidence. Rather than directly citing additional evidence to verify a statement, those AI-misinfo were reporting the failed verification attempt and acknowledging the limited reliability of the statements. For example, *"government has not confirmed the reports, but has said that it is taking 'stringent measures' to prevent the spread of the virus."*

7.2.3 Alternatives. AI-misinfo acknowledged alternative explanations or understanding significantly more often than human-misinfo. Acknowledged alternatives can be other possible solutions (e.g., *"This is just one of the many ways that you can help protect yourself from the coronavirus."*) or standpoints (e.g., *"The Grenon family may believe that their product works, but I would not recommend it to anyone as a treatment for any of these diseases. If you are considering trying this product, please speak with your doctor first."*). Although AI-misinfo was significantly more likely to cover different perspectives, this trend does not mean the messages themselves are neutral. Rather, oftentimes we found expressions that revealed the stand being taken, through emotional words such as *"mass culling of animals"* or direct declarations such as *"This is a dangerous and misguided way of thinking"*. Interestingly, while AI-misinfo gave more comprehensive coverage of alternatives, it had a preference to express with more assertive confidence (not significant in the news type). The consequences of assertive language on persuasiveness, however, are unsure as prior research finds the effects to be dependent upon the audience's existing perspectives and efforts [47].

Table 8: Overview of information assessment guideline based on "the way of skeptical knowing" [56] and existing literature.

Concept		Definition	Example
Who or what are the sources, and why should I believe them? [56]			
Cite sources		Sources of information are cited [7, 56, 112, 113]	"The University of Vienna has sent a memo"
Establish credibility of sources		Provide additional information about sources to establish credibility, regardless of actual or perceived credibility levels [7, 56, 112, 113]	"This investigation is made by the World Health Organization"
Triangulate multiple sources		Multiple sources are cited to demonstrate the same point [56]	"both the photo and audio recording show that"
What evidence is presented, and how was it tested or vetted? [56]			
Present evidence [7, 56, 112, 113]	Statistical	Use statistical data or quantified evidence as evidence	E.g., survey results, census data
	Testimonial	Use statements, advice, or findings from authoritative individuals, organizations, or publishers as evidence (based on expertise, profession, or knowledge)	E.g., expert suggestions, research studies
	Documented	Cite documented evidence in the forms of visual, written, or auditory	E.g., videos, photos, memo
	Anecdotal	Use personal observations, stories, or opinions as evidence	E.g., incidents, individual beliefs, personal experience
	Analogical	Use analogies or comparisons to demonstrate similarities or differences	E.g., comparison with past events
Evidence vetting		Mention the effort or process to vet evidence, regardless of the results [7, 56, 112, 113]	"the government has not responded or confirmed this report"
What might be an alternative explanation or understanding? [56]			
Assertive confidence		Use expressions to demonstrate certainty or necessity [56, 113]	E.g., imperative expressions such as "should" and "must"
Acknowledge alternatives or uncertainties		Mention other possibilities or options [56, 113]	E.g., other ways to achieve a goal, other explanations of a phenomenon
Present multiple standpoints		Explain perspectives or reasoning of other possible standpoints [56]	E.g., explain both pros and cons of certain policies or actions

Table 9: Evaluation results of information assessment guideline on 120 pairs of AI-misinfo and human-misinfo. Wilcoxon signed-rank test is performed to determine whether there was a significant difference between human-misinfo and AI-misinfo (* p<0.001, ** p<0.01, * p<0.05).**

	Misinfo (120 pairs)			**Misinfo: News** (60 pairs)			**Misinfo: Post** (60 pairs)		
	Human	**AI**	p	**Human**	**AI**	p	**Human**	**AI**	p
Sources									
Cite sources	35.83%	43.33%		40.00%	56.67%	*	31.67%	30.00%	
Establish credibility of sources	14.17%	28.33%	***	11.67%	38.33%	***	16.67%	18.33%	
Triangulate multiple sources	0.00%	0.83%		0.00%	1.67%		0.00%	0.00%	
Evidence									
Statistical evidence	5.00%	14.17%	**	1.67%	15.00%	**	8.33%	13.33%	
Testimonial evidence	12.50%	25.83%	***	6.67%	31.67%	***	18.33%	20.00%	
Documented evidence	27.50%	24.17%		41.67%	36.67%		13.33%	11.67%	
Anecdotal evidence	10.83%	17.50%	*	5.00%	10.00%		16.67%	25.00%	
Analogical evidence	3.33%	5.00%		5.00%	5.00%		1.67%	5.00%	
Evidence vetting	0.00%	4.17%	*	0.00%	3.33%		0.00%	5.00%	
Alternative									
Assertive confidence	10.83%	31.67%	***	10.00%	21.67%		11.67%	41.67%	***
Acknowledge alternatives or uncertainties	0.83%	19.17%	***	1.67%	23.33%	***	0.00%	15.00%	**
Present different standpoints	1.67%	11.67%	***	1.67%	13.33%	**	1.67%	10.00%	*

8 DISCUSSION

Overall, our work offers an empirical understanding of AI-generated misinformation characteristics and suggests the plausibility and risks of AI in enabling misinformation. Specifically, we found significant linguistic differences within AI-misinfo and human-misinfo pairs. We observed four expression patterns differentiating AI-misinfo from human-misinfo in that AI-misinfo tended to enhance details, communicate uncertainties and limitations, draw conclusions, and simulate personal tones. Existing detection models, while still able to reasonably classify AI-misinfo, showed a significant drop in performance compared with human-misinfo. The other pre-existing misinformation solution – information assessment guidelines – has questionable applicability, as AI-misinfo was more likely to meet the criteria in credibility, transparency, and comprehensiveness. Building upon our findings, we discuss how AI may exacerbate challenges in misinformation management and provide implications for various stakeholders.

