Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering

Sue Lim* and Ralf Schmälzle

Department of Communication, Michigan State University, East Lansing, MI, United States

Introduction: This study introduces and examines the potential of an AI system to generate health awareness messages. The topic of folic acid, a vitamin that is critical during pregnancy, served as a test case.

Method: We used prompt engineering to generate awareness messages about folic acid and compared them to the most retweeted human-generated messages via human evaluation with an university sample and another sample comprising of young adult women. We also conducted computational text analysis to examine the similarities between the AI-generated messages and human generated tweets in terms of content and semantic structure.

Results: The results showed that AI-generated messages ranked higher in message quality and clarity across both samples. The computational analyses revealed that the AI generated messages were on par with human-generated ones in terms of sentiment, reading ease, and semantic content.

Discussion: Overall, these results demonstrate the potential of large language models for message generation. Theoretical, practical, and ethical implications are discussed.

KEYWORDS

health communication, message generation, artificial intelligence, prompt engineering, social media, folic acid (FA)

1. Introduction

Consider the following two messages: *"Every woman needs #folicacid every day, but even more women need it in their first trimester! Folic acid is an essential nutrient needed for the formation of new blood cells in the womb, as well as the developing nervous system,"* and *"The risk of neural tube defects is reduced if women consume 400 micrograms of #folate every day before and during early pregnancy."* Could you tell which of these messages was written by an AI system? We will resolve this question at the end of this paper, which will discuss the theoretical potential, practical, and ethical implications of AI for communication science, here focusing first on the generation of health awareness messages.[1]

---

1   Importantly, we anticipate that the scope and impact of AI in communication will go far beyond health awareness messaging, including AI-based health persuasion (focusing on attitude and behavior change), AI in political communication, advertising, as well as interpersonal domains. However, this paper focuses exclusively on the generation of brief messages intended to raise awareness; we view this as a high-impact and decidedly pro-social domain in which communication and AI can connect. Throughout the paper, we will emphasize that at this stage, influencing attitudes and behavior via AI-generated messaging systems is a possibility (and likely already ongoing), but this empirical research paper focuses only on the proximal goal of raising awareness.

In this paper, we first introduce the theoretical underpinnings of AI-based message generation. The following section explains the application context, health communication, and specifically health awareness campaigns on social network sites. The third section brings these two research streams together, discussing how AI language models can generate health awareness messages and introducing recent research in this area. We then present a series of studies with the hypotheses and research questions. Finally, we discuss the results and their implications.

1.1. AI message generation: theory, method, and relevance for communication

Theoretical breakthroughs in deep learning (DL), most of which happened between 2010 and 2020, have equipped computers with impressive capacities. Deep learning uses artificial neural networks that are stacked across multiple layers (hence: deep) and are trained to perform specific tasks (hence: learning). The tasks include automatic speech recognition (e.g., for voice dictation) and image recognition (e.g., labeling objects in photographs). DL-fueled applications are integrated in every cell phone and computer (Schmidhuber, 2015; Hassabis et al., 2017; Chollet, 2021). While these capacities were initially geared more toward perceptual and receptive tasks like speech and image recognition, the field soon expanded to include generative, productive tasks, such as text or image synthesis. For example, every text processing software now includes the ability to auto-complete words from their beginning or even suggest entire sentence continuations.

To provide readers with a non-technical intuition of the underlying principles, we ask you to consider the sentence "I take my coffee with cream and …." Most people will predict the next word as "sugar." Logically, given the semantic context set up by the preceding words, the word sugar is very likely—certainly far more likely than e.g., "socks." Language models (LMs) used for natural language processing (NLP) are powered by artificial neural networks trained to complete next-word prediction tasks like the example above. During the training process, large amounts of texts are gathered and fed into the neural networks, and the neural networks repetitively perform next-word-prediction by computing conditional probabilities that a word will come after another word (or sets of words). Once this training process is complete, the model will have learned the hidden statistical relationships between words in a large corpus of natural language (Hirschberg and Manning, 2015). When prompted with a text like "I take my coffee with cream and …," then for every word processed during the training process, the neural network will be able to calculate the probability that the word will come after the text. If trained well, the language model will assign the word "sugar" with the greatest likelihood of being the next word. This principle can be applied over multiple scales, from letters in words (like when a word starts with "th" it is likely that the next letter will be an "e") to words in sentences (like in the example above), and beyond.

A subfield of DL called generative deep learning (generative DL; Goodfellow et al., 2016; Chollet, 2021), takes this idea of next-word prediction to the next level by using pre-trained language models for text generation (Gatt and Krahmer, 2018; Keskar et al., 2019;

Tunstall et al., 2022). The texting apps and email programs that can suggest next-word candidates or autocompletion options for sentences are powered by such language models. One can enter a sequence of words in the system, and the LM will compute conditional probabilities of the next word, so that the texting or the email system will select the word with the highest probability to either recommend to the user or to autocomplete the text. Recent advancements in generative DL have led to so-called large language models (LLMs) that can move beyond simple next-word-prediction tasks to generate complex content that appears natural to humans (Rashkin et al., 2020; Tunstall et al., 2022; Bubeck et al., 2023; Liu et al., 2023). Examples of these highly proficient text generation systems include OpenAI's ChatGPT, along many others (e.g., Google's BARD and Huggingface's Bloom). Although many limitations still exist (Bender and Koller, 2020), recent advancements in this area are very relevant for communication science and have many theoretical implications.

First, LLMs' capability to generate messages[2] that closely mimic human natural language provides new opportunities for communication scientists to study principles of message generation. Message generation, as conceived here, consists of developing the prompts to feed to the LLM (similar to message production; Greene, 2013) and adjusting the statistical parameters of the LLM (mimicking the process of language production via the human neural system). Indeed, although we can rely on many principles and theories for communication message design (Miller et al., 1977; Witte et al., 2001; Cho, 2011), the act of coming up with a novel and grammatically correct message in natural language is still a largely creative endeavor (Gelernter, 2010) that humans accomplish via intuition (Hodgkinson et al., 2008). As such, we must acknowledge that there is a gap in our understanding of how humans naturally come up with new messages.[3] However, with LLMs, which are created to mimic natural human language, researchers can get a glimpse into the

---

2 We acknowledge that definitions of communication, message, and theory can vary, and there are many longstanding debates about what counts and what does not count as communication, message, or theory. Clearly, the field of NLP is closer to engineering and computer science and has to a large degree ignored communication science (e.g., Bender and Koller, 2020). Nevertheless, there have always been theoretical connections, like between computer-mediated communication and voice assistants, or using computational methods to analyze communication. The advent of text generation methods—although they still certainly lack many if not the most central characteristics of human communication (e.g., intent, goals, pragmatics, etc.)—is clearly bringing the fields into closer contact. As such, we believe that it will be in communication scientists' best interest to engage with this development, which is already influencing key aspects of communication in the 21st century.

3 To be clear, one could certainly view the entire literature on rhetoric as being focused on message production and delivery, like the notion of 'inventio' in Aristotle's theory, which can be considered the birthplace of communication science. Likewise, work on communication goals and strategic communication also have action- or production-oriented elements. However, these approaches are geared toward more abstract levels, and they are all silent about the actual nuts and bolts of message generation.

message generation process and are increasingly able to strategically steer this process via prompt engineering and related approaches (von Werra et al., 2020). Demonstrating this potential of human-aided prompt engineering in combination with LLM's capability for message generation is one of the key goals of this paper.

A second argument as to why AI-based message generation research is theoretically relevant is because theory and method evolve in a synergistic relationship (Greenwald, 2012). In particular, technological advances often promote new theory and stimulate theoretical advances, which then provide the basis for new methods. Applying this to the current context, we expect that the abilities to generate messages (see the paragraphs above) hold promise to expand our theoretical understanding of what has so far remained enigmatic, namely how new, coherent messages can be created. Within health communication, we have already seen how computational methods have been used to test and expand existing theories (Chen et al., 2020; Rains, 2020; Lee et al., 2023). In this sense, putting method and theory in opposition is—although common—actually misleading.[4] Instead, we expect that the nexus between AI, message communication, and message reception research will lead to synergistic benefits between theory and method.

