Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages

ELISE KARINSHAK, University of Georgia, USA
SUNNY XUN LIU, Stanford University, USA
JOON SUNG PARK, Stanford University, USA
JEFFREY T. HANCOCK, Stanford University, USA

Artificial Intelligence (AI) is a transformative force in communication and messaging strategy, with potential to disrupt traditional approaches. Large language models (LLMs), a form of AI, are capable of generating high-quality, humanlike text. We investigate the persuasive quality of AI-generated messages to understand how AI could impact public health messaging. Specifically, through a series of studies designed to characterize and evaluate generative AI in developing public health messages, we analyze COVID-19 pro-vaccination messages generated by GPT-3, a state-of-the-art instantiation of a large language model. Study 1 is a systematic evaluation of GPT-3's ability to generate pro-vaccination messages. Study 2 then observed peoples' perceptions of curated GPT-3-generated messages compared to human-authored messages released by the CDC (Centers for Disease Control and Prevention), finding that GPT-3 messages were perceived as more effective, stronger arguments, and evoked more positive attitudes than CDC messages. Finally, Study 3 assessed the role of source labels on perceived quality, finding that while participants preferred AI-generated messages, they expressed dispreference for messages that were labeled as AI-generated. The results suggest that, with human supervision, AI can be used to create effective public health messages, but that individuals prefer their public health messages to come from human institutions rather than AI sources. We propose best practices for assessing generative outputs of large language models in future social science research and ways health professionals can use AI systems to augment public health messaging.

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

Additional Key Words and Phrases: natural language processing, large language models, AI-mediated communication, persuasion, public health messaging, message factors

ACM Reference Format:
Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (April 2023), 27 pages. https://doi.org/10.1145/3579592

1 INTRODUCTION

Public health messaging is critical in promoting beneficial public health outcomes. Public health campaigns are of significant societal importance, as effective messages nudge people to make better decisions on behalf of themselves and their communities; campaigns address diverse behavioral goals and achieve varying levels of success (e.g., [24, 73, 94]). Public health messaging research focuses on composing messages to convey high-stakes information and persuade

---

Authors' addresses: Elise Karinshak, elise.karinshak@uga.edu, University of Georgia, USA; Sunny Xun Liu, Stanford University, USA, sunnyxliu@stanford.edu; Joon Sung Park, Stanford University, USA, joonspk@stanford.edu; Jeffrey T. Hancock, Stanford University, USA, hancockj@stanford.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

people to engage in healthy behaviors [55]. The COVID-19 pandemic and the need to vaccinate the world population has been perhaps the highest stakes public health communication campaign in generations.

Despite the U.S. government allotting billions toward CDC (Centers for Disease Control and Prevention) vaccination-related activities [81] and the Department of Health and Human Services investing a historic $250 million into coronavirus campaign efforts [59], vaccine hesitancy and low vaccination rates remain across most demographic segments [18, 77]. Public health communication faces many challenges in today's media ecosystem, with substantial misinformation regarding vaccinations on the Internet, and people's trust in governments and health institutions on the decline. Individuals continue to express concerns relating to vaccine development, safety, and whether vaccination is necessary or effective [62]. In communication surrounding COVID-19, key challenges facing practitioners include the prevalence of misinformation, politicization of the crisis, and health inequality [60].

At the same time, Artificial Intelligence (AI) technologies have made important, transformative advances in generating human language. There are a number of important research themes that have emerged in the CSCW literature, including the perceptions of AI systems (e.g., [49, 80]), their role in social media and digital campaigns (e.g., [11, 13, 70]), the impact of AI on health information seeking (e.g., [2, 82]), and the ethical implications of AI and data in public health communication and persuasion (e.g., [6, 17, 64, 69]). In particular, a class of AI referred to as large language models (LLMs) have made substantial progress in generating language. These models have been trained on textual data consisting of trillions of words, are composed of billions of parameters, and can be adapted to many downstream tasks [8], to include writing poetry, code, and tweets [20, 56]. Such models have powerful generative capacities and have achieved notable success in content generation across a range of contexts [5, 8]. Is it possible that this kind of AI could augment public health communication practices to develop more effective vaccination messages?

In this instance of AI-mediated communication, defined as the "mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals" [36], we examine the potential for leveraging LLMs, a class of models within AI, in content creation for public health messaging. The present research assesses potential contributions of LLMs, specifically the GPT-3 model, to public health communication workflows and provides guidance for future integration; additionally, the analysis provides methods for assessing the persuasive quality of LLM-generated messages.

Today, public health communication practitioners are guided by analytics in campaign development and strategy. Data-driven campaign development generally adheres to a workflow with several key activities: situational research and analysis to understand stakeholders and their priorities; identification and monitoring of emerging opportunities and threats; selection of target audiences and target audience research, such as understanding psychometric characteristics to identify messages that will resonate with specific audiences; channel selection; content development; and measurement and evaluation [27, 61].

We propose that within this workflow, an LLM may be able to augment several activities by rapidly generating content that is tailored by situation and target audience due to AI's ability to draw upon extensive, relevant verbal user data for content creation [5]. As such, an LLM may augment practitioners' abilities to develop messages that resonate with target audiences. Prior research has demonstrated that AI can support human decision-making and persuasion [35, 42, 44, 50, 58, 92], including in communication tasks within high stakes domains such as health (e.g., [19, 42, 84]). Emerging findings demonstrate that generative AI offers a promising direction for creative content development systems, such as in writing tasks (e.g., [3, 10, 92]). Such AI-supported systems have the potential to support content development by enabling the release of more targeted, effective public health content, or augmenting or strengthening content development workflows.

To examine this potential for generative AI in supporting public health communication, the present research examines how LLMs can serve as a tool for content creation in public health messaging. For example, when given the prompt "Persuade someone to receive a coronavirus vaccine:" (along with two examples of messages), a prominent language model generated the following message:

> The most important thing you can do to protect yourself and those around you from the coronavirus is to get the vaccine as soon as possible. The sooner you get vaccinated, the sooner you can start to do things that you may not have been able to do before—like travel, attend work or school, go out in public, visit friends or family, etc. The best time to get vaccinated is now!

However, such models are nondeterministic and can produce wildly varying content, which can be less effective or contain outright wrong information. For instance, when the same prompt was submitted again, the same model produced the following message, which has low levels of linguistic appropriacy and logical argumentation [19] and clearly could not be used as a public health message:

> Take the information you learned from the vaccine insert with you to your health care provider to help this process go smoothly.

Given these strengths and limitations of LLMs in content generation, the present research evaluates how such models can contribute to public health messaging. In particular, we examine an LLM's potential to create effective and persuasive COVID-19 vaccination messages, as well as the impact of source labels (traditional sources versus AI) on individuals' perceptions of persuasiveness. Such analysis provides insight into how practitioners can leverage AI to augment existing public health messaging workflows and provides methods for future research evaluating the persuasive quality of AI-generated messages.

2 LITERATURE REVIEW

As AI models become increasingly sophisticated, scholars are exploring the implications of these technologies across systems and industries in which such technologies can be applied. Emerging discussion focuses upon the future of AI in persuasion [19, 26, 42, 66] and human-AI decision-making and human collaboration, especially in high-stakes domains (for review, see [48]).

Previous work suggests that AI can deliver various elements of persuasion, yet the effects of the persuasion differ depending on the designs of AI systems, how they are implemented, content topic, and intended recipient. For example, medical professionals readily accept AI advice without higher suspicion than advice from their peers [31]. Yet, when participants encounter health messages labeled as AI-generated, these messages were only rated as more appropriate and more effective when their contents focus on how rather than why to perform a health-related action; participants believed that AI lacks consciousness, a human's unique ability to form goals and intentions [44]. This lack of consciousness, or perceptions of dehumanization, can be detrimental to persuasion. A recent study found that COVID-19 data generated by AI (versus humans) reduced people's intentions to take preventive measures, such as wearing masks and social distancing, because people did not connect human lives, but numbers with AI-generated data [41].

This literature, along with the emerging body of AI-mediated communication research, suggests that AI has the potential to meaningfully contribute to communication systems, although there are serious concerns about bias and misuse with such systems. Buchanan et. al [9], for example, demonstrate how LLMs such as GPT-3 can be leveraged to effectively automate disinformation at a large scale. In the context of sensitive information, such as health communication, research has identified biases in algorithms and their underlying data along with a lack of transparency in how AI functions [51].

One approach to mitigating these concerns is to entirely forgo automated systems and rely instead on how AI technologies can augment, or support, human communication practices. Research on "algorithm-in-the-loop" AI systems, in which humans have discretion and ultimate review, suggests this form of human-AI collaborations as a promising direction for future systems (e.g., [34]), such as providing support in drafting and editing processes (e.g., [3]). At an institutional level, adopting AI technologies such as LLMs could offer significant advantages to public health organizations in their communication efforts. The ability of LLMs to access and synthesize insights from a large breadth of data and language information [5] could replace more time-consuming and costly traditional research methods seeking to understand target audiences, such as surveys and focus groups, streamlining content development processes. This approach may enable practitioners to redirect limited resources toward additional channels and content development for multimodal platforms (e.g., videos on TikTok, images on Instagram), which have proven essential strategic elements of previous and future public health communication efforts (e.g., [45]).

2.1 Present Research

In this research, we build upon and extend prior research by evaluating AI performance in the context of an important real-world application: the development of COVID-19 vaccination public health messages. Specifically, we evaluate GPT-3, which has achieved state-of-the-art performance among LLMs. The research comprises three studies across two phases: Phase I involves one study, focusing on the system side of the equation and examining message generation and benchmarking, while Phase II involves two studies, focusing on the human side and examining how people perceive GPT-3-generated messages. In Phase I, we conduct a study that systematically characterizes how an LLM can be used to produce pro-vaccination messages and evaluate these messages along baseline criteria (e.g., accuracy, relevance, and attempted persuasion). In Phase II, we conduct two studies to examine perceptions (perceived effectiveness, argument strength, and positive attitudes evoked) of GPT-3-generated messages by experimentally comparing GPT-3-generated messages with messages created by the CDC (Study 2) and by examining the influence of source labels (e.g., "CDC", "doctors", "AI") (Study 3). The preregistration document detailing each of these studies is located at the following URL: https://osf.io/nm6s8.

The pattern of results across the studies suggests that with careful human oversight, LLMs like GPT-3 can augment public health messaging. The results from Study 1 revealed that GPT-3's message generation capacity is highly variable, producing both highly effective pro-vaccination messages and poor quality messages requiring substantial editing or that were even factually incorrect. The results from our perceptual studies suggest that 1) the best GPT-3 messages from Study 1 are perceived by audiences as effective and persuasive, but 2) our participants trusted messages less if they believed they were written by AI than if the source of the message was a human institution, namely the CDC. This pattern suggests that AI should only be used to augment public health messaging, with strong input and oversight by health officials, and not as a method to automate public health messaging.