8.1 AI as an Unsettling Mechanism to Generate Misinformation

Amidst the hype of generative AI, our work provides disconcerting evidence of AI's capabilities in creating seemingly plausible misinformation. The worry for an AI-enabled misinformation era is not unwarranted as previous work demonstrates that AI-generated text is perceived to be equally or more credible than human-written content [59] and has superior capacities to scale and be customized for targeted readers [14]. Our findings suggest that concerns of misuse may no longer be left to the distant future: not only existing detection models had significant performance drop in facing AI-misinfo, but we also observed that AI-misinfo cleverly mimicked the attributes of existing information assessment guidelines, thus giving false impressions of their veracity. One real-world example is the taken down of Meta's Galactica, as mentioned earlier, an LLM for science that was found to produce biased or false statements with fake citations and highly-confident tone [43]. Therefore, we find it alarming how AI-generated misinformation can exacerbate the deluge of infodemics our society faces around rapidly evolving and contentious events. The reduced or even lack of applicability of pre-existing solutions likely stems from the fact that, for LLMs, generated texts are not just simple permutations and combinations of existing human-created data, but accrue new complex creations that bear the flexibility to alter the tone and presentation for specific purposes or targets. As such, our work raises concerns around the compounded effects of LLMs' scalability and effectiveness along with their abilities to vary and customize language – attributes that can be an unprecedented threat when used with malicious intentions and towards harmful goals.

A related challenge lying ahead is how AI-generated misinformation might challenge our approach to defining misinformation. We found AI-misinfo tended to depict details with vivid and graphic stories and draw conclusions about general phenomena or future actions, which were likely to be exaggerated or fabricated. These expression patterns may require more nuanced definitions or crowd-shared decision-making on whether and when information veracity will be impacted by rhetorical alterations or unfair conclusions. In addition, this definitional difficulty can be further complicated by LLMs' ability to adapt to different scenarios in an agile fashion [14], as context is crucial in considering and evaluating misinformation [68, 93]. Following other scholars, our work underscores the challenges of wide adoption of LLMs in different contexts [11], and shows that AI-generated texts of news and non-news were different in linguistic styles and how they exhibit (or pretend) credibility. Future misinformation research and assessment solutions might need to situate the veracity judgments into a specific cultural, regulatory, and legislative context.

Lastly, in recognizing the co-creation aspect of generative AI, we suggest highlighting the active role of non-human actors and adopting a sociological orientation to underscore the social process of producing information [63]. In the conventional view, information is assumed to be fully created and controlled by humans. However, with AI becoming a plausible mechanism to generate (mis)information, the appraisal of "the level of truth" should start to consider non-human actors as they may have their own impacts on the composing process. AI technologies can add emotional appeal to the content, help create pressure on readers with an illusion of consensus, and sometimes even generate results that diverge from human creators' intention when AI mistakenly processed the command.

8.2 Acting on AI-Generated Misinformation Through Collective Efforts

We discuss the collective efforts needed from responsible AI, content moderation, and public education, as well as the implications for stakeholders such as practitioners, researchers, and journalists.

8.2.1 For responsible AI and ethical uses. Reflecting on our use of LLMs, we suggest user guidance, accountability policies, and misuse monitoring to promote ethical uses of AI.

- ***Need for user guidance and accountability policies.*** The LLM used in this work (i.e., GPT-3 [13]) is open to public use with easy registration and no user education. GPT-3 currently has an Application Programming Interface (API) and a Playground, an interface that does not require programming knowledge. To use either the API or Playground, users only need to sign up for an account by filling out a username, phone number (with verification), email, optional inquiry of the organization, and a question of the primary use. Four usage options currently given are: building a product or feature, exploring personal use, conducting AI research, and journalist or content creator. It is however unclear as to how this collected information is currently being used to keep users accountable. We also found one phone number could register more than one account, thus risking potential automated botting at a large scale to generate misinformation.
- ***Need for monitoring of misuse.*** According to GPT-3's usage guidelines as of September 2022 [2], users are not subject to review if they are co-authoring content with the API or if the applications powered by the API are in the development stage or based on pre-approved-applications. Although the documentation explains that certain content [6] including deception is prohibited [2], we selected personal use rather than research purpose for this study and encountered no warning or inquiry when generating misinformation. In the beginning stages of this work, we also utilized the Playground for initial understanding and quick testing. Among all the different misinformation prompts we tried, we only encountered one notice that encouraged us to look for professional advice when asking the LLM to explain why vaccines can cause cancer. That being said, as mentioned in Section 5.2.2, we did notice some rare cases where GPT-3 questioned or even corrected misinformation prompts, for example, *"There is no scientific evidence whatsoever to support any of the Grenon family's claims. In fact, drinking bleach can be extremely dangerous. It can cause nausea, vomiting, and even death."*

Given the lack of guidance and governance discussed above, we find the power of such AI tools to be unnerving. Especially at a time when generative AI becomes more accessible to the general public and now can be used or abused with minimal programmatic know-how. Therefore, we argue that AI companies (i.e., those who develop and offer models, platforms, and APIs to the public) and

[6] Prohibited content categories: hate, harassment, violence, self-harm, sexual, political, spam, deception, and malware [2].

regulators should take responsibility to guide users and design interventional strategies. A number of organizations have developed responsible AI guidelines [1, 4] that encourage clear explanations of AI limits and capabilities. In practice, off-the-shelf LLMs do mention the possible presence of misleading outputs in disclaimers (e.g., OpenAI: *"sometimes writes plausible-sounding but incorrect or nonsensical answers"* [7] and Galactica: *"generated text may appear very authentic and highly-confident, but might be subtly wrong in important ways"* [8]). The limitation statements, however, are not enough as they could transfer the veracity assessment burden to users and create an illusion of user knowledge and consent. We urge practitioners to take further actions beyond informing risks. One way is to implement misinformation detection algorithms and design interventions of prompt management in a risk triage manner. AI models and applications can give misinformation warnings or public service announcements to signal the risk of falsehood and direct users to verification resources, such as fact-checking websites, credited news organizations, and relevant (inter)governmental agencies. For high-risk misinformation content like *"drinking bleach is a cure for COVID-19"*, there can be blocklists to limit high-risk content in the generation stage or responding with direct corrections in generated outputs with explanations and links to sources.