Third, as methods promote theory and vice versa, we often see the emergence of entirely new fields of research. For instance, when computers came into contact with communication, the field of CMC arose, which in turn stipulated theory. As AI emerged into communication research, especially within the field of CMC, AI-mediated communication arose as a field (Hancock et al., 2020). Similarly, theory and methods from communication and neuroscience are converging, giving rise to the field of communication neuroscience (Huskey et al., 2020; Schmälzle and Meshi, 2020). Together, these three arguments underscore why the nexus of AI and communication is a theoretical wellspring:[3] DL-approaches for text generation help explain a henceforth mystic process and they provide us with a principled approach to generate thousands of novel messages, which we can analyze. This development has started to impact many fields of communication, including health communication.

Finally, LLMs' capability of generating hundreds of messages within a short period of time also provides new opportunities for communication researchers to test existing theories regarding message processing and effects. Within communication, a plethora of research regarding message effects (e.g., emotional appeals; Nabi et al., 2020; Turner and Rains, 2021, message tailoring; Noar et al., 2007), information processing (e.g, Elaboration Likelihood Model; Petty et al., 1986), and persuasion (e.g., source factors; O'Keefe, 2015, social norms; Lapinski and Rimal, 2005) exist. Thus, researchers can adjust the prompt texts according to formative research and/or existing theories to generate specific types of

messages, and then use computational and human evaluation methods to test the theories. There is already evidence of researchers leveraging LLMs for persuasive message generation within health communication (Karinshak et al., 2023; Zhou et al., 2023). While we do not test specific existing message effects theories in this paper, we address the feasibility of using these machines to generate social media health messages. We turn to this in the next section.

1.2. Health communication, awareness campaigns, and the message generation bottleneck

The need to communicate about health is clear, compelling, and constantly high: About 60% of worldwide deaths and a large share of the global burden of disease are due to preventable factors (Mokdad et al., 2004; Giles, 2011; Ahmad and Anderson, 2021). By using health communication to inform and influence individuals, we hope to be able to reduce these numbers and help people to live healthier and happier lives (Schmälzle et al., 2017). The COVID-19 pandemic has made this all particularly salient, although there are many other topics beyond infectious disease and vaccinations where health communication comes into play (e.g., general health promotion and disease prevention, nutrition and lifestyle, substance use, risk behaviors, etc.; Thompson and Harrington, 2021).

Health communication research is complex and multifaceted, so this section will only focus on the topic of health-related mass media campaigns (Atkin and Silk, 2008; Rice and Atkin, 2012). Even within this more confined area, we will only focus on awareness messaging. Health awareness campaigns are aimed at increasing the public's knowledge about a particular health issue and ways to prevent it. Though much of health communication research examines attitude and behavioral change, increasing the public's knowledge about the health issue sets the foundation of any health campaign.[5] One must know about the health problem to have an attitude or act upon it. Existing literature provides examples of awareness influencing attitudes and behavior (e.g., Yang and Mackert, 2021). In addition, simply increasing public knowledge about a health issue can help alleviate the extent of the problem. Thus, campaigns' focus can also be to engage the public and increase knowledge (Bettinghaus, 1986).

Mass communication channels, particularly social media platforms, provide cost-effective tools to spread awareness about

---

4 Paraphrasing Lewin, who once said that "There's nothing so practical as a good theory" (1943), Greenwald (2012) states that "There is nothing so theoretical as a good method." History clearly supports both statements, like when the development of telescopes massively propelled astronomical theories, or the cross-pollination of neuroscience and mathematical theories in the 1940s, which paved the way for the deep learning revolution in the last decade.

---

5 Please note that the goal of this study focuses only on generating messages that are capable of raising awareness about health issues. Clearly, additional goals of health communication and persuasion, particularly attitude and potentially even behavior change, are within the realm of possibilities, but these are not the goal of this study. To our knowledge, these goals have not yet been addressed by research and we deliberately chose to focus on the arguably simpler goal of raising awareness and dealing with the complexities of attitudes and behavior at a later point. That said, however, awareness about a health issue is the most obvious and important starting point because people cannot prevent risks of whose existence they are not aware (Bettinghaus, 1986; McGuire, 2001; Schmälzle et al., 2017).

health issues to large audiences (Shi et al., 2018; Willoughby and Noar, 2022). On social media, people can connect with others and have social networks that extend beyond geographical boundaries. Users can share information and pass it on to another through their social networks. Social networks, then, allow information to be disseminated to a large number of people at a rapid pace. Existing literature has shown that social media can be effectively used to raise awareness on health prevention issues such as Covid (Chan et al., 2020), skin cancer (Gough et al., 2017), or the famous ALS ice-water bucket challenge (Wicks, 2014; Shi et al., 2018). In addition, the true power of social media is when content becomes "viral," achieving a significantly high amount of awareness (e.g., ALS ice-water bucket challenge). Health researchers have already begun examining how to leverage this phenomenon on social media to expand the research of health messages (Thackeray et al., 2008; Kim, 2015; Wang et al., 2019). Thus, if leveraged well, communicators can potentially reach billions of people from all over the world through awareness campaigns.

On the other hand, the rapid advancements in social media have created additional challenges for health professionals (Shi et al., 2018; Willoughby and Noar, 2022). For one, each social media platform has different characteristics of predominant users, making it a challenge to truly understand what types of people the campaigns will reach on the platforms. Second, features that make content viral on social media platforms constantly change. This dynamic nature of social media presents a significant challenge for health researchers. While large companies may have the financial resources to hire prominent influencers, health campaigners often lack these resources. In addition, crafting health messages often takes time and resources to conduct formative research, create messages, and then pretest the messages (Snyder, 2007; Rice and Atkin, 2012). By the time messages are created, the content may no longer be interesting enough to the audience. Thus, generating and disseminating quality health content fast enough to match the speed of developments in social media has become a challenging task for campaigners. A message generation system that can generate awareness messages within a few minutes, then, can help reduce the bottleneck of message creation.

1.3. Promoting folic acid awareness as a test case for AI health message generation

This study will use folic acid (FA) as a test case to examine whether a state-of-the-art AI system is capable of generating health messages to promote awareness about this important topic. Folic acid, or folate, is a type of vitamin B9 that is essential for new cell generation and building DNA [Geisel, 2003; Folate (Folic Acid), 2012; CDC, 2022]. While it is suggested that folic acid may help prevent health issues in adults (e.g., stroke; Wang et al., 2007), it is most known to prevent birth defects such as neural tube defects (NTDs) in newborn babies (Scholl and Johnson, 2000; CDC, 2022). NTDs refer to defects in the spinal cord, brain, and areas around them, and they are part of the five most serious birth defects in the world (Githuku et al., 2014). This defect is especially dangerous because it could form in the fetus during the early stages of pregnancy, sometimes even before the mother is aware of

the pregnancy. To prevent NTDs, CDC recommends that women consume at least 400 mcg of folic acid a day, especially during the early stages of pregnancy (Gomes et al., 2016; CDC, 2022). Despite the significance of consuming folic acid, there is a general lack of awareness regarding this issue. Spreading awareness about the issue can therefore help decrease the rate of birth defects (Green-Raleigh et al., 2006; Medawar et al., 2019).

Extant folic acid awareness campaigns showed some effectiveness in increasing knowledge about folic acid. For example, a review by Rofail et al. (2012) showed that existing folic acid campaigns increased knowledge of sources of folic acid. Other extant literature showed that folic acid awareness campaigns not only increased knowledge but also promoted an increase in folic acid consumption (Amitai et al., 2004). Taken together, folic acid remains an important health topic that suffers from a chronic lack of awareness because too few women of childbearing age are aware of the link between folic acid and NTDs during early pregnancy (Medawar et al., 2019). Thus, folic acid is an ideal health issue to test the value of an AI-message engine.

1.4. Using AI to generate health awareness messages

In this section, we will shift back to AI-based message generation and review how newly available text-generation technologies have already been applied to health communication in very few prior studies. To our knowledge, only one paper in communication has used NLP models to generate health awareness messages. Schmälzle and Wilcox (2022) introduced a message machine capable of creating new awareness messages. Specifically, they used a GPT2-LM to generate short topical text messages, now commonly called tweets. This was done because twitter is one of the most prominent social media sites, and specifically text-focused. Moreover, twitter is used heavily for health promotion, although the messages that were generated could also be posted elsewhere, especially on Facebook or Instagram. In order to generate topic-focused (i.e. about Folic Acid) social media messages, Schmälzle and Wilcox (2022) used a fine-tuning strategy. Fine-tuning refers to retraining a language model with a dataset so the model can generate messages with similar structure and content as the dataset (one can think of it as expert training in a specific domain). Specifically, they fine-tuned a GPT-2 based model on messages about folic acid and then used it to generate novel messages about this topic. They compared AI-generated messages against human messages and found that AI-generated messages were on par, and even minimally higher compared to the human-generated messages in terms of quality and clarity.