3 PHASE I: SYSTEMATIC GENERATION AND EVALUATION OF AI CONTENT

AI offers a significant opportunity for the generation of copy in marketing and advertising contexts, and public health campaigns are one of many practical applications for generation of creative and informative copy. In this research, we assessed AI's baseline ability to generate public health content through message generation and benchmarking. By systematically evaluating AI output quality, we observed the strengths, shortcomings, and consistency of AI in generating such content. This exploratory analysis provides a foundational understanding of the current capabilities of AI and offers insight into how public health professionals can effectively leverage AI in content generation.

Natural Language Processing (NLP), a subfield of artificial intelligence, explores how computers can accomplish tasks related to language, such as speech-to-text translation and text generation. Large language models (LLMs), which are trained on massive, broad datasets and leveraged for a variety of tasks [7], have gained increasing prominence. A variety of LLMs for language tasks, such as ELMo, Gopher, and GPT-3, have been introduced in recent years [5, 75]. In the present paper, we employ the third-generation Generative Pre-trained Transformer (GPT-3), an autoregressive LLM released in 2020 that has gained attention for its impressive capabilities [8]. It is capable of generating quality, humanlike text by probabilistically predicting successive words.

The GPT-3 model has 175 billion parameters that were trained through unsupervised learning on trillions of words of Internet data from 2016 until 2019. Additionally, unlike models such as GPT-2 (its precursor), GPT-3 adopts a "task-agnostic" approach and does not require fine-tuning (task-specific training); the general pre-trained model can be used for diverse text-generation tasks [8]. Because GPT-3 does not rely on task-specific training, the model can be quickly and easily adopted for diverse purposes [72].

GPT-3 achieved excellent performance on metrics established for evaluating artificial intelligence models. In preliminary testing of GPT-3's output quality, when tasked with determining whether a news article was human-written or GPT-3-generated, participants only achieved an accuracy of 52%. Research to date suggests that GPT-3 can produce quality output and that people cannot effectively distinguish between GPT-3-generated and human-written content [12]. GPT-3 is also able to generate text at a massive scale. According to OpenAI, GPT-3 generates an average of 4.5 billion words per day [72].

Content generation with GPT-3 is heavily dependent on prompting methods. Prompts (the text input fed to the model) and request parameters affect how the model generates output (for review of prompt methodology, see [54]). For the present study, we adopted a prompting method that employed a simple text-based prompt for the GPT-3 system to generate a pro-vaccination message. In Study 1, we carefully describe the prompt methodology along each step. While more advanced prompt development strategies are available, such as discrete prompts and continuous prompts [54], these methods require more advanced technical skills and are less likely to be used in practice by communication professionals than the intuitive, manual method we describe. The prompting method used in Study 1 was designed to be similar to how an advertising or public health professional might use GPT-3.

A variety of evaluation approaches have been used to characterize GPT-3 output, as these approaches vary by task [91]. In the context of public health messaging, we assessed whether messages were accurate and relevant, as these are common metrics for evaluating AI output quality [91] and minimum standards for public health content [67]. We also assessed whether persuasion is attempted; in this case, desirable output attempted persuasion since the messages were intended for a pro-vaccination context [55, 71]. While more advanced computational approaches for reviewing natural language output exist [19, 42], this exploratory analysis is intended to provide an understanding of whether AI-generated output satisfies the minimum standards for a public health message, and to quantitatively characterize these messages to better understand the nature of the output that could be used with and without modification.

RQ1: What are the qualities of GPT-3 messages regarding accuracy, relevance, and persuasion?

Table 1. GPT-3 output content coding definitions

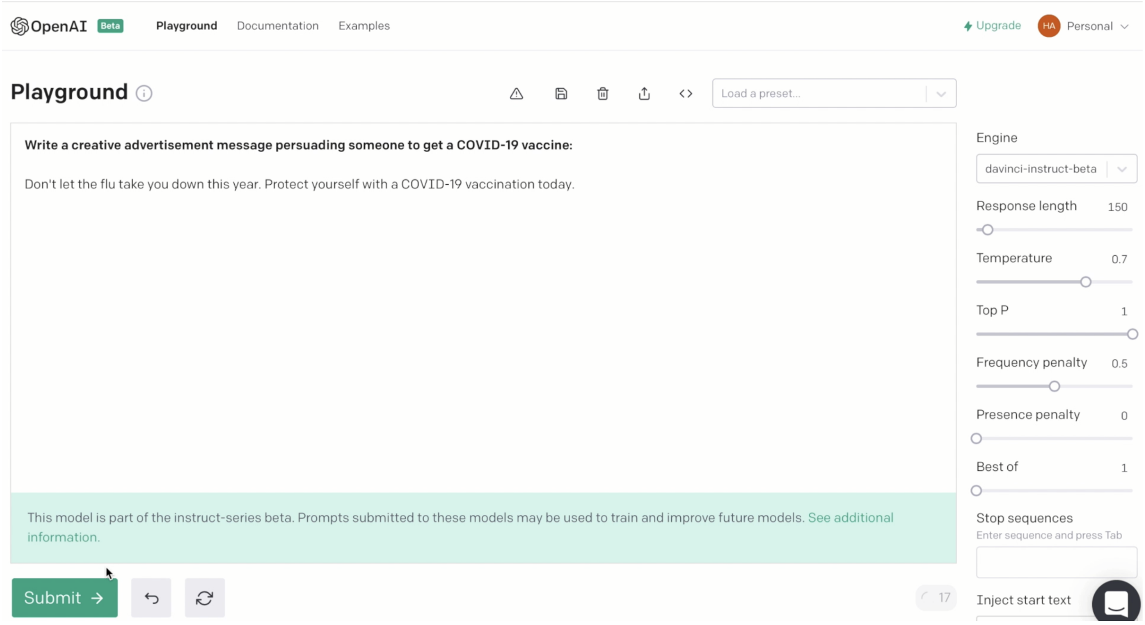
Dimension	Variable Type	Description
Accurate	binary	Whether factual information is correct enough to be used as a public health message without edits. If factual content is not included, noted as accurate (ex. if message expresses opinion).
Relevant	binary	Whether the message is relevant to vaccination and not does not describe vaccines for illnesses other than COVID-19 (unless providing COVID-19-relevant context).
Attempts Persuasion	binary	A reasonable person reading the message would understand that the message is promoting a desirable public health behavior (such as vaccination).
Edits Required	discrete scale (0 - 5+)	The estimated minimum number of words to adjust for the completion to be accurate, relevant, and attempt persuasion. Includes word insertion, deletion, or modification (ex. changing the subject of the message from "flu" to "COVID-19"). 5+ indicates that significant change is necessary (ex. adding a sentence).
Complete	binary	Whether the thought is complete (versus cut off); this value served as a quality filter, as some completions exceeded the token length limit.

3.1 Method

We used GPT-3 to generate pro-vaccination public health messages; the output messages were then coded for quality by standards for public health messages. GPT-3's performance depends on the request parameters and the prompt, which both directly affect the resulting completion, or instance of GPT-3 output. To achieve appropriate output and accurately assess GPT-3's capabilities, these factors should be reasonably optimized [54]. The desired GPT-3 output was COVID-19 vaccination messages that were accurate [91], relevant [91], and attempted persuasion [55, 71]. Finally, required human editing, assessed as the estimated minimum number of words to modify to satisfy each dimension, was observed; this metric characterizes the degree of editing necessary. Definitions for these dimensions are provided in Table 1. See Appendix A for details related to setting the GPT-3 parameters.

3.1.1 Prompt Engineering. To generate messages, GPT-3 requires a prompt. Study 1 uses manual template engineering for the prompts [54], which involves the creation of intuitive natural language as this is the simplest prompt engineering process and replicates how GPT-3 could be used in practice by communicators without technical expertise. A variety of possible prompts (varying in format, length, and supplied examples) were iteratively tested to characterize output generated by GPT-3 (see Figure 1). These were tested with wording variations, such as “COVID-19” versus “coronavirus” and “convince” versus “persuade”. Prompts were developed in the following progression: zero-shot prompts, few-shot prompts, and prompts for davinci-instruct-beta. “Zero-shot” prompts provide simple instructions and no additional context. “Few-shot” prompts incorporate examples of desired output (such as public health messages from federal and state government sources). Previous research has achieved more substantive output with few-shot prompts [8]. Finally, we generated messages with zero-shot prompts for the davinci-instruct-beta engine, which was developed to more effectively interpret instructions (in comparison to other engines).

3.1.2 Coding Procedure. The generated messages were saved and coded along the following dimensions: accuracy [91], relevance [91], attempts persuasion [55, 71], and the estimated amount of human editing required (see definitions in Table 1).



Write a creative advertisement message persuading someone to get a COVID-19 vaccine:

Don't let the flu take you down this year. Protect yourself with a COVID-19 vaccination today.

This model is part of the instruct-series beta. Prompts submitted to these models may be used to train and improve future models. See additional information.

Engine: davinci-instruct-beta
Response length: 150
Temperature: 0.7
Top P: 1
Frequency penalty: 0.5
Presence penalty: 0
Best of: 1
Stop sequences: Enter sequence and press Tab
Inject start text

Fig. 1. GPT-3 interface and request parameters. *Note: Prompt and completion are illustrative. For final prompts and corresponding output, see section 3.2.*

3.2 Results

We used three approaches to generate messages: zero-shot prompts, few-shot prompts, and davinci-instruct-beta (see Table 2 and Appendix A for detailed results).

The first round of prompts tested were "zero-shot" prompts, which provide simple instructions and no additional context, such as: *"You should get a COVID-19 vaccine because:"* These prompts were also run with a sentence or short paragraph of context: *"COVID-19 is a highly contagious respiratory disease caused by SARS-CoV-2. This ad message convinces you to get a COVID vaccine:"*. The resulting output occasionally satisfied the accuracy, relevance, and persuasion dimensions, but the output was often short and casual with little factual support or reference to benefits of vaccination. Of the 10 initial prompts and three trials for each prompt (for a total of 30 GPT-3 completions), four (13.3%) completions were accurate, relevant, and could be used as persuasive messages without edits. Of messages requiring edits, the average edits needed to fix accuracy was 3.73 words and greater than 5 words for relevance and persuasion. Of messages requiring edits, the average number of words to adjust was 4.13 for accuracy, 2.84 for relevance, and greater than 5 for persuasion.