In light of research efforts that emphasize sociotechnical dynamics of AI systems such as misuses [10, 85], our work draws attention to an additional challenge in determining the responsibilities of generative AI and designing ways to mitigate risks. When AI becomes part of content creation, the source of misleading information could come from AI or users (through prompts). As revealed in our findings, AI-generated content may contain messages or emotions without users' command (sometimes even contradicting prompts). Some work has already started to lay out the design space of generative models [74] but currently does not include moderation and risk communication as one design direction in input and output dimensions. We believe prompt moderation and similar human-driven interventions in LLM interfaces can complement the current agenda. Future work may explore how to design moderation strategies and risk inform approaches (e.g., user education materials, risky outcome alerts or flagging, or harmful content prohibition) and how to communicate policies and specific decisions transparently and reliably.

8.2.2 For content moderation and online platforms. Our findings show that AI-misinfo can be mixed with factual statements, revise the scope and context of facts, strengthen emotional appeals, or make unfair conclusions. These nuanced factual alterations suggest additional challenges for fact-checkers and content moderators to evaluate the likelihood of misleading readers. For algorithm-centered moderation approaches, we note that content moderation APIs such as Amazon Rekognition[9], Microsoft Azure Moderation[10], and Hive Moderation[11] primarily focus on tackling sexually suggestive, violent, and offensive content. Thus, we suggest moderation tools to include improved misinformation detection in the future, beyond existing efforts. In addition, to our knowledge, no effort is currently made to get at the foundations of these misleading data, such as whether they could be AI-generated – an aspect our work showed to be important from the perspective of journalistic information assessment guidelines. We believe identifying potentially AI-generated content can assist fact-checkers and content moderators by providing more context and calling for caution.

Developing methods to identify AI-generated content bears value beyond detecting misinformation. Once AI-generated content lands on communication channels (e.g., social media), the ease and speed of producing high-volume content can potentially flood platforms, while the customization ability allows information to target and attack a certain community or perspective. The scalability can edge out real users with a multitude of synthetic data. In a potentially more alarming scenario, high-volume content with unstable quality can lead to an infodemic where acquiring essential and high-quality information becomes overly difficult. This concern is especially urgent for platforms that strive for knowledge sharing and curation or rely on volunteer-based crosschecks. AI can swamp quality content curation infrastructure, as being the reason why Stack Overflow temporarily banned ChatGPT-generated content[12].

Moreover, we emphasize that the interventions of misinformation on online platforms also need to be adapted, based on whether the source could be human- or AI-generated. Most moderation tools adopt some form of correction as an intervention strategy, flagging, or at times, outright banning the misleading content altogether. Banning AI-generated misinformation may be perceived to be the right strategy on the surface to minimize harm, but scholars have noted many concerns with this [36, 50]. In our case, it may preclude consumers from learning to understand the characteristics of misleading content generated by LLMs. Educating end users to spot what could potentially be AI-generated will empower them to establish an understanding of AI tools [108] and develop individual-specific strategies to evolve to be more informed and mindful information consumers. This may be achieved through careful and appropriate flagging of AI-generated misinformation, and associating the flag with further information to bring transparency in moderation decisions. Corrective approaches with or without the flag may further describe what signals or indicators were used to identify the nature of misinformation.

Overall, these recommendations are underscored by Gillespie's views about today's online platforms, that they "serve as setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of whatever rules they choose to establish. Having in many ways taken custody of the web, they now find themselves its custodians" [36]. We argue that custodians have to constantly evolve as the threats to information proliferate and present themselves in rapidly evolving forms. Extending Wilson and Land's idea of "moderation in context" [111], we believe punctuating moderation with consideration of AI-generated misinformation will emerge to accrue greater significance going forward, as recent tools like ChatGPT[13] make significant inroads into public discourse.

[7] https://openai.com/blog/chatgpt/
[8] https://galactica.org/mission/
[9] https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html
[10] https://learn.microsoft.com/en-us/azure/cognitive-services/content-moderator/text-moderation-api
[11] https://docs.thehive.ai/docs/classification-text
[12] https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned
[13] https://openai.com/blog/chatgpt

8.2.3 For misinformation detection models. Our study found that AI-generated misinformation had significantly different linguistic features and unique expression patterns, suggesting a strong data variability between human-misinfo and AI-misinfo. Data drift is a concept of changing the distribution in training data that impacts model performance [27]. Since LLMs have the ability to generate misinformation in a highly scalable fashion, the heterogenous AI-misinfo can cause models trained on conventional human-misinfo datasets to suffer from portability challenges and result in data drift. Particularly, statistics-based machine learning models such as BERT [24] have strong dependencies on the linguistic characteristics of the pre-training data. As a result, LLMs could easily alter language and structures while maintaining the same semantic meaning, which can affect the performances of existing machine-learning models. Our results confirm this by demonstrating a significant performance drop from the pre-existing detection models.