A second paper used NLP to generate persuasive messages in the context of Covid-19 vaccination. Karinshak et al. (2023) used GPT-3, a successor model of GPT-2, to generate messages to encourage getting the Covid-19 vaccine. One of the features of GPT3 is that it enables so-called zero-shot learning, that is generating messages from prompts without prior fine-tuning. The results showed that participants rated AI-generated messages as more effective, having better arguments, and they affected post-exposure attitudes more strongly.

Together, these studies suggest that AI message generation is feasible and promising for health communication. However, with only two published studies that we are aware of, it is clear that many questions remain unanswered, perhaps even unasked, and that the rapid pace at which AI systems evolve require further research.

1.5. The current study and hypotheses

The current study tested people's perceptions of AI-generated vs. human-generated messages in terms of quality and clarity. Compared to prior work, this study implemented several key innovations: a newer and more powerful LLM named Bloom, a novel message generation strategy (prompting), and a more challenging comparison standard (comparing AI-generated messages to the most shared messages).

This study used the Bloom model, the most recent and largest open-source LLM available to researchers (Scao et al., 2022). In addition to being openly available, Bloom allows for prompting, which refers to inputting the beginning part of a text (e.g., *"A good night's sleep is important because … "*). Then the machine provides the text output that begins with the prompt (e.g., *"A good night's sleep is important because your body needs to recover"*). Simply put, the prompting technology provides a context, thereby constraining the topic of the generated text to the situation provided in the prompt. Studying the mechanisms and effects of prompting is currently an active topic, known as prompt engineering, which holds much potential for communication scientists (Lin and Riedl, 2021; Liu et al., 2023).

This study also improved the process for selecting human-generated messages as comparison standards for AI-generated ones. Schmälzle and Wilcox (2022) study randomly selected human-generated messages from the scrapped tweets. In this study, we randomly selected human-generated tweets as well, but this time, from the retweeted messages only. Retweeting represents an objective outcome that is easy to assess and one of the gold standards of messaging success (DellaVigna and Gentzkow, 2010; Rhodes and Ewoldsen, 2013). In other words, retweeted messages contain elements deemed more worthy of sharing by audiences (Dmochowski et al., 2014; Pei et al., 2019). Thus, by comparing AI-generated messages to retweeted human-generated messages, this study examined the feasibility of using the message engine to not simply generate new FA messages, but to generate those that are at least on par with the content people share with others.

With these innovations and modifications, we examined the potential of AI LLMs for health awareness generation. Specifically, we postulated the following hypothesis based on the previous literature:

H1: AI-generated messages would be at least on par with the retweeted human-generated content in terms of quality and clarity.

We tested our hypothesis by conducting two rounds of online study, one using a college sample and another using the young adult women (ages 25-35) sample. In addition, we examined the general characteristics of the AI-generated messages compared to the human generated messages exhibited through a mix of qualitative assessment and detailed computational analyses of the generated messages.

RQ1: How do AI-generated messages compare to the retweeted human-generated content in terms of semantic structure, sentiment, and readability?

RQ2: What is the feasibility of awareness message generation in general and particularly the novel strategy, i.e., using the Bloom LLM with a prompting approach?

2. Method

In this section, we provide details about the human and computational analysis methods we used to test H1 and answer RQ1. First, we describe the architecture of the message generation system (called message engine) and how the system was used to generate messages (via so-called prompt engineering), Then we outline the two participant samples we recruited for our human evaluation study as well as the methods used for the computational analyses.
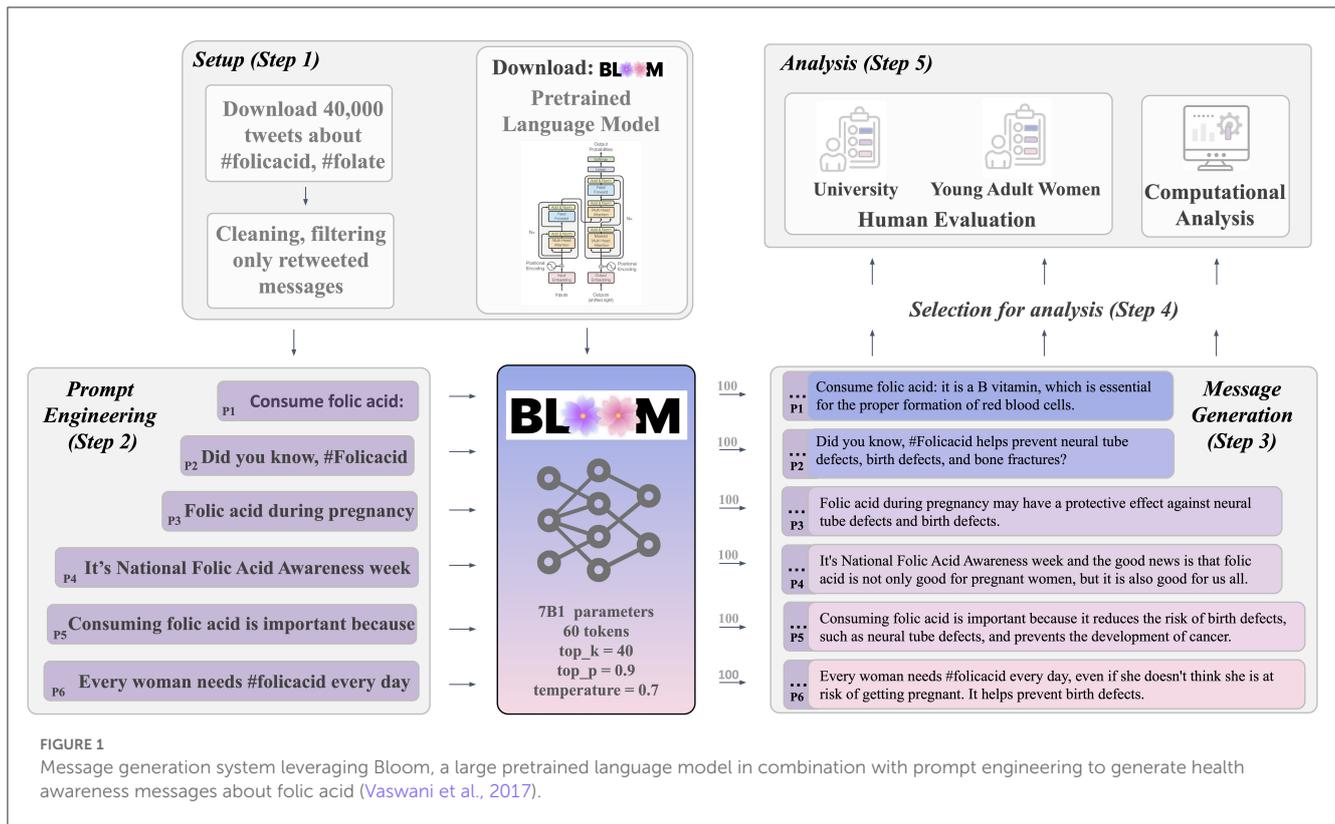
2.1. Description of Bloom, state-of-the-art NLP model

As mentioned above, we used Bloom, the latest and largest open-source multilingual large language model. Bloom is powered by a transformer-based ANN architecture that is similar to that of OpenAI's GPT-3. For feasibility, we used the second largest version (7B1) with 7 billion neural network parameters for this study. Bloom 7B1 was trained on 1.5 TB of pre-processed text from 45 natural and 12 programming languages (Scao et al., 2022) including, for example, the Wikipedia and the Semantic Scholar Open Research Corpus (Laurençon et al., 2022). As previously mentioned, instead of fine-tuning this model as done in previous research (Schmälzle and Wilcox, 2022), here we used a prompting-based generation strategy. See Figure 1 for a conceptual diagram of message generation via Bloom.

2.2. Message generation protocol

2.2.1. Step 1: collecting retweeted human-generated tweets

The first step consisted of scraping FA messages from Twitter. Using the snscrape package (Snscrape, 2021) in python, we collected all tweets with #folic acid or #folate on Twitter (no date constraints). This yielded a base corpus of 42,646 raw messages, which were exported into a csv-file and sorted by retweeted count. The top 50 tweets were selected and filtered for duplicates and non-English text. Next, to exclude messages that had obvious other goals than promoting FA awareness (e.g. selling nutritional supplements), we also removed tweets that included shopping-related words (e.g., buy, order, sale). This resulted in 28 tweets, to which we then added the next two most retweeted to obtain the 30 most retweeted human-generated tweets about FA. We also visually inspected these tweets as well as their sources.