We then ran the following few-shot prompt 27 times (for 10 of these trials, the prompt was run with "COVID-19" instead of "coronavirus"). In the prompt, the "—-" breaks separate the distinct examples supplied to the model.

> Persuade someone to receive a coronavirus vaccine.
> —-
> When you get a coronavirus vaccine, you are choosing to protect yourself and make a difference for your children, parents, grandparents, and other loved ones. Millions of people in the U.S. have already received a coronavirus vaccine. For a community to be fully protected, most community members need to get the

> vaccine. Getting vaccinated to prevent the coronavirus will help protect you from the coronavirus, and it may also protect the people around you.
>
> —-
>
> Vaccines prevent the spread of the coronavirus and will help bring this pandemic to an end. The coronavirus vaccines are safe and effective. After you've been fully vaccinated, you can start to do some things that you had to stop doing because of the pandemic.
>
> —-

These 27 prompt completions generated 52 messages, 13 of which contained incomplete sentences due to the length constraint. Of the remaining messages, 16 (30.8%) were accurate, relevant, attempted persuasion, and could be used without modification. Of messages requiring edits, the average edits needed to fix accuracy was 3.73 words and greater than 5 words for relevance and persuasion. Although the examples above focus on keeping loved ones and the community safe and ending the pandemic, the resulting output also addressed a wide variety of relevant topics, such as FDA approval, how immune systems respond to vaccines, COVID-19 symptoms, and preventative measures for unvaccinated individuals.

Finally, the davinci-instruct-beta engine was tested with zero-shot prompts. This engine was designed to interpret the supplied instructions without examples of desired output. The following prompt was run 10 times: *"Persuade someone to receive a coronavirus vaccine in a paragraph:"*. Of the 10 resulting messages, 2 (20.0%) were accurate, relevant, attempted persuasion, and could be used without modification. Of messages requiring edits, the average number of words requiring edits to satisfy accuracy was 4.33 and greater than 5 words for relevance and persuasion. With this prompt and the davinci-instruct-beta engine, GPT-3 produced some quality messages, but the model's overall performance was inconsistent. Sample GPT-3 messages corresponding to each prompt design are provided in Table 2.

3.3 Discussion

Our results indicate that GPT-3 achieved inconsistent performance when generating public health messages. Regarding RQ1, GPT-3 is capable of creating accurate, relevant, and persuasive messages, but these criteria are inconsistently met. Few-shot prompts and the davinci-instruct-beta engine achieved more promising output. This pattern of results has important implications for the use of LLMs by health communication practitioners. LLMs offer notable advantages in message generation, but require substantial human review. In public health workflows, communication practitioners should be able to use LLMs to develop prompts and adjust the request parameters based on their desired output for the campaign. In public health contexts, messaging outputs must be accurate, relevant, and persuasive, and ideally they should be creative and compelling. Furthermore, campaigns may have more specific audience objectives (e.g., writing content that appeals to Gen Z), and models should be able to facilitate message generation tailored for specific audiences. Our results, however, highlight that practitioners will need to carefully select the examples supplied by few-shot prompts, including examples incorporating desired tone, message length, and structure. Practitioners should expect an iterative process of development and review of output before finalizing the prompts and parameters, rather than expecting an automated form of message development.

The study highlighted several drawbacks of using LLMs in public health messaging as well. It is important to recognize the time required for prompt optimization and the number of prompts needed. The process of developing and adjusting prompts and request parameters manually can become time-consuming, depending on the task-specific requirements and desired output. In this case, since COVID-19 was an unfamiliar and specific topic, GPT-3 generally

Table 2. Sample GPT-3 output

Prompt Design	Request Parameters	Sample Output
Zero-shot	Engine = davinci Response-length = 100 Temperature = 0.7 Top P = 1.0 Frequency penalty = 0 Presence penalty = 0 Best of = 1	You are at risk for infection with rubella virus, which is transmitted via the airborne route. – This year I didn't have to take a day off work because I was sick with the flu. I got a flu shot. – A COVID-19 vaccine could save your life.
Few-shot	Engine = davinci Response-length = 150 Temperature = 0.8 Top P = 1.0 Frequency penalty = 0.5 Presence penalty = 0 Best of = 1	The COVID-19 virus can cause serious illness, disability, or death. To protect your loved ones, get the vaccine. For more information about the COVID-19 pandemic and vaccine, visit www.cdc.gov/covid19 . – The new CAIV-19 vaccine is a nasal spray that has been approved by the FDA to prevent infection with COVID-19. Like other CAIV-19 vaccines, it is made from weakened viruses that are called "live" or "attenuated" viruses. The viruses in these vaccines are not able to cause illness because they have been weakened. As with other CAIV-19 vaccines, you should not receive this
Zero-shot	Engine = davinci-instruct-beta Response-length = 135 Temperature = 0.7 Top P = 1.0 Frequency penalty = 0.5 Presence penalty = 0 Best of = 1	The coronavirus is a virus that can cause life-threatening pneumonia and sometimes leads to death. The vaccine has shown to be very effective in preventing the virus from spreading so it's a very important vaccine. If you are at risk for the coronavirus and have not been vaccinated, get vaccinated today! – The coronavirus vaccine is a vaccine that protects against the Coronavirus. Generally, it is recommended to receive the vaccine if you are near someone who is infected to help prevent them from spreading the infection.

achieved a low level of consistency and required increasingly specific prompts. In practice, once an effective prompt or set of prompts is developed, these prompts can be reused to generate content at scale–practitioners could develop a curated library of prompt templates and their corresponding settings to streamline this process for subsequent campaigns. Once effective prompts have been created, an LLM can then rapidly generate message drafts and facilitate content creation processes, offsetting the initial time requirement for prompt development.

More closely examining GPT-3's performance, the model's training on data from 2016 to 2019 affected accuracy and relevance, as it was not trained on COVID-19-specific information; however, it was trained on information relating to coronaviruses generally and other previous pandemics and vaccines (such as the flu and flu vaccines, Ebola, MERS, etc.). The model's lack of specific COVID-19 data most directly impacted its ability to reference factual information, undermining the accuracy and relevance of the messages. Specifically, GPT-3 achieved 73.3% accuracy and 40.0% relevance with zero-shot prompts, 57.7% accuracy and 88.5% relevance with few-shot prompts, and 40.0% accuracy and 80.0% relevance with the davinci-instruct-beta engine. Messages incorporating highly-specific COVID-19 facts were more likely to reference incorrect information, such as statistics for other vaccines; however, GPT-3 produced reasonable output for themes such as community safety and returning to normal, as these are less reliant on specific figures. GPT-3's performance varied by the vaccination theme addressed. With more general topics, the model was able to reasonably perform without the most up-to-date information, instead drawing on more generic facts and discussion. Finally, relating to message persuasiveness, messages that failed to satisfy this dimension tended to discuss health-related information without providing a call to action.

GPT-3's performance related to COVID-19 vaccination messaging demonstrates its ability to generate copy addressing topics it was not directly trained to discuss. Although GPT-3 is limited to this training dataset and is not constantly updated to reflect modern developments, its ability to draw on transferable knowledge suggests that it is able to partially overcome this limitation and address topics beyond the scope of its training data. This study demonstrated GPT-3's ability to reference relevant information and write meaningful messages without event-specific data.

4 PHASE II: PERCEPTIONS OF AI-GENERATED MESSAGES

After leveraging GPT-3 to generate public health messages in Phase I, we conduct two studies to explore how people perceive these AI-generated messages. Key criteria for the effectiveness of any public health messages are changes in receivers' beliefs, attitudes, and behaviors. Thus, Study 2 assesses how people interpret these GPT-3-generated messages, as compared to professional, human-written public health messages. Study 3 builds upon Study 2 by exploring whether the effectiveness of the messages changes when receivers are informed that messages are AI-authored relative to human-authored.

5 STUDY 2: CDC VERSUS GPT-3-PRODUCED MESSAGE QUALITY

Message factors, such as imagery, word choice, creativity, music, and mode of communication, significantly affect persuasiveness [78]. Analysis of message construction at the linguistic level demonstrates how language characteristics contribute to persuasiveness. Verbal immediacy, for example, refers to how directly the communicator conveys a threat to the receiver, which can increase persuasiveness [93]. More immediate language contains denotative specificity, spatial immediacy, temporal immediacy, and lack of qualifiers, and these are developed through pronoun choice, verb tense, and object referents [55]. These characteristics are closely related to strategic linguistic agency, the assignment of action or change when conveying a threat, which is a significant factor in vaccination message interpretation [4]. Furthermore, in health-specific analysis, fear appeals and efficacy promotion are widely-embraced persuasive techniques [47]. More recently, Duerr & Gloor [19] added to the traditional persuasion literature by providing the following categories for persuasion and their determinants from a generative perspective: benevolence (creating value for the recipient, e.g., addressing a desired outcome), linguistic appropriacy (congruence in stylome between author and recipient, e.g., usage of amplifiers), logical argumentation (consistency of logic, e.g., consistency and use of connectives), and trustworthiness (e.g., agreeableness, authority). Together, this persuasion literature informs message composition, messaging strategy, and anticipated perceptions in a public health context.

Recipients form a variety of audience segments and differ in their interpretations. In the context of COVID-19 pro-vaccination messages, vaccination status is likely an indicator of vaccine beliefs, with those having received the vaccine likely having a more positive attitude towards the vaccine than those who did not. Because the vaccine was widely available at the time of our studies, most individuals were able to choose whether to receive a vaccine (although individual circumstances such as health conditions or a specific location of residence also affected behavioral outcomes). Despite vaccine availability, a significant number of individuals have not yet received vaccines, due to concerns such as vaccine safety and effectiveness [62]. In addition, demographic and psychographic characteristics often serve as bases for segmentation, as they are significant indicators of participants' attitudes and behaviors [86]. Vaccination rates vary by demographics, such as age, race, education level, and population density in location of residence [30]. Vaccination rates are also related to political beliefs, as the COVID-19 pandemic and vaccination became an increasingly politicized topic in public discourse. Individuals' attitudes were more likely to align with the stance of their preferred political party [30]. Attitudes affect how individuals interpret and judge messages [21]; thus, we expect demographic dimensions and political beliefs to be related to message ratings.

Together, these content and audience considerations inform human-authored communication in traditional public health communication workflows. In contrast, generative AI draws upon training data to automatically generate messages instead of explicitly considering these factors. To assess the effectiveness of AI-generated public health messages, Study 2 empirically evaluates how individuals rate the quality of the messages generated in Study 1 by comparing GPT-3 output to professional public health messages. The CDC (Centers for Disease Control and Prevention) was selected as a standard for such content, as it is a prominent and trusted source of public health information [29, 65]. Furthermore, the CDC has been described as a "benchmark organization for the use of social media for health" [79]. Using established scales from public health communication literature, messages from the CDC and those generated by GPT-3 were compared to assess perceptions of AI-generated messages.