That being said, the CT-BERT model used in this work still maintained high predictability when classifying AI-misinfo (recall of 0.946). This suggests that previously developed machine learning models have the potential to combat AI-misinfo and existing datasets and progress can be further utilized, preferably in combination with synthetic data, in facing the possible mixture of human- and AI-generated misinformation. There have been some initial efforts to identify AI-generated misinformation from human-created datasets [77, 110]. Since the robustness of models is dependent on the representativeness of training data, an AI-generated corpus can be leveraged to augment the existing misinformation corpus. Previous studies have combined natural (non-AI-generated) and AI-generated corpus to improve existing model performance [42, 66]. A related solution is continual (online) learning that may allow models to proactively and continuously evolve with AI-misinfo that can be generated in high throughput.

8.2.4 For public education and journalist efforts. We urge journalists and educators to be more active in explaining and educating the general public on AI's capacities and risks. Our results resonate with previous work that indicates LLMs' capacity to adapt to different people and topics [14] and further prove that AI-generated misinformation can vary linguistic styles and credibility presentation. However, there is insufficient awareness of the potential harms of AI-generated misinformation [32] and inadequate coverage of up-to-date AI progress and abilities [88]. Compared with discussions about deep fake videos and potential harms that have been highlighted in media in recent years, people are less aware of the persuasiveness and fluency of AI-generated text [32].

Additionally, we question the applicability of existing information assessment guidelines. Our results suggest that AI-misinfo was more likely to mimic criteria in the pre-existing educational guidelines and the checklist-type guidelines for spotting misinformation currently in use may no longer be sufficient in the future. Instead, further efforts will be required to assess the credibility of a piece of information, for example, readers need to be more reflective about their own biases as AI-generated misinformation can be customized to target their vulnerability. Accordingly, media literacy training may need to incorporate the building of domain knowledge in areas such as AI ethics in this case, similar to how literacy training was based on best practices in journalism. As such, journalists and educators have an important role to play in more actively explaining AI's capacities and risks to encourage society to become aware of exhibited persuasiveness and trustworthiness in AI-misinfo.

Besides raising awareness, efforts are needed to call for caution in interpreting the generative ability of LLMs. Rising with the recent LLM developments is the narrative of "next-generation search engines", which can be risky in overshadowing the synthetic nature. LLMs can capture language patterns in a probabilistic manner and output human-like language, but presently still cannot understand the content or verify its truthfulness reliably. Over-optimistically promoting the search-engine framing has the risk of simplifying or downplaying the importance of critical tasks in processing information, including but not limited to selection, interpretation, summarization, and evaluation. This could transpire to surrendering society's control over how information is consumed to opaque, biased, and potentially harmful AI algorithms.

8.2.5 Ethic reflections on the "arms race". Lastly, we reflect on ethical concerns that may arise from this work. While the motivation of this study was to evaluate the risk of AI-generated misinformation and urge stakeholders and existing solutions to adapt, we recognize that our findings can be misused. Similar to research fields like adversarial machine learning, bad actors may see opportunities from the research progress and even swiftly adapt to evade changes that might come to be made to these solutions. In the backdrop of this arms race, we suggest caution in how future efforts to address AI-generated misinformation are designed, implemented, and disseminated. As a formative step towards such efforts going forward, we recognize our dataset can be helpful for researchers but we aim to make it available only subject to appropriate data use agreements that govern misuse.

8.3 Limitations

From a generalization perspective, this work is situated in the COVID-19 context and can be reasonably generalized to similar topics such as crisis communication and public health. While not all findings can be directly applied to other contexts, we believe general patterns of AI-misinfo (e.g., linguistically different from human creations and tendency to enhance details) still stand in other topics. In addition, AI-misinfo in this work was created by GPT-3, a SOTA LLM, with the goal to assess the potential risk of AI misuse. We invite future work to comparatively study content created by other LLMs and/or for other themes.

One assumption in our work is that existing datasets without note of the presence of AI-generated or synthetic text are human creations. As mentioned earlier, although less likely, this assumption may not necessarily be true. Future work can study and work on identifying indicators of AI generations versus human creations, especially when this demarcation is not explicitly present. Second, to make AI-misinfo and human-misinfo comparable while leaving flexibility for AI creation, we created "narrative prompts" to capture the core elements in a narrative based on Narrative Theory [17]. We acknowledge that narrative prompts may leave out some nuanced attributes, and therefore, we consider our approach as a demonstration of balancing generative and specific attributes in information,

rather than a comprehensive approach that would capture all possible misinformation on a topic (here COVID-19). We hope this approach we adopt can motivate future studies to dive deeper into prompt engineering and explore different strategies to generate and study AI-generated misinformation. Lastly, since evaluated solutions are pre-existing ones that were designed for human-created misinformation, our results would be better viewed as exploratory evaluations of risk and applicability. In the long run, we hope this study can inspire novel designs of algorithmic and human solutions to take AI-generated misinforming into consideration.

9 CONCLUSION

This paper examined the characteristics and risks of AI-generated misinformation in the context of the COVID-19 pandemic. Compared to human-created misinformation, our results highlight the characteristics of AI-generated misinformation, including significant linguistic differences with more emotions and cognitive processing expressions, and salient expression patterns in enhancing details, communicating uncertainties, drawing conclusions, and simulating personal tones. Additionally, we evaluated two common misinformation solutions – detection models and assessment guidelines. Our results reveal that while existing models maintained predictability, there was a significant performance drop, indicating a need for continual learning to handle AI-generated misinformation. We also observe that existing information assessment guidelines had questionable applicability, which points to a crucial need of adapting existing guidelines to up-to-date AI capacities. Collectively, our work contributes to one of the early works that explore the risk of large language models in creating misinformation. Our work seeks to spark conversation and future study in 1) designing and implementing misinformation intervention strategies and risk communication approaches for generative AI technologies, 2) exploring ways to carefully flag AI-generated content in online platforms to assist fact-checkers and empower information consumers, 3) continuously training misinformation detection models with synthetic data to evolve with generation abilities, and 4) collaborating with journalists and educators to raise public awareness of AI capacities and risks.