FIGURE 1
Message generation system leveraging Bloom, a large pretrained language model in combination with prompt engineering to generate health awareness messages about folic acid (Vaswani et al., 2017).

We found that half (15) of the 30 tweets came from official sources, including Center for Disease Control (CDC), World Health Organization (WHO), and the United States Department of Health and Human Services (HHS). One can argue that these sources were sending the tweets with the goal of health promotion/education, and that their writing quality reflects at least the average (or better) message quality achieved on social media these days by such sources. The tweet IDs can be found in the Supplementary material.

2.2.2. Step 2: prompt engineering

As prompting can significantly influence the generated messages (Liu et al., 2023), prompts were crafted based on the sentence structure of the most retweeted messages. Starting from the most retweeted message, the beginning phrases of each of the tweets were examined and selected as long as the phrase did not contain too much information about folic acid. Candidate prompts were discarded if we observed the following red flags: too many of the same messages without much informative content, too many names of official organizations or countries, or clearly false information. Some phrases were also modified and tested as prompts. For example, instead of using "Consume #folicacid;," we modified the prompt to "Consume folic acid:" Using this procedure, we engineered the following prompts: "It's National Folic Acid Awareness week," "Every woman needs #folicacid every day," "Did you know, #Folicacid," "Consuming folic acid is important because," "Consuming folic acid:" and "Folic acid during pregnancy."

2.2.3. Step 3: AI message generation

Using the transformer python package, we provided the constructed prompts to the Bloom language model to generate two pools of sample messages: A first (smaller) pool of 600 messages was generated and tested in the human and computational evaluation studies. Second, we expanded this smaller pool into a large pool of 3000 messages (for computational evaluation only, see Supplementary material). Other than varying the input prompts, the generation command used fixed values for the parameters *maximum result length, sampling, temperature, top-k,* and *top-p.* The *maximum result length* specifies the amount of text to be generated and was set to 60 tokens (~45 words) to roughly match the character limit of Twitter messages. The other parameters allow the precision and randomness of the generated messages. By setting the parameter *do_sample = True,* we instructed the model to use the sampling approach, which means that the model will calculate the conditional probability of each word that was stored via training and randomly selected based on the probability. Sampling was chosen over other methods in order to add some variation into the text generation, modeling after the variations that exist in natural human language (Holtzman et al., 2019; Wolf et al., 2020). We then determined the *sampling temperature,* or how much randomness is allowed in generating the text after the prompt. The *temperature* was set to 0.7 to follow the recommended levels for multiple or longer text generation (Misri, 2021). Next, top-k and top-p sampling strategies were used in combination in order to allow for variation in generated text while preventing words with low calculated probabilities from being selected (Fan et al., 2018; von Platen, 2020; top_k = 40, top_p = 0.9).

Using these parameters, we first generated a message pool for the human evaluation studies (see step 4). Next, based on reviewer feedback, we expanded this smaller pool of 600 messages inta a large pool of 3,000 messages. Because it is not feasible to submit this number of messages to human evaluation, the results of the computational analysis of these messages will be reported in the Supplementary material.

2.2.4. Step 4: selecting AI-generated messages for evaluation

After generating the messages, the following procedure was used to select 30 messages for comparison against human generated messages. Using a random number generator, 10 messages were randomly selected from each of the generated messages for each prompt source. Among the 10 messages, messages were discarded if we found the following exclusion criteria: (1) clear false information, (2) specific references to non-US specific organizations (e.g., NHS, UK, Scotland, Netherlands), (3) references to sources that could not be verified [e.g., "the recommended intake of 400 mg/day is based on a meta-analysis of randomized controlled trials (RCTs)], and (4) recipes that include folic acid. In addition, two of the 60 randomly selected messages were discarded because they included characters in different languages or had repetitive phrases. If more than 5 messages passed the exclusion criteria, then we selected the messages that included more content than hashtags or differed from other included messages. Finally, the selected messages were cleaned (e.g., &amp changed to &, end part of the messages deleted if the sentence was not complete). These messages were then used for the human and computational analyses (and the large pool was analyzed computationally—see Supplementary material).

2.3. Human evaluation

2.3.1. Participants

University Sample: Participants were recruited from a study pool and received course credits as compensation for the study, which was approved by the local review board. Responses from participants who completed the survey at an unrealistically fast speed (<4 mins) or failed to complete the survey were discarded. The final dataset included $N = 109$ respondents ($m_{age} = 19.6$, $sd_{age} = 1.42$), with 70% being female ($n_f = 84$). Since folic acid is especially significant for stages during pregnancy, the high proportion of female participants fit the purpose of our study. We also conducted a power analysis using the pwr package in R (Champely, 2020) for a one-sample and one-sided t-test, with effect size $d = 0.3$ and significance level $\alpha = 0.05$. Sample size of 100 was enough to detect significance at the power level of.9. Moreover, we relied on evidence suggesting that a sample of this size is more than sufficient to select best-performing messages as candidates for campaigns (Kim and Cappella, 2019).

Young Adult Women Sample: To have a sample with a less narrow age demographic than the college student sample, we also recruited participants from Prolific, a subject recruitment platform for online studies (Palan and Schitter, 2018). Specifically,

we used the pre-screening questions via Prolific to restrict our participant pool to females ages 25-35 who lived with their spouses or significant others. We specifically selected females ages 25-35 because in the U.S the median age for the first pregnancy is 30 years in age (Morse, 2022). We selected an age range around 30 in an attempt to capture realistic ages during which women may start planning their first child. Although we did not specifically ask about family planning because of privacy concerns, the reasoning was that those who lived with their spouses or significant others may naturally be more inclined to consider pregnancy, and certainly more so than the university sample. Moreover, given that a large number of pregnancies occur spontaneously within this age range, this sample overlaps with the primary audience of folic acid awareness campaigns. A total number of 40 participants responded to the survey ($m_{age} = 31.15$, $sd_{age} = 2.99$), a number chosen in line with the recommendations of Kim and Cappella (2019) for message evaluation study sample sizes. Each participant received $2.50 for their participation in the study (about $9.50 per hour).

2.3.2. Procedure

Both studies (university sample and young adult women sample) were conducted online via Qualtrics, and participants were not told in advance which of the messages were AI-generated or human generated; they were only told that the purpose of the study was to evaluate health messages related to folic acid. Once participants consented to the study, they were asked to evaluate the quality and clarity of the messages in blocks. Message order was randomized within each block and approximately half of the sample started with the first question, the other half started with the second block. The two survey items were adopted from Schmälzle and Wilcox (2022). The quality of the message measure asked, *"How much do you agree that the content and the quality of this message is appropriate to increase public knowledge about folic acid,"* and responses varied from 1 (strongly disagree) to 5 (strongly agree). The clarity of the message measure stated, *"Please evaluate the following messages in terms of whether they are clear and easy to understand,"* with 1 meaning very unclear and 5 meaning very clear.

Survey data were downloaded and further processed in python. Specifically, we computed by-message averages for each question and averaged across the blocks of AI- and human-generated messages, respectively. To answer H1, we used the scipy-package (Virtanen et al., 2020) to test for differences between AI- and human-generated messages in each measure (α- level of.05). Furthermore, the participant sample was also split by sex (due to the relevancy of folic acid during pregnancy).

2.3.3. Computational evaluation

The goal of the computational analyses was to examine the characteristics of the AI- and human-generated messages. Specifically, we examined the AI-generated and human-generated messages via several common text analytic methods in python in R, including N-gram analysis, semantic analysis, readability analysis, topic modeling, and sentiment analysis. Through n-gram analysis and semantic analysis, we hoped to examine the similarities in the word distributions and the general attributes of the sentences. The readability analysis examined how easy the AI-generated messages

TABLE 1 Results from the online survey that asked participants to rate AI-generated and human-generated messages in terms of clarity and quality.