Existing research has assessed AI perceptions through a variety of applications, finding that individuals' reactions to AI vary by context. AI assistance is perceived negatively in interpersonal communication [39] and job applications [57], and professional meteorologists are perceived as significantly more credible than AI reporters [88]. Artwork from human and AI sources are also perceived differently, with human-created artwork achieving higher quality ratings for composition, degree of expression, and aesthetics than AI-generated art [40]. Kim & Duhachek [44] find that in persuasive tasks, people perceive artificial intelligence as more appropriate for low-construal settings. Together, this literature suggests people generally have more negative perceptions of AI than humans.

Study 2 is an assessment of perceptions of AI-generated message quality (relative to CDC-written messages). The study controls for source (none displayed) to isolate the message-level quality comparison. In previous research, GPT-3 has demonstrated the ability to write convincingly humanlike copy [12]; however, its output quality is limited by rudimentary extrapolation capabilities, resulting in inconsistent factual accuracy (a challenge particularly salient for applications in healthcare) [5]. Furthermore, generated text can output suboptimal wording (such as unnecessary repetition and unorthodox phrasing) [8]. Given the CDC's expertise in public health communication and GPT-3's current limitations and lack of COVID-19-specific information, although we anticipate that AI models may eventually improve the quality of messages, we conservatively hypothesize the following relationships for cognitive beliefs (H1) and affective attitudes (H2) perceptions between content quality indicators for GPT-3 and the CDC:

> H1: Participants will perceive CDC-written COVID-19 vaccine messages to be of higher quality than GPT-3-produced COVID-19 vaccine messages.
>
> H2: Participants will have more positive attitudes towards COVID-19 vaccines after exposure to CDC-written COVID-19 vaccine messages versus GPT-3-produced COVID-19 vaccine messages.

Additionally, given the relationship between vaccination status, political beliefs, and demographics with vaccine attitudes, as well as the role of attitudes in message judgment, we assessed:

> RQ1: Would the messaging effects differ between those who have received vaccines and those who have not?

> RQ2: How would demographic variables and people’s political party preferences impact their perceptions of message quality?

5.1 Method

We conducted an *a priori* power analysis to determine required sample size. For a two-tailed t-test with an effect size of 0.2, power of 0.8, and alpha of 0.05, the resulting required minimum sample size was 788.

5.1.1 Participants. We recruited participants from Amazon Mechanical Turk (mTurk) who completed more than 5000 hits and who had an approval rate of 95% and above. We recruited additional respondents to allow the removal of low quality data and the respondents who failed the manipulation check; in total, 873 mTurkers participated. The study was approved by the university IRB (IRB # 51574). After reviewing a consent form and giving their consent, participants started the study. Each participant was paid 75 cents for their participation. We excluded 21 participants who completed the study in less than a minute. Our pilot test indicated that it took around 3 minutes to complete the survey. Finishing the survey within a minute showed the participants’ lack of attention. The final sample size was 852 (47.4% female). 30.5% of the participants were between the ages of 25-34, followed by 35-44 (29.9%), 45-59 (21.2%), 60-69 (12.6%), 18-24 (3.3%), and 70-79 (2.5%). 78.5% of the participants were White, followed by Asian (10.4%), African American (7.6%), Latino (6.7%), and others (1.3%). 44.2% of the participants had a Bachelor’s degree, followed by high school graduates (24.5%), Associate degree in college (14.1%), Master’s degree (13.3%), Ph.D. or higher (2.5%), and others (1.4%). Each participant was randomly assigned to review one message (either written by the CDC or GPT-3).

5.1.2 Stimulus Sampling. We collected CDC messages from examples of social media posts related to vaccine messages (featured on the CDC website in their official COVID-19 Vaccination Communication Toolkit) and recent CDC posts on social media. The 10 messages selected cover a variety of topics relevant to vaccine messaging, including: the emerging Delta variant, the safety of communities and loved ones, and people’s desire to return to normal. Given the official nature of this content, these CDC messages are assumed to be of consistent, professional quality; thus, curation is not necessary for ensuring the sample messages are of reasonable quality. For the GPT-3 stimuli, after review of the GPT-3 output from Study 1, 10 of the messages that satisfied the criteria of accuracy, relevance, and attempting persuasion were selected; the curated messages addressed a variety of vaccination themes and were intended to represent AI’s best output (Appendix B). These messages served as the stimuli representing GPT-3 output. Although the CDC messages and the GPT-3 messages are obviously not the same, they match in the following key dimensions: length (CDC message average word count: 65.1, GP3-3 message average word count 49.2), topic (COVID-19 vaccination), and intended persuasion goal (get vaccinated).

5.1.3 Procedure and Measures. Respondents were randomly presented with one message and asked to rate the message on scales for perceived message effectiveness [63], perceived argument strength [95], and post-exposure attitudes [95]. These dependent variables are established indicators of message quality in public health persuasion research. Perceived message effectiveness and argument strength are cognitive indicators, while attitude is an affective indicator. After completing the message ratings, the participants completed demographic items and debriefing information before receiving payment.

Perceived message effectiveness [63] assesses persuasiveness, believability, and processing; the scale contained 4 items (ex. “This message made me stop and think”), and participants rated how strongly they agreed with each statement on a 5-point scale (labeled “strongly disagree” to “strongly agree”) (M = 3.47, SD = 1.00, alpha = .83). Perceived

Table 3. Sample Stimuli Messages

GPT-3	CDC
While no vaccine can guarantee complete protection, CDC and FDA continue to monitor the safety of all COVID-19 vaccines. If you are vaccinated, you can decrease your risk of getting sick with COVID-19.	COVID-19 vaccination helps keep you from getting COVID-19. The vaccines currently available in the United States are effective at preventing COVID-19, and are important tools to stop the pandemic.
Imagine, everyone in your community getting a COVID-19 vaccine. You are safer and healthier. And even more importantly, your community of loved ones is safe and healthy.	You're fully vaccinated 2 weeks after getting your last COVID19 vaccine. Get vaccinated as soon as you can so you can get back to doing the things you love. Get vaccinated.

argument strength [95] measures the message's ability to evoke thoughts consistent with the desired outcome; the scale was composed of 9 items (ex. "The message gave a reason for getting a COVID vaccine that is important to me.") and participants rated the statements on a 5-point scale ("very weak" to "very strong") (M = 3.74, SD = 1.02, alpha = .87). Post-exposure attitudes [95] measure the positivity versus negativity of the respondent's opinion toward vaccination after reading the message; this was assessed with a 9-point semantic differential scale with 4 items (ex. "beneficial/harmful") (M = 2.45, SD = 2.30, alpha = .98).

5.2 Results & Discussion

Our hypotheses predicted that participants would perceive CDC messages to be of higher (a) argument strength, (b) perceived effectiveness, and (c) evoke more positive attitudes towards vaccination than GPT-3 messages. We conducted one-way analysis of variance (ANOVA) tests. In contrast to this prediction, GPT-3 messages outperformed CDC messages on argument strength, F(1, 807) = 9.03, $p < 0.01$, perceived effectiveness, F(1, 820) = 4.43, $p < 0.05$, and in evoking more positive attitudes, F(1, 822) = 17.77, $p < 0.001$.

Regarding RQ1, results of two-way ANOVA tests suggested that vaccinated individuals rated the messages higher than unvaccinated individuals regardless of source (perceived strength, F(1, 807) = 354.37, $p < 0.001$; effectiveness, F(1, 820) = 169.38, $p < 0.001$; attitude, F(1, 822) = 782.33, $p < 0.001$). The interaction between message creator and vaccination status was not significant on message qualities (perceived strength, F(1, 807) = 2.28, ns; effectiveness, F(1, 820) = 2.01, ns; attitude, F(1, 822) = 782.33, $p < 0.001$), suggesting that GPT-3's advantage over CDC messages was similar for vaccinated and unvaccinated participants on message qualities. The interaction between message creator and vaccination status, however, was significant for attitudes toward the message, F(1, 822) = 8.62, $p < 0.01$. Unvaccinated people had a significantly lower attitude toward CDC messages than the rest of the conditions. Thus, the prediction that CDC messages would be more persuasive and effective than GPT-3 messages (H1) and the prediction that participants would have more positive attitudes from CDC messages relative to GPT-3 messages (H2) was not supported. Instead, the GPT-3 messages were evaluated more positively on each measure of message quality relative to the CDC messages, and this advantage for the GPT-3 messages was even greater for the attitudes of unvaccinated individuals.

RQ2 explores how demographic variables and political party preferences impact respondents' perceptions of message quality. We conducted six separate linear regression analyses for the CDC and GPT-3 conditions. Note that the demographic variables were highly correlated. Collinearity statistics indicated that age and education had high Variance Inflation Factor (VIF) scores; thus, the results must be interpreted with caution. As Table 5 shows, age, education, and political party all predicted the message quality variables in the same way, with older adults, more educated, and more

Table 4. Message ratings by creator and vaccination status

Creator	CDC				GPT-3			
Vaccination Status	Vaccinated		Unvaccinated		Vaccinated		Unvaccinated	
	M	SD	M	SD	M	SD	M	SD
Message Effectiveness	3.73	0.79	2.73	1.05	3.77	0.83	2.98	1.17
Argument Strength	4.07	0.70	2.76	1.05	4.17	0.67	3.05	1.22
Attitude	3.46	1.05	-0.39	2.51	3.61	0.94	0.50	2.61
N	284		123		266		135	

Table 5. Regression: Demographics and message ratings

DVs	Argument strength		Message effectiveness		Attitude	
	CDC	GPT-3	CDC	GPT-3	CDC	GPT-3
Demographics						
Age	.16**	.23***	.17**	.25***	.37**	.41***
Gender	.08	.04	.11*	.06	-.10	.11
Living area	.08	.12**	.01	.15***	-.05	.12
Education	.19***	.29***	.12*	.27***	.34***	.45***
Employment	.04	.02	.02	.01	.11	.08
Political orientation						
Political view	-.12***	-.09*	-.04	-.04	-.45***	-.35***
Political party	-.07*	-.09**	-.09**	-.10**	-.24***	-.26***
Adjusted R2	92.6%	92.1%	91.6%	91%	57.7%	62.9%

Democrat-leaning participants rating perceptions of the messages more highly. Importantly, the pattern of results across CDC and GPT-3 suggests that there were no differences by creator (see Table 5).

Overall our results suggest that GPT-3 is capable of writing quality pro-vaccination messages. The best performing GPT-3 messages selected from our prompt analysis outperformed CDC messages on key health persuasion criteria, including perceived argument strength and effectiveness, and evoked more positive attitudes towards vaccination; this advantage held across audience demographics.