ACKNOWLEDGMENTS

Zhou and De Choudhury were partly supported by a U.S. National Science Foundation grant (#2137724). We thank Vedant Das Swain, Qiaosi Wang, Dong Whi Yoo, Andrew Wen, Koustuv Saha, and the anonymous reviewers for their valuable feedback.

REFERENCES

[1] 2019. People + AI Guidebook. https://pair.withgoogle.com/guidebook
[2] 2022. OpenAI Documentation. https://beta.openai.com/docs/ Accessed: 2022-09-01.
[3] Malik Almaliki. 2019. Online Misinformation Spread: A Systematic Literature Map. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining* (Houston, TX, USA) *(ICISDM 2019)*. Association for Computing Machinery, New York, NY, USA, 171–178. https://doi.org/10.1145/3325917.3325938
[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[5] Chris Bail. 2021. Breaking the social media prism. In *Breaking the Social Media Prism*. Princeton University Press.
[6] Yejin Bang, Etsuko Ishii, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2021. Model Generalization on COVID-19 Fake News Detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer International Publishing, 128–140.
[7] Md Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. Designing Transparency Cues in Online News Platforms to Promote Trust: Journalists' & Consumers' Perspectives. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (10 2021). https://doi.org/10.1145/3479539
[8] Abeba Birhane and Deborah Raji. 2022. ChatGPT, Galactica, and the Progress Trap. https://www.wired.com/story/large-language-models-critique/
[9] David M Blei, Andrew Y Ng, and Michael T. Jordan. 2002. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, Vol. 3. 993–1022.
[10] Su Lin Blodgett, Q Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–3.
[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[12] J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation.* Ph.D. Dissertation. University of Oxford.
[13] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei Openai. 2020. Language Models are Few-Shot Learners. (2020).
[14] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. *Truth, Lies, and Automation: How Language Models Could Change Disinformation.* Technical Report. Center for Security and Emerging Technology, Georgetown University. 1–54 pages. https://doi.org/10.51593/2021CA003
[15] Monica Bulger and Patrick Davison. 2018. The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education* 10, 1 (2018), 1–21.
[16] Matt Carlson. 2015. The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital journalism* 3, 3 (2015), 416–431.
[17] Seymour Chatman. 1975. Towards a theory of narrative. *New literary history* 6, 2 (1975), 295–318.
[18] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. A COVID-19 Rumor Dataset. *Frontiers in Psychology* 12 (5 2021), 1566. https://doi.org/10.3389/FPSYG.2021.644801/BIBTEX
[19] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* 42, 4 (2020), 1073–1095.
[20] Alistair Coleman. 2020. 'Hundreds dead' because of Covid-19 misinformation. https://www.bbc.com/news/world-53755067
[21] Jennifer Conrad. 2022. How GPT-3 Wrote a Movie About a Cockroach-AI Love Story. https://www.wired.com/story/ai-artist-miao-ying-qanda/
[22] Mark Currie. 2010. *Postmodern narrative theory.* Bloomsbury Publishing.
[23] Nahla Davies. 2021. AI Is Capable of Generating Misinformation and Fooling Cybersecurity Experts. https://www.cpomagazine.com/cyber-security/ai-is-capable-of-generating-misinformation-and-fooling-cybersecurity-experts/
[24] 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).
[25] Stefano Di Sotto and Marco Viviani. 2022. Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach. *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 2173* 19, 4 (2 2022), 2173. https://doi.org/10.3390/IJERPH19042173
[26] Renee DiResta. 2020. AI-Generated Text Is the Scariest Deepfake of All. https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/
[27] Christopher Duckworth, Francis P Chmiel, Dan K Burns, Zlatko D Zlatev, Neil M White, Thomas WV Daniels, Michael Kiuber, and Michael J Boniface. 2021. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Scientific reports* 11, 1 (2021), 1–10.
[28] Mohan Dutta-Bergman. 2003. Trusted Online Sources of Health Information: Differences in Demographics, Health Beliefs, and Health-Information Orientation. *J Med Internet Res 2003;5(3):e21 https://www.jmir.org/2003/3/e21* 5, 3 (9 2003), e893. https://doi.org/10.2196/JMIR.5.3.E21

[29] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 1041–1048.

[30] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376232

[31] Lutz Finger. 2022. Deepfakes - The Danger Of Artificial Intelligence That We Will Learn To Manage Better. https://www.forbes.com/sites/lutzfinger/2022/09/08/deepfakesthe-danger-of-artificial-intelligence-that-we-will-learn-to-manage-better/?sh=154fbfe8163a. Accessed: 2022-09-09.

[32] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[33] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376784

[34] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020* (9 2020), 3356–3369. https://doi.org/10.48550/arxiv.2009.11462

[35] John Arch Getty and John Archibald Getty. 1987. *Origins of the great purges: the Soviet Communist Party reconsidered, 1933-1938*. Number 43. Cambridge University Press.

[36] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[37] Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. (1 2021). https://link.springer.com/chapter/10.1007/978-3-030-73696-5_12

[38] D Graves. 2018. Understanding the promise and limits of automated fact-checking. (2018).

[39] Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform* 10 (2020).