	University sample			Young adult women sample		
	AI-generated	Human-generated	t (*p*-value)	AI-generated	Human-generated	t (*p*-value)
Clarity	3.77 (0.55)	3.22 (0.43)	4.32 (<0.001)	4.38 (0.58)	3.79 (0.51)	4.27 (<0.001)
Quality	3.65 (0.42)	3.12 (0.33)	5.39 (<0.001)	4.08 (0.38)	3.36 (0.51)	6.11 (<0.001)

Left: University Sample; Right: Young Adult Women Sample. The table shows average ratings for each concept (scales range from 1–5, standard deviations in brackets) and the results of the t-test comparing ratings for AI vs. human-generated messages.

were to understand compared to human messages (Flesch, 1946; DeWilde, 2020). Using topic modeling (Blei, 2012) and sentiment analysis (Hutto and Gilbert, 2014; Hirschberg and Manning, 2015), we examined the similarities and differences in the discussed topics and the general sentiment of the messages. These analyses were carried out using python and R packages including spacy, textacy, vader, and the sentence-transformers (Hornik and Grün, 2011; Hutto and Gilbert, 2014; Reimers and Gurevych, 2019; DeWilde, 2020; Matthew et al., 2020).

3. Results

This study examined the feasibility to use the Bloom message engine to generate clear, effective, and novel awareness messages. In this section, we show how the AI-generated messages compared to retweeted human-generated messages in terms of clarity and quality from the online survey studies with the university and young adult women samples (H1). Next, we describe characteristics of the generated messages and present the results from the computational methods (RQ2).

3.1. Results from the human evaluation study (H1)

We first analyzed the data from the university sample. The results from the one-sample *t*-test supported H1 (see Table 1). In fact, AI-generated messages were rated significantly higher than human-generated tweets for both dimensions—message clarity ($t_{clarity}$ = 4.32, p < 0.001) and message quality ($t_{quality}$ = 5.39, p < 0.001). Figure 2 also shows the result for each dimension for the full population. On average, the participants rated AI-generated and human-generated messages as relatively clear and easy to understand and containing quality content appropriate to increase FA awareness (ratings > 3 out of 5). However, the left graph of Figure 2A shows that the distribution of the AI-generated messages was concentrated toward the top (around 4) while human-generated messages had a slightly more even distribution, concentrated around 3-3.5. The distribution for females appeared wider than for males, with more AI-generated messages rated higher than human-generated messages. In addition, for males, AI-generated messages on average seemed similarly rated as the human-generated messages in terms of clarity. This may be the case because many FA messages referred to the importance of

consuming the vitamin during pregnancy, which may not apply to college-age males.

Next, we computed the same analysis for the young adult women sample. As for the analyses of the university sample, the results from the one-sample *t*-test comparing ratings for the young adult women sample supported H1 (see Table 1). In fact, AI-generated messages were rated significantly higher than human-generated tweets for both dimensions—message clarity ($t_{clarity}$ = 4.27, p <0.01) and message quality ($t_{quality}$ = 6.11, p <0.01). On average, the participants rated AI-generated messages as relatively clear and easy to understand and containing quality content appropriate to increase FA awareness (ratings > 3 out of 5). However, the ratings for the human-generated tweets had a wider range depending on the tweet. Figure 2B also shows that the distribution of the AI-generated messages was concentrated toward the top (above 4) while human-generated messages had a slightly more even distribution, concentrated around 3–3.5.

3.2. Qualitative characteristics of the generated messages and computational analyses (RQ2)

In terms of usability, the Bloom message engine was much more efficient and simpler to use compared to the GPT-2 model used in the previous study. The main appeal of Bloom is its ability of zero-shot learning, or the capability of generating messages with prompting, and the fact that it does not require a resource-intensive fine-tuning process. Because of this feature, we were able to start the message-generation process right away. While harnessing the message engine requires some familiarity with python coding, online resources with explanations of the coding made the process easy to learn (e.g., Wolf et al., 2020; Theron, 2022; Tunstall et al., 2022). The most time-consuming part was preparing for prompt engineering (e.g., scraping and cleaning the tweets, which took a few days to complete). However, now that the codebase exists, the prompt engineering process can be replicated quickly and made more flexible and comfortable for users. Once the final prompts were determined, the actual message generation process was fast. Loading the Bloom 7B1 model into Google Colab and generating 600 messages took a little over an hour. Overall, the model was fast and efficient to use, and we can expect that AI message generation systems will become even simpler to use in the future as the technology becomes commodied.
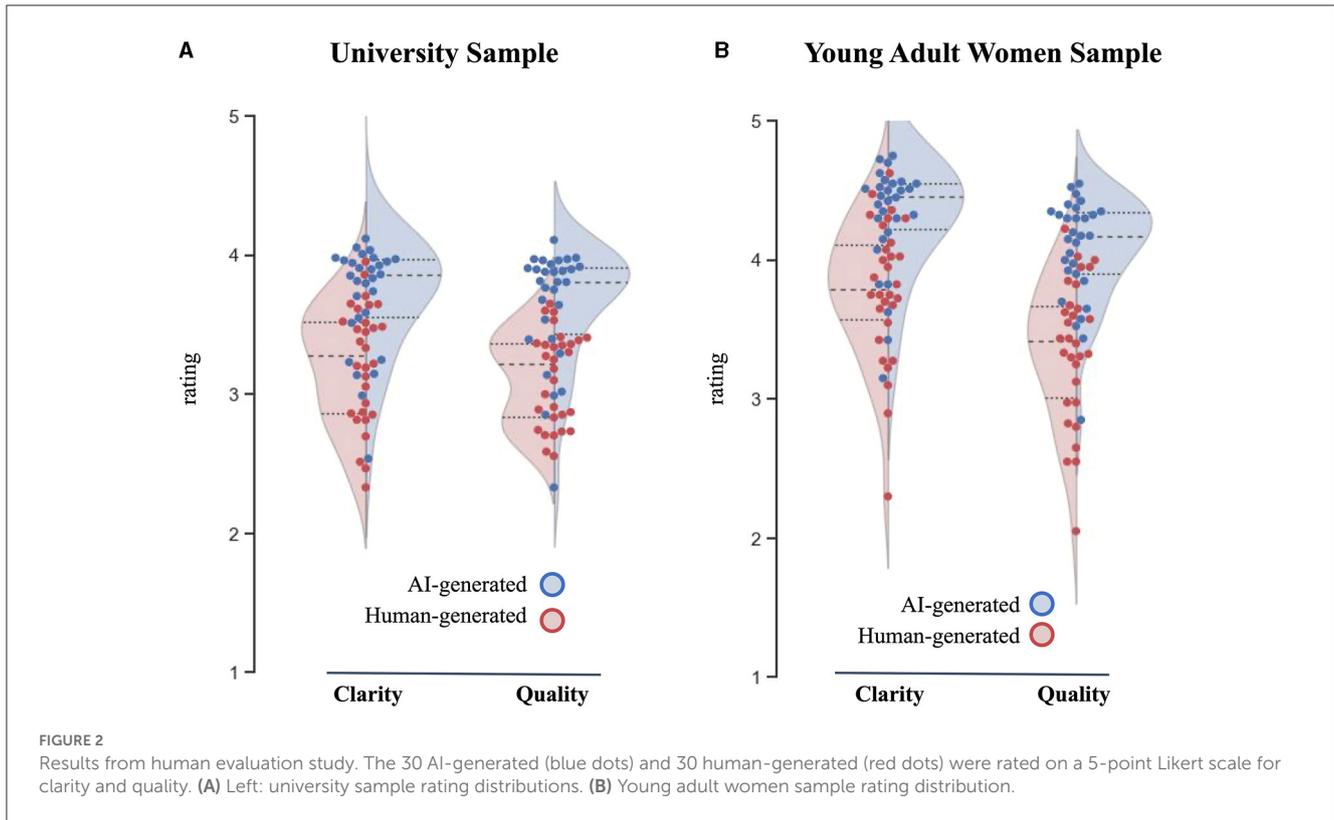
FIGURE 2
Results from human evaluation study. The 30 AI-generated (blue dots) and 30 human-generated (red dots) were rated on a 5-point Likert scale for clarity and quality. **(A)** Left: university sample rating distributions. **(B)** Young adult women sample rating distribution.

TABLE 2 Means (standard deviations) and results of a *t*-test comparing the means of AI- vs. human-generated messages for computational analyses of reading ease (Flesch reading ease, higher scores are more readable, scale ranges up to 100) and message sentiment (Vader Compound Score, scores range from −4 to 4, with higher scores indicating more positive sentiment).