To understand message factors which could be driving this observation, we used Linguistic Inquiry and Word Count (LIWC) [89] to explore linguistic differences between messages created by the CDC and GPT-3. Table 6 shows the results of the LIWC analysis. The comparison revealed that the CDC messages used significantly more dictionary words in the LIWC dictionary than the GPT-3 messages, while GPT-3 messages had more words related to risk and drive. With the small sample size of messages (N=10 per message creator), statistical power is limited. Table 6 also reveals that GPT-3 and CDC messages numerically differ in affiliation, achievement, power, authenticity, and tone in a pattern that suggests that GPT-3 messages were more colloquial in style and perceived as more authentic and positive. While remaining cognizant of the limitation of the small sample of messages, these data are consistent with previous research showing that conversational language is more effective in communicating crisis messages [28], and that factors affecting message persuasiveness include benevolence, linguistic appropriacy, logical argumentation, and trustworthiness [19].

Table 6. LIWC: CDC versus GPT-3 linguistic composition

	Condition	N	Mean	Std. Error Mean	t
WC	CDC	10	65.1	7.60	1.34
	GPT-3	10	49.2	9.14	
Analytic	CDC	10	79.82	7.54	.56
	GPT-3	10	73.71	7.97	
Clout	CDC	10	73.28	6.28	-.80
	GPT-3	10	81.06	7.50	
Authentic	CDC	10	11.57	4.08	-.82
	GPT-3	10	19.32	8.50	
Tone	CDC	10	34.11	9.97	-.24
	GPT-3	10	38.20	13.69	
WPS	CDC	10	16.13	1.38	0.73
	GPT-3	10	14.54	1.69	
Sixltr	CDC	10	26.24	1.98	1.12
	GPT-3	10	22.88	2.26	
Dic	CDC	10	69.45	2.83	-3.05**
	GPT-3	10	81.29	2.65	
drives	CDC	10	8.17	1.39	-2.8*
	GPT-3	10	14.93	1.97	
affiliation	CDC	10	1.80	0.52	-1.25
	GPT-3	10	3.49	1.24	
achieve	CDC	10	0.55	0.25	-1.73
	GPT-3	10	1.64	0.58	
power	CDC	10	1.00	0.30	-0.95
	GPT-3	10	1.92	0.92	
reward	CDC	10	3.37	1.10	-0.57
	GPT-3	10	4.13	0.75	
risk	CDC	10	1.88	0.60	-3.89**
	GPT-3	10	5.78	0.81	

The observed difference in composition supports these findings and could be contributing to GPT-3's success. Further research on a larger sample of messages is required to more specifically isolate and analyze the difference between GPT-3 and CDC messaging.

These differences in the linguistic composition of GPT-3 and CDC messages suggest that GPT-3 is effectively able to generate content that resonates with public audiences. This may be explained by its ability to broadly draw upon and synthesize data directly from these audiences, including more casual forms of language, such as language data from the Internet, to generate content that is less formal and more casual than the CDC messages. AI models that are trained on trillions of words taken from the Internet may be better able to capture and leverage linguistic and thematic trends that appeal to a broader set of audiences by mimicking popular content, including their structure and styling.

As expected, our results also demonstrate that political beliefs, age, and education are key demographic differences related to how favorably individuals rate messages. These differences are consistent across both the CDC condition and the GPT-3 condition, suggesting that participants' demographic characteristics were not driving the finding that the GPT-3 messages outperformed the CDC messages.

6 STUDY 3: MESSAGE SOURCE AND PERCEIVED QUALITY

Building upon Study 2, which presented messages to participants without source attribution, Study 3 examines the role of message source, a well-known factor that influences message perceptions. The source of a message, including a source's reputation, credibility, and expertise, significantly impacts how a message is perceived [23, 90]. This is also true in the case of vaccination campaigns, which is one reason that the CDC and prominent public health officials play a major role in vaccination messaging. Individuals cite doctors and the CDC among their preferred sources for reliable vaccination information during COVID-19 [65]. As an emerging technology, people tend to have a limited understanding of AI and generally hold a negative view of AI's role in human communication (e.g., [39, 42, 43]). Given that Study 2 examined how people rate AI-written and human-written vaccination messages without any source context, an important question is whether source labeling affects message interpretation for AI-generated messages. Here, we refer to "actual source" as the entity that created the message (e.g., the CDC or GPT-3), and to "labeled source" as the label displayed to participants indicating the source of the message.

Prior research suggests that doctors and the CDC are among the most trusted health information sources during the COVID-19 pandemic [29, 65]. Thus, we assessed ratings of CDC and GPT-3 created messages with a manipulated source label displayed: 1) doctor, 2) CDC, 3) Artificial Intelligence, or 4) no source label (as the control). The messages were the same as those used in Study 2 and were created by either the CDC or GPT-3.

Our first objective for Study 3 was to replicate the findings from Study 2 that showed that, contrary to our initial expectations, GPT-3-generated messages were rated as having stronger arguments, being more effective, and perceived more positively than CDC-written messages. A comparison of the control conditions across actual sources (CDC vs. GPT-3) represents a direct replication of Study 1.

Our next objective was to examine how actual source (CDC vs. GPT-3) and labeled source (CDC, GPT-3, doctor, control) influenced perceptions of argument strength, effectiveness, and attitude toward vaccination. Existing research demonstrates that source and source credibility significantly affect psychological and behavioral outcomes. In the context of AI news generation, AI credibility perceptions are positively related to individuals' social trust and public discussion frequency [50], as well as individuals' social media usage habits [38]. Relating to chatbots, perceptions of credibility are positively affected by literacy and user trust [83]. Such factors provide insight to psychological processes affecting AI perceptions, and these AI perceptions vary by individual preferences and circumstances. Given that doctors and the CDC are among the most trusted vaccination sources during the COVID-19 pandemic [65], we expected messages labeled as coming from the CDC and doctors to be perceived of higher quality than messages from AI. Thus, we hypothesize the following source effects:

> H3: Participants will rate messages from the CDC and doctor sources more highly on perceived message quality and attitude than the messages from AI sources.

> RQ1: Are there interaction effects between the creator of the messages and the source of the messages?

Finally, we are interested in individual differences in trust. The public's trust in public health organizations and communicators is essential to encouraging desired behavioral outcomes [19, 90]. In this study, we measure trust [87] associated with each source to compare trust in AI versus prominent public health communicators. We ask how message ratings differ based on source labels and the moderating role of trust.

> RQ2: How does trust moderate the message effects?

6.1 Method

The number of participants was determined through power analysis for an F test with an effect size of 0.2, power of 0.8, alpha of 0.05, numerator degrees of freedom of 10, and 8 groups. The resulting required minimum sample size was 416. Since the unvaccinated population was the population we were most interested in regarding the effects of the messaging, we calculated the effect size using the sample size of possible unvaccinated participants. Based on the results of Study 2, approximately one-third of the total population was not vaccinated; the resulting sample size is 416 * 3 = 1248 respondents. We recruited 1600 respondents anticipating removal of low quality data and respondents who failed the manipulation check of correctly identifying sources. The study was approved by the university IRB (IRB # 51574). After reviewing a consent form and giving their consent, participants started the study. Each participant was paid 75 cents for their participation.

6.1.1 Manipulation Check. We asked participants to write down the source of the message they reviewed as a manipulation check. One hundred and twenty participants failed to identify the correct source of the message and were excluded from the dataset.

6.1.2 Participants. We recruited mTurkers who completed more than 5000 hits and who had an approval rate of 95% and above to participate in our study. Sixteen hundred and thirty-three mTurkers participated. We excluded 120 participants who failed the manipulation check and 17 participants who completed the study in less than a minute. The final sample size was 1496 (49.4% female), 31.9% of the participants were between the age of 35-44, followed by 45-59 (27.5%), 25-34 (26.7%), 60-69 (8.1%), 18-24 (3%), and 70-79 (2.5%). The majority of the participants were White, followed by African American (9.2%), Asian (8.3%), Latinos (4.7%), and others (2.4%). 43.5% of the participants had a Bachelor's degree, followed by high school graduates (24%), Associate degree in college (14.6%), Master's degree (13.8%), Ph.D. or higher (2.3%), and others (1.7%).

6.1.3 Procedure. We used a 4 (labeled source: CDC, Doctor, GPT-3, control) by 2 (actual source: CDC vs. GPT-3) between-subjects experimental design. Participants were randomly assigned to review one message generated by GPT-3 or the CDC, and the message was presented along with one of the four labeled source conditions: "AI (Artificial Intelligence)", "CDC (Centers for Disease Control and Prevention)", "Doctor", or source was not mentioned as a control condition. The vaccination messages from Study 2 were used. These messages were slightly adapted to be entirely independent of source (e.g., links to additional CDC information and CDC hashtags were removed).

6.1.4 Measures. The scales for perceived message effectiveness, perceived argument strength, and attitudes used in Study 2 were again used as indicators of message quality. We also used an adapted measure of trustworthiness (based on Soh [87]) to measure trustworthiness with the following statements on 5-point scales ("strongly disagree" to "strongly agree"): "I trust CDC/doctors/AI", "I have confidence in CDC/doctors/AI", "It's safe to trust CDC/doctors/AI" (M = 3.60, SD = 1.05, alpha = .97).

6.2 Results & Discussion

We first examined whether the results from Study 2 were replicated, in which GPT-3 messages had higher ratings of argument strength, perceived message effectiveness, and more positive attitudes than CDC messages. We analyzed the data from conditions without source labels. The results from the present study replicated those from Study 2: GPT-3

Table 7. Message ratings by source and creator

Actual	CDC								GPT-3							
Labeled	CDC		Doctors		AI		None		CDC		Doctors		AI		None	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Msg. Eff.	3.59	0.87	3.45	0.86	3.45	0.93	3.42	1.01	3.50	1.04	3.47	0.96	3.28	0.95	3.61	0.93
Arg. Str.	3.82	0.94	3.70	0.86	3.63	0.95	3.60	1.05	3.81	1.01	3.79	0.94	3.60	0.92	3.93	0.95
Attitude	2.76	2.13	2.52	2.12	2.49	2.22	2.29	2.40	2.49	2.19	2.73	1.98	2.38	2.38	2.80	2.10

messages relative to CDC messages were once again perceived as having higher argument strength, $t(435) = -3.48$, $p < .001$, as more effective, $t(441) = -2.03$, $p < .05$, and as evoking more positive attitudes $t(440) = -2.36$, $p < .05$.