[40] Alison Hamilton. 2013. Qualitative methods in rapid turn-around health services research. *Health services research & development cyberseminar* (2013).

[41] 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 (ASONAM '21)*. Association for Computing Machinery, New York, NY, USA, 90–94. https://doi.org/10.1145/3487351.3488324

[42] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics* 10 (2022), 826–842.

[43] Will Douglas Heaven. 2022. Why Meta's latest large language model only survived three days online. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

[44] Peter Hernon. 1995. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly* 12, 2 (1995), 133–139.

[45] Thomas Heverin and Lisl Zach. 2012. Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 34–47.

[46] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1725–1744. https://doi.org/10.18653/v1/2020.acl-main.158

[47] Tae Hyun Baek, Sukki Yoon, and Seeun Kim. 2015. When environmental messages should be assertive: examining the moderating role of effort investment. *International Journal of Advertising* 34, 1 (1 2015), 135–157. https://doi.org/10.1080/02650487.2014.993513

[48] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu Hena Mostofa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene* 103, 4 (10 2020), 1621. https://doi.org/10.4269/AJTMH.20-0812

[49] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (4 2021). https://doi.org/10.1145/3449092

[50] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[51] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (11 2018). https://doi.org/10.1145/3274351

[52] Khari Johnson. 2022. DALL-E 2 Creates Incredible Images—and Biased Ones You Don't See. https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media/

[53] S. Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist* 65, 2 (2021), 371–388. https://doi.org/10.1177/0002764219869406

[54] Brian Felipe Keith Norambuena and Tanushree Mitra. 2021. Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (1 2021). https://doi.org/10.1145/3432927

[55] Will Knight. 2021. GPT-3 Can Write Disinformation Now—and Dupe Human Readers. https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/

[56] Bill Kovach and Tom Rosenstiel. 2011. *Blur: How to know what's true in the age of information overload*. Bloomsbury Publishing USA.

[57] Bill Kovach and Tom Rosenstiel. 2021. *The Elements of Journalism, Revised and Updated 4th Edition: What Newspeople Should Know and the Public Should Expect*. Crown Publishing Group (NY).

[58] Peter M Krafft and Emma S Spiro. 2019. Keeping rumors in proportion: managing uncertainty in rumor systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[59] Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117.

[60] Kathleen M Kuehn and Leon A Salter. 2020. Assessing digital threats to democracy, and workable solutions: a review of the recent literature. *International Journal of Communication* 14 (2020), 22.

[61] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. 591–602.

[62] Samuli Laato, AKM Islam, Muhammad Nazrul Islam, and Eoin Whelan. 2020. Why do people share misinformation during the Covid-19 pandemic? *arXiv preprint arXiv:2004.09600* (2020).

[63] Bruno Latour and Steve Woolgar. 2013. An anthropologist visits the laboratory. In *Laboratory Life*. Princeton University Press, 43–104.

[64] 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.

[65] Allison A Lewinski, Matthew J Crowley, Christopher Miller, Hayden B Bosworth, George L Jackson, Karen Steinhauser, Courtney White-Clark, Felicia McCant, and Leah L Zullig. 2021. Applied rapid qualitative analysis to develop a contextually appropriate intervention and increase the likelihood of uptake. *Medical Care* 59, 6 Suppl 3 (2021), S242.

[66] Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association* 28, 10 (9 2021), 2193–2201. https://doi.org/10.1093/JAMIA/OCAB112

[67] Brooke Fisher Liu, Logen Bartz, and Noreen Duke. 2016. Communicating crisis uncertainty: A review of the knowledge gaps. *Public Relations Review* 42, 3 (9 2016), 479–487. https://doi.org/10.1016/J.PUBREV.2016.03.003

[68] Bridget Lockyer, Shahid Islam, Aamnah Rahman, Josie Dickerson, Kate Pickett, Trevor Sheldon, John Wright, Rosemary McEachan, Laura Sheard, and Bradford Institute for Health Research Covid-19 Scientific Advisory Group. 2021. Understanding COVID-19 misinformation and vaccine hesitancy in context: Findings from a qualitative study involving citizens in Bradford, UK. *Health Expectations* 24, 4 (2021), 1158–1167.

[69] Linqi Lu, Jiawei Liu, Y. Connie Yuan, Kelli S. Burns, Enze Lu, and Dongxiao Li. 2021. Source Trust and COVID-19 Information Sharing: The Mediating Roles of Emotions and Beliefs About Sharing. *Health Education and Behavior* 48, 2 (4 2021), 132–139. https://doi.org/10.1177/1090198120984760

[70] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.

[71] 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.

[72] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, 262–272.

[73] Steve Mollman. 2022. ChatGPT gained 1 million users in under a week. Here's why the AI chatbot is primed to disrupt search as we know it. https://fortune.com/2022/12/09/ai-chatbot-chatgpt-could-disrupt-google-search-engines-business/

[74] Meredith Ringel Morris, Carrie Jun Cai, Jess Scon Holbrook, Chinmay Kulkarni, and Michael Terry. 2022. The Design Space of Generative Models. (2022).

[75] 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.

[76] Moin Nadeem, Anna Bethke, and Siva Reddy. [n.d.]. StereoSet: Measuring stereotypical bias in pretrained language models. ([n. d.]). https://stereoset.

[77] Ahmadand Najee-Ullah, Luis Landeros, Balytskyi Yaroslavand, and Chang Sang-Yoon. 2022. Towards Detection of AI-Generated Texts and Misinformation. In *Socio-Technical Aspects in Security*, Parkin Simonand and Luca Viganò (Eds.). Springer International Publishing, Cham, 194–205.