	AI-generated	Human-generated	*t* (*p*-value)
Flesch reading ease	63.4 (17.9)	68.4 (20.3)	0.99 (0.32)
Sentiment	0.25 (0.53)	0.23 (0.49)	0.16 (0.87)

Having affirmed the feasibility of the message engine, we will now discuss qualitative characteristics of the messages. First, the generated messages showed differing characteristics based on the structure of the prompts. When prompts included #folicacid, the generated messages also included hashtags and included features that are characteristic of tweet messages. For instance, the prompt *"Did you know, #Folicacid"* generated text such as *"Did you know, #Folicacid is one of the most important invitations in our daily diet?...So make sure to have it daily! #FolicAcid #VitaminB..."* On the other hand, when prompts did not include hashtags, the tone appeared to be more formal, and the content tended to include more factual information about folic acid. The prompt, *"Consuming folic acid is important because,"* for example, generated text such as *"Consuming folic acid is important because it helps prevent neural tube defects, which are birth defects in the brain and spinal cord. Folic acid is also necessary for healthy blood cells and nerve cells, and for maintaining healthy red blood cells...."* These results point to how prompt technology can be used to strategically control qualities and tone of the AI-generated content, which is a fundamental feature.

Next, the results from the computational analysis[6] showed that AI-generated and human-generated messages were similar in terms of word distributions, topics discussed, semantics, readability, and sentiment (see Table 2; Figure 3). For instance, the uni- and bi-grams show that both AI-generated and human-generated messages contain the words related to neural tube and birth defects, focused generally on folic acid, pregnancy, and health topics. An analysis of text readability statistics (Flesch, 1948) showed that AI-generated messages were as easy to read, if not easier to read, compared to the human-generated messages ($m_{Flesch-ScoreAI}$ = 68.4, $m_{Flesch-Scorehuman}$ = 63.4; *n.s.*). Next, a sentiment analysis using the valence-aware dictionary approach (VADER) showed no significant differences in terms of message sentiment, although nominally the AI-generated messages again ranked slightly higher (i.e., more positive sentiment; $m_{VaderCompoundAI}$ = 0.25, $m_{FVaderCompoundhuman}$ = 0.23; *n.s.*). Overall, these analyses showed that there were no major differences between the AI-generated and human-generated messages.

Next, we focused on topics covered and semantic properties of the messages: A topic modeling analysis showed that both AI and human-generated messages discussed similar topics centered around pregnancy, birth defects, and essential vitamins. One observation is that the AI-generated messages appeared to weigh the topics of pregnancy and risk of birth defects more heavily.

---

6 Note that these results refer to the first pool of messages that was also evaluated in the human evaluation studies. A computational analysis of the much larger pool of AI-generated messages can be found in the Supplementary material. We chose to report the analysis for the more restricted pool of messages here to match the human evaluation studies.
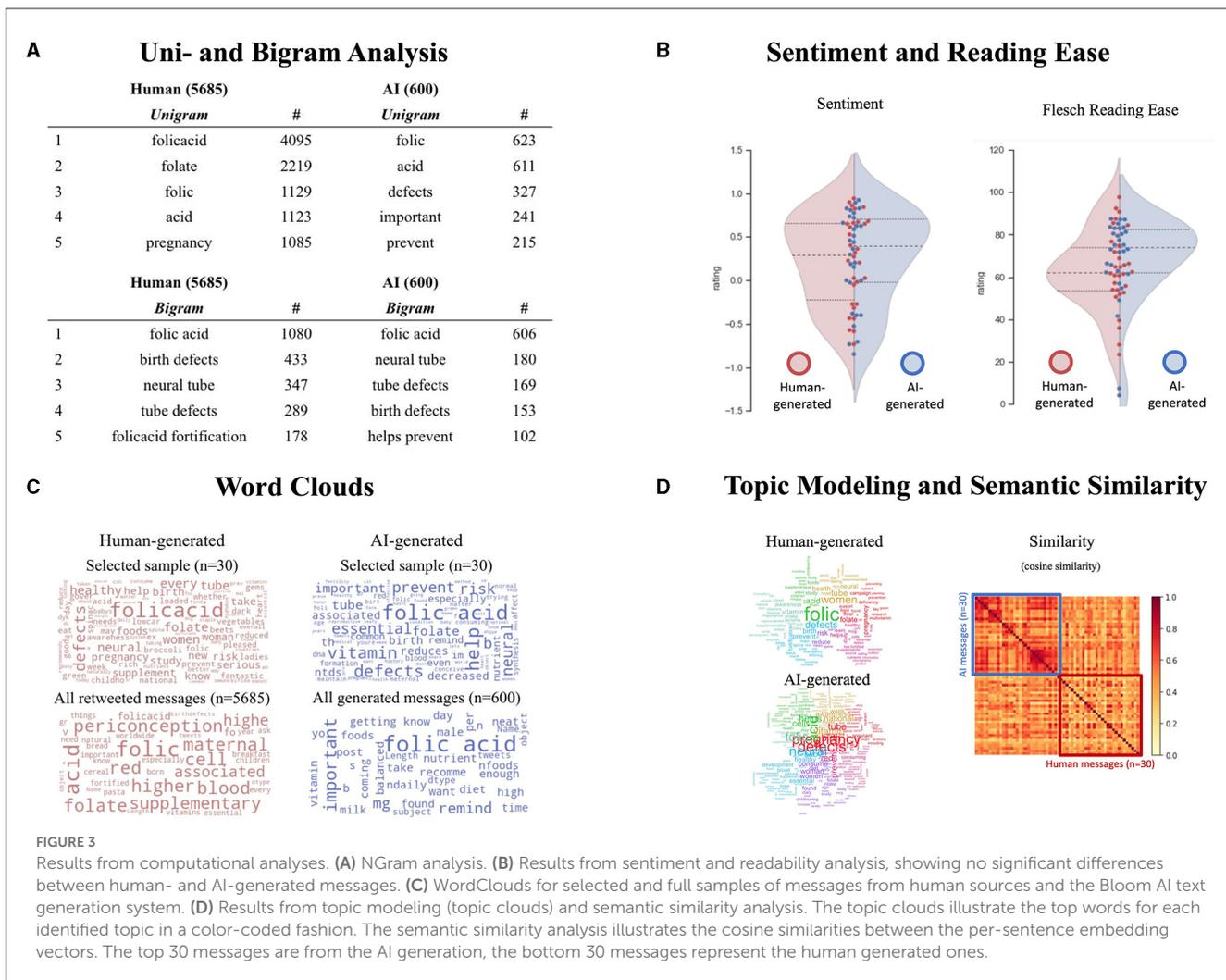
FIGURE 3
Results from computational analyses. **(A)** NGram analysis. **(B)** Results from sentiment and readability analysis, showing no significant differences between human- and AI-generated messages. **(C)** WordClouds for selected and full samples of messages from human sources and the Bloom AI text generation system. **(D)** Results from topic modeling (topic clouds) and semantic similarity analysis. The topic clouds illustrate the top words for each identified topic in a color-coded fashion. The semantic similarity analysis illustrates the cosine similarities between the per-sentence embedding vectors. The top 30 messages are from the AI generation, the bottom 30 messages represent the human generated ones.

Lastly, we conducted an analysis of semantic similarities between all messages using the sentence-transformer package. This analysis asked how similar the sentence vectors of each of the 60 messages (30 AI-generated and 30 human messages) were to each other, as assessed via cosine similarity. As can be seen in Figure 3, the AI-generated messages were more similar to each other ($m_{similarityAI} = 0.59$, $m_{similarityhuman} = 0.43$; $p < 0.001$; note that while we removed prompts for other analyses, we included prompts here because the semantic similarity analysis requires entire sentences). As illustrated by Figure 3, a block-like structure separates the AI-generated messages from the rest, and within the AI messages, subblocks or similarity clusters exist. Importantly, all messages are fairly similar to each other in this semantics measure, with an average similarity score of about 0.5. In sum, these analyses confirm that the AI-generated messages were on topic, readable, and generally positive in tone.

4. Discussion

This study examined the feasibility of using the Bloom message engine to generate clear and high-quality awareness messages and

tested the potential of prompt-engineering to steer the message creation process. Overall, the Bloom message engine was easy to use and proficient in generating messages. Across two samples, human raters perceive AI-generated messages as easy to understand and appropriate to increase public knowledge about folic acid.

The most important result is that both human evaluation studies clearly supported hypothesis 1: Human evaluators rated the AI-generated messages as more clear and as higher in quality compared to the retweeted human-generated messages. This result was even more pronounced in the young adult women sample, whose pre-screened demographics ensured that the sample matches the audience of folic acid awareness campaigns. But the same result was also present among the university sample, and also in the men within this sample. Thus, the findings are robust across different groups.