We then explored the impact of labeled source (the CDC, doctors, AI, and none) to test H3, which predicts that participants will rate messages from the CDC and doctor sources more highly than the messages from AI sources. The results revealed that there was a significant effect of labeled source for argument strength, $F(3, 1461) = 2.85$, $p < .05$, and message effectiveness, $F(3, 1486) = 2.70$, $p < .05$. AI-labeled messages were perceived less positively compared to CDC-labeled messages [argument strength: $t(709) = -2.77$, $p < .01$; message effectiveness: $t(722) = -2.56$, $p < .05$] and messages without labels [argument strength: $t(801) = -2.15$, $p < .05$; message effectiveness: $t(810) = -2.23$, $p < .05$]. These data partially confirm H3. Messages labeled with the CDC as the source were perceived more positively than AI sources and no source, although the doctor source was not more effective than AI or the control sources.

Regarding RQ1, the interaction between actual source and labeled source was significant [argument strength: $F(3, 1461) = 3.10$, $p < .05$; message effectiveness $F(3, 1461) = 2.68$, $p < .05$]. GPT-3-created messages without a label were rated as having higher quality than CDC-created messages with an AI label [argument strength; $t(405) = 3.17$, $p < .01$; message effectiveness, $t(396) = 3.52$, $p < .001$] or without a label [argument strength $t(435) = 3.48$, $p < .001$] and GPT-3-created messages with an AI label [argument strength $t(392) = 3.50$, $p < .001$]. Regarding attitudes, participants' ratings are similar across all conditions. As such, H3 was partially supported. Overall, participants rated GPT-3-created messages higher than CDC-created messages, except when the message was actually labeled as from an AI source. Participants also rated messages labeled as from the CDC higher than messages labeled from GPT-3. That is, our participants displayed a preference for AI written messages but a dispreference for messages labeled as sourced from AI (see Table 7).

RQ2 explored a possible moderating role of trustworthiness. We first conducted a one-way ANOVA to explore people's trust towards the CDC, doctors, and AI. The results showed that people trust doctors ($M = 3.98$, $SE = .05$) significantly more than they trust the CDC ($M = 3.77$, $SE = .06$) and AI ($M = 3.10$, $SE = .05$), while they trust AI significantly less than the CDC and the doctors [$F(2, 1048) = 76.51$, $p < .001$]. Refer to Table 8 for message ratings.

We added trustworthiness as a covariate and ran two-way (source by content creator) ANOVA for each of the outcome variables. For each of the variables, the main effect of source was significant (all p's < .001) [argument strength: $F(2, 1024) = 11.29$; message effectiveness: $F(2, 1043) = 14.12$; attitudes: $F(2, 1034) = 69.88$], with AI-labeled source messages rated as having higher perceived strength, higher message effectiveness, and the most positive attitude toward vaccination relative to CDC or doctor-labeled source messages (all p's < .001). This pattern of results indicates that a lack of trust in AI as a source can explain why our participants had a dispreference for AI-labeled messages.

Overall, Study 3 replicated the findings from Study 2, finding that GPT-3 messages again received higher ratings than CDC messages for message effectiveness, argument strength, and attitude positivity. Study 3 also revealed the impact

Table 8. Message ratings with trustworthiness as a covariate

Labeled Source	Doctors		CDC		AI	
	M	SE	M	SE	M	SE
Trust	3.98	.05	3.77	.06	3.10	.05
Message Effectiveness	3.27	.05	3.46	.04	3.61	.04
Argument Strength	3.52	.04	3.71	.04	3.91	.04
Attitude	2.14	.10	2.41	.10	3.07	.10

of source on message ratings–messages from human sources, doctors, followed by the CDC, received higher ratings. Despite the effectiveness of AI-generated messages, AI-labeled messages received lower ratings. However, independent of source labels, AI-generated messages achieved higher ratings (Studies 2 and 3), and when controlling for trust, messages with AI as the source received higher ratings. This finding suggests that respondents interpreted AI-labeled messages to be of substantive quality but assigned lower ratings to messages with AI labeled as the source due to lower trust in AI for vaccination information. This relationship between source factors and message interpretation is consistent with previous findings [43, 90].

As AI becomes increasingly integrated with communications efforts, organizations can disseminate AI-generated messages either with themselves or AI cited as the source. Practitioners and scholars will need to determine how to ascribe ownership and responsibility for these messages. Practitioners should also determine context-specific trusted sources for releasing messages, as optimal sources will vary by campaign and target audiences. These results may change as individuals become more familiar with AI technologies and these technologies improve in performance, in which case individuals may gain trust in AI [85].

7 GENERAL DISCUSSION

Together, these three studies provide an exploratory analysis of AI in public health message generation, characterizing AI output (Study 1) and examining human perceptions to human-authored and AI-generated content without and with source labels (Studies 2 and 3). These three studies applied GPT-3, a state-of-the-art LLM, in the context of public health messaging, demonstrating that such technology can generate high-quality persuasive content. Furthermore, we assessed how online users perceive such content–GPT-3's most effective output was able to outperform CDC messages. These studies 1) highlight the potential contribution of using GPT-3 in public health communication, 2) provide methods for assessing the persuasive quality of model-generated messages, 3) suggest that the displayed source will affect how they are perceived, and 4) provide guidance for public health communication practitioners on how LLMs may augment their messaging workflows. Our findings demonstrate that models such as GPT-3 can be useful tools in augmenting content generation, but given the model's inconsistent message generation quality, only under systematic human review.

In addition to contributions to public health strategy, this research also supports the larger body of human-AI research. Gaps identified by Lai et. al. [48] include the misalignment between tasks assessing AI capabilities and realistic tasks for usage of AI tools, especially for deep learning models. Furthermore, previous literature primarily focuses upon communication on one-on-one contexts, such as chatbot interactions, rather than development of content for more general audiences. This series of studies addresses these gaps by evaluating LLM performance in communication workflows through a practical, real-world application, providing insight to the performance of a state-of-the-art deep learning model and assessing perception from a mass communication perspective.

7.1 Potential for AI in Public Health Messaging

GPT-3's performance in comparison to CDC content reveals that LLMs have the potential to support existing creative processes and content development workflows in collaboration with human practitioners. With the unique ability to draw upon massive amounts of textual data, LLMs are able to synthesize extensive, diverse sources in content generation almost instantly in contrast to existing workflows, which require extensive, expensive consumer research processes [61]. This synthesis enables such models to potentially expedite this consumer research process and develop content reflecting public language and sentiments. Furthermore, this access to data allows GPT-3 to linguistically mimic data from public discourse, which our language results from Study 2 suggest should allow it to generate content in a conversational tone that resonates with audiences. GPT-3 can be adopted to mediate public health messaging by quickly generating message drafts which practitioners then refine for accuracy, relevance, and the appropriate call to action. Examining generated prompts may provide practitioners with valuable examples of potential messaging strategies, aiding the creative development process.

Although prompt development can take a significant amount of time, once effective prompts have been identified, GPT-3 can easily generate a large number of unique messages. Furthermore, prompts can be developed for a variety of tasks, such as creating more targeted messages. Although the idea of creating and disseminating personalized messages at scale is counterintuitive, such content-generation technology would allow communication practitioners to create diverse content, and they can tailor this content toward individuals [15, 92]. As GPT-3 and similar models for content generation continue to improve, they will enable communication practitioners in creating targeted messages at scale. Thus, future research can explore GPT-3's ability to generate targeted content and the effectiveness of targeted messages.

Our results highlight, however, that substantial human review is required for LLMs to produce high quality and effective messages. GPT-3's message generation was inconsistent, indicating that an automation approach for LLMs in public health communication is unlikely to be successful. Instead, our pattern of results argues for an augmentation approach for LLMs in this context, with an interactive process involving human review.

Indeed, with LLMs' capacity for prolific copy generation, misinformation and disinformation are a significant concern. Because GPT-3 was trained on information from the internet, biased and incorrect information is part of the model's training base [1]. Research demonstrates that larger models achieve lower levels of output accuracy, tending to mimic common misconceptions [37, 53]. NaturalQS, a performance evaluation dataset testing the model's ability to answer questions such as those entered in search engines, is an indicator of factual accuracy. On this assessment, GPT-3 underperforms in comparison to RAG, another recently developed model which relies on fine-tuning [52], with the models achieving scores of 29.9 and 44.5 respectively [46]. This production of factually inaccurate text raises serious concerns of misinformation generation [25]. In fact, OpenAI notes their concern of GPT-3 being used for malicious purposes [74] and prior research demonstrates GPT-3's ability to create disinformation at scale [9]. Taken together with the results from Study 1, research to date highlights an augmentation approach for LLMs in public health messaging, and reiterates the necessity for human review of content created by GPT-3.

With appropriate human review, future LLMs can strengthen model performance by refining training data, ensuring that the data is high quality, up-to-date, and does not reflect systemic biases, as the training data directly contributes to model output. Specifically, by developing models that are able to draw upon more current data (rather than data from 2016-2019), AI models can better capture and leverage linguistic and thematic trends (e.g., mimicking a popular content structure, phrase, or meme) to appeal to specific audiences. With the transition to short format video in social media, rich content will become available in non-verbal formats, such as TikTok videos (which are dominated by trends), future models can be developed and trained on multimodal data sources. These new models could augment public health messaging by generating novel content with more current data on linguistic trends.

7.2 Ethical considerations and implications

Future applications for LLM augmentation in messaging (and specifically, public health messaging) should be analyzed through both practical and ethical lenses. Are models able to create accurate and appropriate content that meaningfully appeals to a specific recipient? The release of inappropriate content in such a high-stakes context could contribute to misinformation, confusion, or offense, with severe consequences at a societal and national level. Another important question that emerges from an ethical lens is as messages become more personalized, at what point is the messaging persuasive versus manipulative? While public health practitioners seek to encourage positive health outcomes, it is important for content to remain factual and truthful. While professionals distinguish between persuasive and manipulative or misleading messaging, AI has not been developed to distinguish between these modes and thus may cross the line from persuasion to manipulation.

LLMs can also be explicitly misused to create disinformation campaigns with cheaply developed, personalized content. Malicious content creation can include manipulation of niche audience segments through targeted campaigns, harassment of individuals [5], or targeting based upon stereotypes or offensive language. Divisive misuse could contribute to future societal wedging and intercultural tensions [9]. As public health officials leveraging LLMs in communication seek to minimize the risk of misinformation and malicious messaging, important guidelines for ethical content development and publication will rely upon systematic human review for the foreseeable future. Human review will require practitioners who are thoroughly skilled in multicultural communication and trained to understand the ethical dimensions concerning the AI technology, allowing them to better anticipate and correct problematic output.