[78] Taylor Nelson, Nicole Kagan, Claire Critchlow, Alan Hillard, and Albert Hsu. 2020. The Danger of Misinformation in the COVID-19 Crisis. *Missouri Medicine* 117, 6 (2020), 510–512. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721433/

[79] Andrea L. Nevedal, Caitlin M. Reardon, Marilla A. Opra Widerquist, George L. Jackson, Sarah L. Cutrona, Brandolyn S. White, and Laura J. Damschroder. 2021. Rapid versus traditional qualitative analysis using the Consolidated Framework for Implementation Research (CFIR). *Implementation Science* 16, 1 (12 2021), 1–12. https://doi.org/10.1186/S13012-021-01111-5/TABLES/5

[80] Cailin O'Connor and James Owen Weatherall. 2019. *The misinformation age: How false beliefs spread*. Yale University Press.

[81] Jeannette Paschen. 2019. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management* (2019).

[82] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an Infodemic: COVID-19 Fake News Dataset.

[83] James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE* 9, 12 (12 2014), e115844. https://doi.org/10.1371/JOURNAL.PONE.0115844

[84] Gordon Pennycook and David Rand. 2021. Reducing the spread of fake news by shifting attention to accuracy: Meta-analytic evidence of replicability and generalizability. (2021).

[85] Lukas Pohler, Valentin Schrader, Alexander Ladwein, and F von Keller. 2018. A Technological Perspective on Misuse of Available AI. *Consciouscoders, August* (2018).

[86] Alfonso J Rodriguez-Morales and Oscar H Franco. 2021. Public trust, misinformation and COVID-19 vaccination willingness in Latin America and the Caribbean: today's key challenges. *The Lancet Regional Health–Americas* 3 (2021).

[87] Kevin Roose. 2022. AI-Generated Art Won a Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

[88] Kevin Roose. 2022. We Need to Talk About How Good A.I. Is Getting. https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html

[89] Julio A Saenz, Sindhu Reddy Kalathur Gopal, and Diksha Shukla. 2021. Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset). https://doi.org/10.21227/b5bt-5244

[90] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)* Icwsm (2020). https://www.aaai.org/ojs/index.php/ICWSM/article/view/7326

[91] Koustuv Saha, John Torous, Eric D Caine, and Munmun De Choudhury. 2020. Psychosocial effects of the COVID-19 pandemic: large-scale quasi-experimental study on social media. *Journal of medical internet research* 22, 11 (2020), e22600.

[92] Mattia Samory and Tanushree Mitra. 2018. 'The Government Spies Using Our Webcams': The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018), 1–24. https://doi.org/10.1145/3274421

[93] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669.

[94] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

[95] Michel Setbon and Jocelyn Raude. 2010. Factors in vaccination intention against the pandemic influenza A/H1N1. *European Journal of Public Health* 20, 5 (10 2010), 490–494. https://doi.org/10.1093/EURPUB/CKQ054

[96] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 312–320.

[97] Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology* 72, 1 (2021), 3–17.

[98] Denis Skopin. 2022. *Photography and Political Repressions in Stalin's Russia: Defacing the Enemy*. Routledge.

[99] Jiao Sun, Tongshuang Wu, Yue Jiang, Ronil Awalegaonkar, Xi Victoria Lin, and Diyi Yang. 2022. Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3502114

[100] Abhijit Suprem and Calton Pu. 2022. Exploring Generalizability of Fine-Tuned Models for Fake News Detection. *arXiv preprint arXiv:2205.07154* (2022).

[101] Edson C. Tandoc. 2019. The facts of fake news: A research review. *Sociology Compass* 13, 9 (9 2019), e12724. https://doi.org/10.1111/SOC4.12724

[102] Edson C Tandoc Jr and James Chong Boi Lee. 2022. When viruses and misinformation spread: How young Singaporeans navigated uncertainty in the early stages of the COVID-19 outbreak. *New Media & Society* 24, 3 (2022), 778–796.

[103] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[104] The Wall Street Journal. [n.d.]. The Difference Between News & Opinion. https://newsliteracy.wsj.com/news-opinion/

[105] Rob Toews. 2020. Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/

[106] 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.

[107] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (3 2018), 1146–1151. https://doi.org/10.1126/SCIENCE.AAP9559/SUPPL{_}FILE/AAP9559{_}VOSOUGHI{_}SM.PDF

[108] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 384, 14 pages.

[109] Jevin D West and Carl T Bergstrom. 2021. Misinformation in and about science. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e1912444117.

[110] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review* 9, 11 (2019).

[111] Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.* 52 (2020), 1029.

[112] World Health Organization. 2020. Let's flatten the infodemic curve. https://www.who.int/news-room/spotlight/let-s-flatten-the-infodemic-curve

[113] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018* (4 2018), 603–612. https://doi.org/10.1145/3184558.3188731

[114] Yixuan Zhang, Nurul Suhaimi, Nutchanon Yongsatianchot, Joseph D Gaggiano, Miso Kim, Shivani A Patel, Yifan Sun, Stacy Marsella, Jacqueline Griffin, and Andrea G Parker. 2022. Shifting Trust: Examining How Trust and Distrust Emerge, Transform, and Collapse in COVID-19 Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 78, 21 pages. https://doi.org/10.1145/3491102.3501889

[115] Jiawei Zhou, Koustuv Saha, Irene Michelle Lopez Carron, Dong Whi Yoo, Catherine R. Deeter, Munmun De Choudhury, and Rosa I. Arriaga. 2022. Veteran Critical Theory as a Lens to Understand Veterans' Needs and Support on Social Media. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 133 (apr 2022), 28 pages. https://doi.org/10.1145/3512980