The finding that AI-generated messages compare favorably against human-generated messages is also noteworthy in light of the fact that human-generated messages comprised messages selected based on their retweet count. In other words, rather than using any messages, retweeted messages arguably pose a higher comparison standard; although we cannot know what the motivations behind retweeting FA-messages were, it seems

plausible that they contain elements that people deem worthy of sharing with others. Therefore, the current results underscore that the Bloom-based message engine can generate awareness messages that surpass even the retweeted human messages. For obvious ethical reasons, we refrained from actually disseminating the AI-generated messages on social media, but we argue that it would be likely that they would also be shared more often. The potential to go viral is one of social media's most notable features when it comes to health communication because viral messages can reach billions of people across the globe to spread awareness about certain topics. In summary, the current results confirm that the Bloom message engine has a high potential to generate clear and high-quality content for awareness messages.

Moving beyond the human evaluation results, qualitative and computational analyses of the generated messages provided further insights about the characteristics of AI-generated content. The computational analyses demonstrated that the AI-generated messages are similar to the human-generated messages in specific quantitative text characteristics like word distributions, readability, sentiment, discussed topics, and general semantic similarity. One distinct feature was that AI-generated messages seemed to contain more words related to prevention. Perhaps this speaks to the differences in tweets vs. AI-generated messages. Messages may exhibit a tendency to have more positive sentiment and focus on the importance of folic acid (e.g., preventing neural tube defects). This all points to the enormous potential of the Bloom model in combination with prompt engineering, whose role in this field is likely to explode over the next few years.

Our study also illustrates that even without additional training with topic-specific datasets (i.e., fine-tuning), the Bloom-based model ("message engine") can generate focused, clear, and appropriate awareness campaign messages. This goes beyond previous work that had relied on fine-tuning (Schmälzle and Wilcox, 2022), which requires costly retraining, time, and is overall more challenging. Since the system used can generate hundreds of messages within a few hours, it can help alleviate the burden of cost and manpower required in creating health communication messages. Moreover, the prompt-engineering strategy employed here allows researchers to steer the content and sentiment of the generated messages (e.g., informal setting to fit social media, or formal to fit professional settings). This feature is extremely promising because it allows communicators to infuse strategy and theory into the AI generation process, which otherwise would remain a "creative black box." Importantly, after the AI-based message generation itself, health researchers could still filter through all messages to select and modify them, or to enter promising candidates back into the system as prompts to try to generate more of their kind. An important note here is that the message engine is not meant to be the decision-maker. Rather, it is a tool that is meant to support the health researcher to save cost, personnel ressources, and time, and to add a replicable, but creative element to the generation process. In all practical use cases, ethics and legal aspects would obviously still require the presence of humans as curators of content.

Another application area of the AI-message engine is that it can be used to generate messages for topics beyond simple awareness messaging (Baclic et al., 2020). As argued above, we deliberately focused on awareness as a fundamental goal of health communication (Bettinghaus, 1986; McGuire, 2001), but clearly, it will be promising to explore the potential of this approach to influence health-related attitudes and behaviors. Furthermore, given the rise of image generation as a visual complement to text generation, it would also be interesting to start generating health-related imagery and combine such messages with AI-generated texts (Crowson et al., 2022). This is especially important because newer social media platforms are increasingly depending on visuals (e.g., Snapchat, TikTok). Overall, we anticipate that this work will find many applications in health communication in the near future, and similar cases could be made for other communication topics (e.g., environmental and political communication) or applied business communication (e.g., AI-copywriting systems, which are already offered commercially).

Going beyond the obvious application potential, we next ask about the theoretical potential and contributions of this work. At first glance, it might be tempting to view AI message generation as a methodological innovation only, but we argue that the advent of natural text generation systems holds immense potential for method-theory synergy (Greenwald, 2012; Weber et al., 2018): Specifically, a healthy mix of *theory*, *measurement*, and *control* is central to the progress of any science (Bechtel, 2008; Chalmers, 2013; Craver and Darden, 2013). Put simply, our theories affect the phenomenon we can conceive of and use to explain the phenomenon in question. Our measures, in turn, determine what phenomena we can "see" and with which precision. Lastly, the ability to control lets us "push things around," manipulating or intervening causally on variables to test our theories. In physics, chemistry, and biology, the triad of *theory, measurement,* and *control* enabled everything from understanding atoms and genetics (on the basic science side) to building mechanical engines and synthetically assembled vaccines (on the applied side). In NLP, the advent of powerful theories that enable machines to generate natural text clearly represents a breakthrough in the quest to understand, measure, and control linguistic or even cognitive capacities (LeCun et al., 2015; Mitchell, 2019). As we saw from the results of this study, this advancement from powerful theories in NLP has many implications for communication (Hassabis et al., 2017; Lake et al., 2017; Dubova, 2022).

Critically, however, we are not claiming that these advances have solved long-standing theoretical questions about the nature of language, let alone symbolic communication—it's quite the opposite. So far, progress in language modeling has focused on language-intrinsic aspects (e.g. syntax and semantics; Bender and Koller, 2020), whereas pragmatic and extralinguistic processes that are critical to communication have been almost ignored. However, the verbal distinction between semantics and pragmatics has always been cumbersome since the underlying phenomena appear to gradually blend into each other. The current work thus represents a step to bridge between NLP and communication, albeit only an initial one. Moreover, the fact that human participants evaluated the AI-generated messages as even better than human-generated ones (and as better as even the most highly shared human-generated messages) underscores the importance of this approach.

In addition to potentially surpassing human-level performance for message generation, another theoretical benefit of AI systems

is that they are rigorously quantitative. While it is often claimed that deep learning is like a "black box" (Marcus and Davis, 2019; Xu et al., 2019; Chen et al., 2020), this is actually not the case: Rather, each of Bloom's over seven billion parameters are computed and thus, in principle, objectively knowable.[7] This again points to a key feature of science because it affords better measurement, better control, and ultimately better theoretical understanding. Indeed, one could perhaps compare AI models to the "Petri dish" in chemistry, i.e., as providing a precisely controlled experimental setting in which researchers can generate synthetic messages under controlled laboratory conditions, and later test how they fare in the market. This promotes insights into properties that make messages effective, an issue that still remains insufficiently understood (Harrington, 2015, 2016; O'Keefe and Hoeken, 2021). In this sense, especially the prompt engineering process, which has been successfully applied here, has barely scratched the surface of what is possible—and needed—to fully realize the potential of this applied fruit of a theoretically matured science of messages.

The emergence of AI-based methods for communication content generation clearly raises important ethical issues. Examples of existing issues include bots in the political domain (which spread misinformation), online troll farms (which stoke conflict), or general misuse of persuasive technologies for target advertising or mood, and emotion intervention (Solaiman et al., 2019; Kreps et al., 2020). At this point, there is no clear regulatory framework for these technologies. Although a few "responsible AI" initiatives exist, the novelty of the topic and the evolving nature of the underlying science prevent any final judgment. However, the creators of Bloom have attempted to address this issue by releasing Bloom under a RAIL license (responsible AI license, Contractor et al., 2020), which includes a number of disallowed use cases, like generating law-violating content. Among these cases is using Bloom for medical purposes. Clearly, given the LLMs' lack of medical knowledge (Schmälzle and Wilcox, 2022), it would be irresponsible to use the model to generate diagnoses and prescribe treatment. On the other hand, using the model to improve general health awareness information seems like a particularly beneficial use case for how LLMs (under supervision by health communication experts acting as message curators) could be used for social good and to improve the quality of public health information. With this in mind, we also want to underscore the importance that communication researchers engage with and help shape this debate.

Like all research, the current paper has several limitations. First, we tested only one topic, folic acid. Although this was an informed choice based on the characteristics of this particular health topic, it will be necessary to expand these findings to other domains. Second, we focused only on health messages designed to raise awareness but ignored more downstream topics like attitude and behavior change. Again, we argue that this provides a logical starting point, but that more work needs to be done to expand

to other topics within health communication. Relatedly, we only generated relatively short messages, such as the ones we find on social media (primarily twitter). However, health communication takes many forms, and as such more work is needed to generate e.g., health-related stories (longer text) or images (no text at all, but a different modality). Third, the computational and human evaluations of the generated messages could be improved. Regarding the computational analyses, we only focused on a number of salient and quantifiable characteristics, like NGrams, sentiment, and reading ease. However, there are additional topics that were not explored, such as similarity of the entire corpora of human and AI-generated messages, other theoretical concepts (e.g. use of politeness instead of general sentiment Yeomans et al., 2019, or the use of specific kinds of persuasive strategies (Armstrong, 2010; Tan et al., 2016; O'Keefe and Hoeken, 2021). Again, this limitation is explained by the early stage of this research—this is only the third study of this kind that we are aware of— and we look forward to future work expanding the scope of computational message analytics. Finally, human evaluation is not without limitations either. In particular, one could criticize the narrow measures and the sample. Regarding the latter, we argue that the populations we sampled from (college students and young adult women living with a partner) overlaps with the target population for a FA awareness campaign (potential parents and particularly women of childbearing age), but it would be advisable to also test how broader audiences respond to these messages.