Indeed, training that includes ethical considerations related to AI implementation and adaptation will be critical. Challenges in AI include those related to equity and bias (e.g., [16, 17, 33]), user privacy (e.g., [6]), algorithmic transparency and explainability ([32, 51]), information and misinformation [9], and the implications of these systems online and offline (e.g., [64]). Furthermore, practitioners face ethical challenges specifically for AI in healthcare and health communication, especially given its high-stakes nature (e.g., [14, 22, 68, 69]. Developing such training for public health communication regarding ethical use and potential misuse, given the important role our results suggest for human oversight of AI, are of paramount importance.

7.3 Limitations

The present research suffers from several important limitations. First, GPT-3's performance was assessed for a topic it was not trained on; due to its training data (internet data from 2016-2019), it was unfamiliar with COVID-19-specific information. While this provides insight for how such models can perform when writing about unknown topics, future research can assess how GPT-3 and human-authored content compare when discussing a topic within the scope of its training. Second, this research focused solely on how unaltered GPT-3 output compared to CDC messages. Research points to the importance of human editing and refinement of drafts by GPT-3 in a more iterative and collaborative process [9, 25]. Further research can assess how human-AI collaboratively generated content is perceived in comparison to purely human-written or model-generated content. Third, this research was focused on purely textual messages. Future research can explore the performance of model-generated content when leveraged in other forms of communication, such as captioning or discussing images or videos in a public health campaign. Finally, this research analyzed single assessments of vaccine messages well after COVID-19 messaging conventions were established. Because raters had likely been exposed to many messages by the time they rated these messages, their responses reflect perceptions of message congruency related to vaccine message expectations. Future research could analyze the impact of official messages versus model-generated messages at other phases in a crisis' development, such as at the onset, providing insight into how artificial intelligence could shift the trajectory of subsequent messaging.

Our record of GPT-3 prompts, request parameters, and the corresponding completions proved useful in analysis of GPT-3's output. In future research assessing generative outputs of language models, we recommend keeping a record of the prompts, request parameters, and corresponding completion or completions (if multiple iterations). Such documentation will provide clarity to the relationship between prompts and model output, allowing further analysis of the body as well as the holistic quality of the output; these records will be helpful in future prompt design and parameter optimization efforts.

8 CONCLUSION

These findings suggest that GPT-3 and similar natural language models have the potential to augment public health communication, such as COVID-19 vaccination, under conditions of human review. We observed that GPT-3 produces messages of inconsistent quality, but that the best of those messages can be leveraged to create high quality public health messages that individuals perceive as effective and persuasive, even relative to CDC messaging, and this effect holds across demographic groups. We also found that the perceived source of a message matters, with participants displaying a dispreference for AI-labeled public health messages, suggesting that how messages are framed when AI is involved in message development will be important. Taken together, our results across three studies suggest that, with human oversight, AI has the potential to contribute to public health communication workflows to effectively and efficiently develop strategic communication campaigns.

9 ACKNOWLEDGMENTS

This research was supported by the National Science Foundation, award AW940028. We thank Hannah Mieczkowski, Maurice Jakesch, Mor Naaman, and members of the Stanford Social Media Lab for providing insightful feedback, Taylor Gindlesperger for contributing to the coding of messages, and Charlie Ekstrom for contributing to literature collection.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* 3, 6 (2021), 461–463.

[2] Sam Addison Ankenbauer and Alex Jiahong Lu. 2021. Navigating the "Glimmer of Hope": Challenges and Resilience among US Older Adults in Seeking COVID-19 Vaccination. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 10–13.

[3] Kenneth C Arnold, April M Volzer, and Noah G Madrid. 2021. Generative Models can Help Writers without Writing for Them.. In *IUI Workshops*.

[4] Robert A Bell, Matthew S McGlone, and Marko Dragojevic. 2014. Vicious viruses and vigilant vaccines: Effects of linguistic agency assignment in health policy advocacy. *Journal of health communication* 19, 10 (2014), 1178–1195.

[5] 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).

[6] Anne Bowser, Katie Shilton, Jenny Preece, and Elizabeth Warrick. 2017. Accounting for privacy in citizen science: Ethical research in a context of openness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2124–2136.

[7] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. (2007).

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[9] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, Lies, and Automation. (2021).

[10] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[11] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[12] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All That's' Human'Is Not Gold: Evaluating Human Evaluation of Generated Text. *arXiv preprint arXiv:2107.00061* (2021).

[13] Scott Counts, Munmun De Choudhury, Jana Diesner, Eric Gilbert, Marta Gonzalez, Brian Keegan, Mor Naaman, and Hanna Wallach. 2014. Computational social science: Cscw in the social media era. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 105–108.

[14] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6, 2 (2019), 94.

[15] Ashok Deb, Anuja Majmundar, Sungyong Seo, Akira Matsui, Rajat Tandon, Shen Yan, Jon-Patrick Allem, and Emilio Ferrara. 2018. Social bots for online public health interventions. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1–4.

[16] Michael A Devito, Ashley Marie Walker, Jeremy Birnholtz, Kathryn Ringland, Kathryn Macapagal, Ashley Kraus, Sean Munson, Calvin Liang, and Herman Saksono. 2019. Social technologies for digital wellbeing among marginalized communities. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 449–454.

[17] Mark Diaz. 2019. Algorithmic Technologies and Underrepresented Populations. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 47–51.

[18] Amiel A Dror, Netanel Eisenbach, Shahar Taiber, Nicole G Morozov, Matti Mizrachi, Asaf Zigron, Samer Srouji, and Eyal Sela. 2020. Vaccine hesitancy: the next challenge in the fight against COVID-19. *European journal of epidemiology* 35, 8 (2020), 775–779.

[19] Sebastian Duerr and Peter A Gloor. 2021. Persuasive Natural Language Generation–A Literature Review. *arXiv preprint arXiv:2101.05786* (2021).

[20] The Economist. [n. d.]. A New AI Language Model Generates Poetry and Prose. https://www.economist.com/science-and-technology/2020/08/06/a-new-ai-language-model-generates-poetry-and-prose..

[21] J Richard Eiser and Camilla J White. 1974. Evaluative consistency and social judgment. *Journal of Personality and Social Psychology* 30, 3 (1974), 349.

[22] Pouyan Esmaeilzadeh. 2020. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC medical informatics and decision making* 20, 1 (2020), 1–19.

[23] Jim AC Everett, Clara Colombatto, Vladimir Chituc, William J Brady, and Molly Crockett. 2020. The effectiveness of moral messages on public health behavioral intentions during the COVID-19 pandemic. (2020).

[24] Matthew C Farrelly, James Nonnemaker, Kevin C Davis, and Altijani Hussin. 2009. The influence of the national truth® campaign on smoking initiation. *American journal of preventive medicine* 36, 5 (2009), 379–384.

[25] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30, 4 (2020), 681–694.

[26] Brian J Fogg. 1998. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 225–232.

[27] Centers for Disease Control and Prevention. [n. d.]. Workplace Health in America 2017.

[28] Karen Freberg, Kristin Saling, Kathleen G Vidoloff, and Gina Eosco. 2013. Using value modeling to evaluate social media messages: The case of Hurricane Irene. *Public Relations Review* 39, 3 (2013), 185–192.

[29] Ilona Fridman, Nicole Lucas, Debra Henke, and Christina K Zigler. 2020. Association between public knowledge about COVID-19, trust in information sources, and adherence to social distancing: cross-sectional survey. *JMIR public health and surveillance* 6, 3 (2020), e22060.

[30] CARY Funk and JOHN Gramlich. 2021. 10 facts about Americans and coronavirus vaccines. *Pew Research Center* (2021).

[31] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.

[32] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[33] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[34] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[35] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1589–1599.

[36] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.

[37] Will Douglas Heaven. 2020. OpenAI's new language generator GPT-3 is shockingly good—and completely mindless. *MIT Technology Review* (2020).

[38] Lennart Hofeditz, Milad Mirbabaie, Stefan Stieglitz, and Jasmin Holstein. 2021. DO YOU TRUST AN AI-JOURNALIST? A CREDIBILITY ANALYSIS OF NEWS CONTENT WITH AI-AUTHORSHIP. (2021).

[39] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[40] Joo-Wha Hong and Nathaniel Ming Curran. 2019. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 2s (2019), 1–16.

[41] Li Huang, Zhi Lu, and Priyali Rajagopal. 2022. Numbers, not lives: Ai dehumanization undermines covid-19 preventive intentions. *Journal of the Association for Consumer Research* 7, 1 (2022), 63–71.

[42] Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 18–33.

[43] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Tae Woo Kim and Adam Duhachek. 2020. Artificial intelligence and persuasion: A construal-level account. *Psychological science* 31, 4 (2020), 363–380.

[45] Andy J King and Allison J Lazard. 2020. Advancing visual health communication research to improve infodemic response. *Health Communication* 35, 14 (2020), 1723–1728.

[46] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[47] Mary E Laffidy. 2021. *Governments' Use of Fear Appeals and Accessible Language in COVID-19 Outreach: Comparing Florida and New York's Messages, March 2020–September 2020*. Ph. D. Dissertation. Northern Arizona University.

[48] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[49] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.

[50] Sangwon Lee, Seungahn Nah, Deborah S Chung, and Junghwan Kim. 2020. Predicting ai news credibility: communicative or social capital or both? *Communication Studies* 71, 3 (2020), 428–447.

[51] Stephan Lewandowsky and Anastasia Kozyreva. 2022. Algorithms, lies, and social media. https://www.niemanlab.org/2022/04/algorithms-lies-and-social-media/.

[52] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401* (2020).

[53] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[54] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021).

[55] Edward W Maibach and Roxanne Parrott. 1995. *Designing health messages: Approaches from communication theory and public health practice*. Sage.

[56] Cade Metz. [n. d.]. Meet GPT-3. It Has Learned to Code (and Blog and Argue).

[57] Hannah Mieczkowski, Jeffrey T Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–14.

[58] Marie-Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation* 9, 1 (2018), 1–14.

[59] Thomas Moore. [n. d.]. HHS plans mega $250m 'defeat despair' COVID-19 campaign.

[60] Xiaoli Nan, Irina A Iles, Bo Yang, and Zexin Ma. 2022. Public health messaging during the COVID-19 pandemic and beyond: Lessons from communication science. *Health Communication* 37, 1 (2022), 1–19.

[61] Krista Neher. [n. d.]. How To Create A Digital Marketing Strategy: Eight Steps To Laser Focus Your Plan.