A LINGUISTIC DIFFERENCES BETWEEN AI-MISINFO AND HUMAN-MISINFO

	Misinfo		Misinfo: News		Misinfo: Post	
Linguistic Features	**Human**	**AI**	**Human**	**AI**	**Human**	**AI**
Language Style						
Analytic (CDI)	88.27 (±15.44)	84.28 (±17.68)	87.32 (±16.45)	90.42 (±11.67)	88.88 (±14.74)	80.39 (±19.64)
Clout	66.02 (±20.26)	67.32 (±17.66)	67.64 (±21.51)	65.57 (±17.46)	65.00 (±19.35)	68.43 (±17.69)
Authentic	22.45 (±22.08)	18.49 (±18.54)	21.46 (±22.04)	22.69 (±20.10)	23.07 (±22.09)	15.83 (±16.94)
Tone*	30.26 (±29.82)	33.02 (±31.53)	26.67 (±26.36)	25.56 (±26.98)	32.53 (±31.61)	37.74 (±33.25)
Informal Attributes						
Informal	0.75 (±1.47)	0.08 (±0.37)	0.25 (±0.71)	0.05 (±0.22)	1.07 (±1.72)	0.10 (±0.44)
Netspeak	0.66 (±1.31)	0.02 (±0.18)	0.25 (±0.71)	0.01 (±0.09)	0.92 (±1.52)	0.02 (±0.22)
Affective Attributes						
Affect	3.06 (±2.86)	4.27 (±2.75)	2.84 (±2.9)	3.60 (±2.39)	3.21 (±2.82)	4.69 (±2.88)
Positive emotion	1.36 (±2.20)	2.04 (±1.99)	1.15 (±2.31)	1.40 (±1.72)	1.49 (±2.11)	2.44 (±2.04)
Negative emotion	1.67 (±2.02)	2.17 (±2.01)	1.69 (±2.00)	2.19 (±1.86)	1.65 (±2.03)	2.16 (±2.11)
Anxiety	0.34 (±1.04)	0.55 (±0.94)	0.49 (±1.32)	0.59 (±1.03)	0.25 (±0.79)	0.52 (±0.88)
Anger	0.59 (±1.41)	0.69 (±1.34)	0.43 (±1.20)	0.61 (±1.20)	0.70 (±1.52)	0.74 (±1.43)
Sad	0.31 (±0.92)	0.36 (±0.94)	0.31 (±0.84)	0.35 (±0.90)	0.31 (±0.96)	0.37 (±0.96)
Cognitive Attributes						
Cognitive process	5.80 (±4.63)	8.57 (±4.87)	5.51 (±4.90)	7.86 (±4.52)	5.98 (±4.44)	9.02 (±5.03)
Insight	1.11 (±1.79)	1.69 (±1.81)	0.91 (±1.59)	1.57 (±1.68)	1.24 (±1.90)	1.77 (±1.89)
Causation	1.26 (±1.90)	1.89 (±2.06)	1.09 (±1.83)	1.62 (±1.74)	1.37 (±1.94)	2.06 (±2.22)
Discrepancy	0.69 (±1.71)	1.23 (±1.70)	0.67 (±1.92)	0.88 (±1.35)	0.71 (±1.55)	1.45 (±1.85)
Tentative	1.19 (±1.89)	1.80 (±1.99)	1.17 (±1.98)	1.77 (±2.05)	1.20 (±1.84)	1.82 (±1.95)
Certitude	0.98 (±1.50)	1.08 (±1.43)	0.72 (±1.39)	0.77 (±1.27)	1.15 (±1.55)	1.28 (±1.49)
Differentiation	1.64 (±2.36)	2.35 (±2.21)	2.11 (±2.77)	2.51 (±2.48)	1.34 (±1.99)	2.25 (±2.01)
Perceptive Attributes						
Perception	2.80 (±3.17)	2.27 (±2.40)	3.46 (±3.35)	2.46 (±2.50)	2.39 (±2.98)	2.15 (±2.32)
See	1.51 (±2.78)	1.05 (±1.68)	2.17 (±3.33)	1.23 (±1.89)	1.09 (±2.26)	0.93 (±1.52)
Hear	0.52 (±1.42)	0.53 (±1.11)	0.58 (±1.46)	0.59 (±1.16)	0.48 (±1.39)	0.49 (±1.08)
Feel	0.51 (±1.59)	0.56 (±1.39)	0.53 (±1.58)	0.55 (±1.55)	0.50 (±1.60)	0.57 (±1.28)
Drives Attributes						
Drive	7.04 (±5.49)	8.08 (±4.02)	6.76 (±5.66)	8.00 (±3.50)	7.22 (±5.37)	8.13 (±4.31)
Affiliation	2.09 (±3.19)	1.67 (±2.14)	2.03 (±3.09)	1.58 (±2.13)	2.12 (±3.25)	1.73 (±2.14)
Achievement	1.27 (±2.14)	1.42 (±1.65)	1.12 (±2.19)	1.26 (±1.58)	1.36 (±2.11)	1.52 (±1.68)
Power	2.98 (±3.09)	3.56 (±2.56)	3.04 (±3.12)	3.74 (±2.49)	2.95 (±3.08)	3.46 (±2.60)
Reward	0.58 (±1.10)	1.03 (±1.26)	0.36 (±0.91)	0.85 (±0.99)	0.71 (±1.18)	1.14 (±1.39)
Risk	0.82 (±1.55)	1.36 (±1.63)	0.91 (±1.67)	1.35 (±1.52)	0.76 (±1.47)	1.37 (±1.69)

* The higher the number, the more positive the tone. Numbers below 50 suggest a more negative emotional tone.