[62] Kimberly H Nguyen, David Yankey, Kelsey C Coy, Kathryn A Brookmeyer, Neetu Abad, Rebecca Guerin, Girija Syamlal, Peng-jun Lu, Brittney N Baack, Hilda Razzaghi, et al. 2021. COVID-19 Vaccination Coverage, Intent, Knowledge, Attitudes, and Beliefs among Essential Workers, United States. *Emerging infectious diseases* 27, 11 (2021), 2908.

[63] Jeff Niederdeppe, Matthew C Farrelly, James Nonnemaker, Kevin C Davis, and Lauren Wagner. 2011. Socioeconomic variation in recall and perceived effectiveness of campaign advertisements to promote smoking cessation. *Social science & medicine* 72, 5 (2011), 773–780.

[64] Fayika Farhat Nova, Pratyasha Saha, Md Shafiur Raihan Shafi, and Shion Guha. 2019. Sharing Of Public Harassment Experiences on Social Media in Bangladesh. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 324–329.

[65] Tennessee Department of Health. 2021. Vaccine Messaging Market Survey Executive Summary Reports. https://www.tn.gov/health/news/2021/4/14/market-study-explores-tennessean-s-perspectives-on-covid-19-vaccine.html.

[66] Harri Oinas-Kukkonen and Marja Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems* 24, 1 (2009), 28.

[67] World Health Organization. [n. d.]. Who principles for effective communications. https://www.who.int/about/communications/principles.

[68] Kirsten Ostherr. 2020. Artificial intelligence and medical humanities. *Journal of Medical Humanities* (2020), 1–22.

[69] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S Lasecki. 2019. Identifying challenges and opportunities in human-AI collaboration in healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 506–510.

[70] Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger Hurts but Starving Works" Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1185–1200.

[71] Richard E Petty, Jamie Barden, and S Christian Wheeler. 2009. The Elaboration Likelihood Model of persuasion: Developing health promotions for sustained behavioral change. (2009).

[72] Ashley Pilipiszyn. 2021. GPT-3 powers the next generation of apps. https://openai.com/blog/gpt-3-apps/.

[73] Jane Pirkis, Alyssia Rossetto, Angela Nicholas, Maria Ftanou, Jo Robinson, and Nicola Reavley. 2019. Suicide prevention media campaigns: a systematic literature review. *Health communication* 34, 4 (2019), 402–414.

[74] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. OpenAI Blog https://openai. com/blog/better-language-models.

[75] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446* (2021).

[76] Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.

[77] Malik Sallam. 2021. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines* 9, 2 (2021), 160.

[78] Charles Harold Sandage and Vernon Fryburger. 1989. Advertising Theory and Practice. (1989).

[79] Shaunak Sastry and Alessandro Lovari. 2017. Communicating the ontological narrative of Ebola: An emerging disease in the time of "epidemic 2.0". *Health communication* 32, 3 (2017), 329–338.

[80] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 153–157.

[81] Kavya Sekar. 2020. Funding for COVID-19 Vaccines: An Overview.

[82] Bryan C Semaan, Scott P Robertson, Sara Douglas, and Misa Maruyama. 2014. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1409–1421.

[83] Donghee Shin. 2021. How do people judge the credibility of algorithmic sources? *AI & SOCIETY* (2021), 1–16.

[84] Donghoon Shin, Subeen Park, Esther Hehsun Kim, Soomin Kim, Jinwook Seo, and Hwajung Hong. 2022. Exploring the Effects of AI-assisted Emotional Support Processes in Online Mental Health Community. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[85] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.

[86] Michael D Slater and June A Flora. 1991. Health lifestyles: Audience segmentation analysis for public health interventions. *Health education quarterly* 18, 2 (1991), 221–233.

[87] Hyeonjin Soh. 2006. *Measuring trust in advertising*. Ph. D. Dissertation. uga.

[88] Patric R Spence, Chad Edwards, Autumn Edwards, Adam Rainear, and Xianlin Jin. 2021. "They're always wrong anyway": exploring differences of credibility, attraction, and behavioral intentions in professional, amateur, and robotic-delivered weather forecasts. *Communication Quarterly* 69, 1 (2021), 67–86.

[89] 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.

[90] Marianne Udow-Phillips and Paula M Lantz. 2020. Trust in public health is essential amid the COVID-19 pandemic. *Journal of Hospital Medicine* 15, 7 (2020), 431–433.

[91] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language* 67 (2021), 101151.

[92] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725* (2019).

[93] Morton Wiener and Albert Mehrabian. 1968. *Language within language: Immediacy, a channel in verbal communication*. Ardent Media.

[94] Ben Young, Sarah Lewis, Srinivasa Vittal Katikireddi, Linda Bauld, Martine Stead, Kathryn Angus, Mhairi Campbell, Shona Hilton, James Thomas, Kate Hinds, et al. 2018. Effectiveness of mass media campaigns to reduce alcohol consumption and harm: a systematic review. *Alcohol and alcoholism* 53, 3 (2018), 302–316.

[95] Xiaoquan Zhao, Andrew Strasser, Joseph N Cappella, Caryn Lerman, and Martin Fishbein. 2011. A measure of perceived argument strength: Reliability and validity. *Communication methods and measures* 5, 1 (2011), 48–75.

A SETTING REQUEST PARAMETERS

GPT-3 request parameters include: engine (for API requests; different engines have different capabilities), response length (maximum number of tokens, or words), temperature (amount of randomness in completions), top P (number of alternatives considered), frequency penalty (penalizes token repetition), presence penalty (penalizes if token previously appears), and best of (selects the best of multiple completions). Tokens are commonly occurring subsets of words, averaging about 4 characters in length. To set the parameters, we began with OpenAI's default values for each parameter and adjusted them based on qualitative observation of the output. The final request parameters for our prompts were: response length as 135 (to control for message length), and temperature as 0.8 [76], frequency penalty as 0.5; other request parameters remained at their default values.

B STIMULI MESSAGES

B.1 Curated GPT-3-Generated Messages

- The most important thing you can do to protect yourself and those around you from the coronavirus is to get the vaccine as soon as possible. The sooner you get vaccinated, the sooner you can start to do things that you may not have been able to do before—like travel, attend work or school, go out in public, visit friends or family, etc. The best time to get vaccinated is now!
- The COVID-19 virus can cause serious illness, disability, or death. To protect your loved ones, get the vaccine.
- COVID-19 is a serious threat to our health. The best way to protect the health of you and your community is to get vaccinated.
- If you get a coronavirus vaccine, you will be taking a step to protect yourself from the coronavirus. The vaccines have been proven safe and effective. Vaccines work by spurring your body's immune system into action. Your immune system protects your body from germs that can make you sick. After getting vaccinated against the coronavirus, your immune system will start making antibodies to fight it off. If you ever come in contact with the coronavirus, your antibodies will go to work right away, so that you don't get sick.
- Your voice matters: Be a part of the effort to stop this outbreak. Don't become ill with the coronavirus or put others at risk. Get a coronavirus vaccine today.
- Be the first to vaccinate. When you get vaccinated, you're helping to protect your community and your family. The vaccine may help protect people around you who don't get a vaccine. What you do makes a difference. Get vaccinated now.
- Imagine, everyone in your community getting a COVID-19 vaccine. You are safer and healthier. And even more importantly, your community of loved ones is safe and healthy.
- Like most other vaccines, coronavirus vaccines are designed to stimulate the body's immune system. The immune system is the body's natural defence against disease and infection that prevents these from becoming a problem. The coronavirus vaccine is designed to protect people who have never been exposed to the virus before while also providing protection for those who have been vaccinated in the past but are again at risk. The coronavirus vaccine is an immunization that protects the carrier from contracting a virus known as coronavirus or from spreading it to others. Coronavirus is a virus that causes serious infections, including severe pneumonia and sometimes even death.
- The most important thing you can do to help stop the spread of COVID-19 is to get vaccinated against COVID-19. Vaccines are the safest and most effective way to prevent COVID-19.

- While no vaccine can guarantee complete protection, CDC and FDA continue to monitor the safety of all COVID-19 vaccines. If you are vaccinated, you can decrease your risk of getting sick with COVID-19.

B.2 Sample CDC Messages

- Two of the COVID-19 vaccines authorized for use in the United States use mRNA. mRNA COVID-19 vaccines teach our cells how to make a piece of a protein to trigger an immune response and build immunity to the virus that causes COVID19. mRNA does not affect or interact with a person's DNA and the cell breaks down and gets rid of the mRNA as soon as it is finished using the instructions.
- You may have symptoms like a fever after you get a COVID19 vaccine. This is normal and a sign that your immune system is learning how to recognize and fight the virus that causes COVID-19.
- How many people need to get a COVID19 vaccine for population immunity? When enough people in a community are protected from getting a disease – because they've already had the disease or they've been vaccinated – that makes it harder for the disease to spread from person to person. This is known as population immunity, and it even protects those who cannot be vaccinated, like newborns. While experts don't yet know what percentage of people would need to get vaccinated to achieve population immunity, vaccination is a safer way to build protection than getting sick with COVID-19.
- While getting COVID-19 may offer some natural protection or immunity, the risk of severe illness and death from COVID-19 far outweighs any benefits of natural immunity. Getting a COVID-19 vaccine will help protect you without having to be sick.
- COVID-19 vaccination helps keep you from getting COVID-19. The vaccines currently available in the United States are effective at preventing COVID-19, and are important tools to stop the pandemic.
- As of July 26, 2021, more than 188.7 million people, or about 56.8% of the U.S. population, have received at least one dose of a #COVID19 vaccine. Of those, 163.2 million, or about half of the population, are fully vaccinated. COVID-19 vaccines are safe and effective at preventing COVID-19, especially severe illness and death. For more COVID-19 vaccination data, visit http://bit.ly/CDT_vaccine.
- You're fully vaccinated 2 weeks after getting your last #COVID19 vaccine. Get vaccinated as soon as you can so you can get back to doing the things you love. Get vaccinated.
- The Delta variant is spreading rapidly in unvaccinated populations. Don't let Delta and other variants stop our progress in the fight against #COVID19. Get vaccinated and do your part to put COVID-19 in the past.
- COVID19 cases, hospitalizations, and deaths are once again going up across the United States. This is due, in part, to the spread of the B.1.617.2 (Delta) variant, accounting for more than 80% of COVID-19 cases. With highly effective vaccines, COVID-19 is now a preventable disease. Read more about COVID-19 trends, variants, and the importance of getting vaccinated in the COVID Data Tracker Weekly Review: http://bit.ly/CDTweeklyreview
- The amount of information—and misinformation—about #COVID19 vaccines can be overwhelming. Help your friends and family who have questions about the vaccines by listening to their concerns without judgment. Identify the root of their concerns. Acknowledge their emotions so they know they've been heard. More: https://bit.ly/talkvaccines.

Received January 2022; revised July 2022; accepted October 2022