5. Conclusions

At the start of this paper, we asked readers to guess which of the two messages came from an AI. The answer is that it was the first message, and the fact that most readers could only guess the correct answer affirms our main findings. Specifically, the Bloom message engine was easy to use, the generated messages were generally on par with human messages in terms of quantitative characteristics, and they were rated as clearer and of higher quality compared to even the most retweeted human messages. Thus, the message engine could be used to alleviate the bottleneck in the message generation process for health awareness messages. Overall, this approach offers fruitful ground to use quantitative methods to examine the generation, content, and reception of messages.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board, Michigan State University. The patients/participants provided their written informed consent to participate in this study.

---

7 Granted, this mindboggling complexity prevents us from understanding the system behavior intuitively, which is what people mean when they call it a black box. However, except for deliberately random parameters, the systems behave deterministically, and, in this sense, the black box metaphor is mistaken.

Author contributions

SL contributed by leading the project, planning and collecting data, and writing most of the manuscript. RS contributed by writing parts of the manuscript. Both authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm. 2023.1129082/full#supplementary-material

References

Ahmad, F. B., and Anderson, R. N. (2021). The leading causes of death in the US for 2020. *Jama* 325, 1829–1830. doi: 10.1001/jama.2021.5469

Amitai, Y., Fisher, N., Haringman, M., Meiraz, H., Baram, N., and Leventhal, A. (2004). Increased awareness, knowledge and utilization of preconceptional folic acid in Israel following a national campaign. *Prev. Med.* 39, 731–737. doi: 10.1016/j.ypmed.2004.02.042

Armstrong, J. S. (2010). *Persuasive Advertising: Evidence-Based Principles*. London: Palgrave Macmillan.

Atkin, C., and Silk, K. (2008). "Health communication," in *An Integrated Approach to Communication Theory*, eds. D.W. Stacks, M. B. Salwen, and K. C. Eichhorn (New York, NY: Routledge).

Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., and Schonfeld, J. (2020). Challenges and opportunities for public health made possible by advances in natural language processing. *Can. Commun. Dis. Rep.* 46, 161–168. doi: 10.14745/ccdr.v46i06a02

Bechtel, W. (2008). Mechanisms in cognitive psychology: what are the operations? *Philos. Sci.* 75, 983–994. doi: 10.1086/594540

Bender, E. M., and Koller, A. (2020). "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Bettinghaus, E. P. (1986). Health promotion and the knowledge-attitude-behavior continuum. *Prev. Med.* 15, 475–491. doi: 10.1016/0091-7435(86)90025-3

Blei, D. M. (2012). Topic modeling and digital humanities. *Digit. Humanit. Q.* 2, 1.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2303.12712 (accessed April 20, 2023).

CDC (2022). *Folic Acid*. Available online at: https://www.cdc.gov/ncbddd/folicacid/about.html (accessed October 5, 2022).

Chalmers, A. F. (2013). *What is This Thing Called Science?* Indianapolis: Hackett Publishing.

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. Available online at: https://CRAN.R-project.org/package=pwr (accessed October 15, 2022).

Chan, A. K., Nickson, C. P., Rudolph, J. W., Lee, A., and Joynt, G. M. (2020). Social media for rapid knowledge dissemination: early experience from the COVID-19 pandemic. *Anaesth.* 75, 1579–1582. doi: 10.1111/anae.15057

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: A unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci.* 14, 580632. doi: 10.3389/fncom.2020.580632

Cho, H. (2011). *Health Communication Message Design: Theory and Practice.* Thousand Oaks: Sage Publications.

Chollet, F. (2021). *Deep Learning with Python (2nd ed.)*. New York, NY: Manning Publications Co.

Contractor, D., McDuff, D., Haines, J., Lee, J., Hines, C., and Hecht, B. (2020). *Behavioral use licensing for responsible AI. arXiv [Preprint]*. Available online at: http://arxiv.org/abs/2011.03116 (accessed October 17, 2022).

Craver, C. F., and Darden, L. (2013). *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago: University of Chicago Press.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., et al. (2022). *VQGAN-CLIP: Open Domain Image Generation and Editing With Natural Language Guidance*. arXiv [Preprint]. Available online at: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136970088.pdf (accessed October 15, 2022).

DellaVigna, S., and Gentzkow, M. (2010). Persuasion: Empirical evidence. *Annu. Rev. Econ.* 2, 643–669. doi: 10.1146/annurev.economics.102308.124309

DeWilde, B. (2020). *Textacy: NLP, before and after spaCy*. Available online at: https://github.com/chartbeat-labs/textacy (accessed September 10, 2022).

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., and Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nat. Commun.* 5, 4567. doi: 10.1038/ncomms5567

Dubova, M. (2022). Building human-like communicative intelligence: a grounded perspective. *Cogn. Syst. Res.* 72, 63–79. doi: 10.1016/j.cogsys.2021.12.002

Fan, A., Lewis, M., and Dauphin, Y. (2018). *Hierarchical neural story generation. arXiv [Preprint]*. Available online at: http://arxiv.org/abs/1805.04833 (accessed October 15, 2022).

Flesch, R. (1946). *The Art of Plain Talk*. New York, NY: Harper and Row, Publishers.

Flesch, R. (1948). A new readability yardstick. *J. Appl. Psychol.* 32, 221-233. doi: 10.1037/h0057532

Folate (Folic Acid) (2012). *Vitamin B9*. Available online at: https://www.hsph.harvard.edu/nutritionsource/folic-acid (accessed October 15, 2022).

Gatt, A., and Krahmer, E. (2018). Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Int. J. Artif. Intell. Res.* 61, 65–170. doi: 10.1613/jair.5477

Geisel, J. (2003). Folic acid and neural tube defects in pregnancy: a review. *J. Perinat. Neonatal. Nurs.* 17, 268–279. doi: 10.1097/00005237-200310000-00005

Gelernter, D. (2010). *The Muse in the Machine: Computerizing the Poetry of Human Thought*. New York, NY: Simon and Schuster.

Giles, J. (2011). Social science lines up its biggest challenges. *Nat.* 470, 18–19. doi: 10.1038/470018a

Githuku, J. N., Azofeifa, A., Valencia, D., Ao, T., Hamner, H., Amwayi, S., et al. (2014). Assessing the prevalence of spina bifida and encephalocele in a Kenyan hospital from 2005-2010: Implications for a neural tube defects surveillance system. *Pan. Afr. Med. J.* 18, 60. doi: 10.11604/pamj.2014.18.60.4070

Gomes, S., Lopes, C., and Pinto, E. (2016). Folate and folic acid in the periconceptional period: recommendations from official health organizations in thirty-six countries worldwide and WHO. *Public Health Nutr.* 19, 176–189. doi: 10.1017/S1368980015000555

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Gough, A., Hunter, R. F., Ajao, O., Jurek, A., McKeown, G., Hong, J., et al. (2017). Tweet for behavior change: Using social media for the dissemination of public health messages. *JMIR Public Health Surveill.* 3, e6313. doi: 10.2196/publichealth.6313

Greene, J. O. (2013). *Message Production: Advances in Communication Theory*. New York, NY: Routledge.

Green-Raleigh, K., Carter, H., Mulinare, J., Prue, C., and Petrini, J. (2006). Trends in folic acid awareness and behavior in the United States: the gallup organization for the march of dimes foundation surveys, 1995–2005. *Matern. Child Health J.* 10, 177–182. doi: 10.1007/s10995-006-0104-0

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspect. Psychol. Sci.* 7, 99–108. doi: 10.1177/1745691611434210

Hancock, J. T., Naaman, M., and Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput. Mediat. Commun.* 25, 89–100. doi: 10.1093/jcmc/zmz022

Harrington, N. G. (2015). Introduction to the special issue: Message design in health communication research. *Health Commun.* 30, 103–105. doi: 10.1080/10410236.2014.974133

Harrington, N. G. (2016). Persuasive health message design. *ORE of Commun.* 13, 7. doi: 10.1093/acrefore/9780190228613.013.7

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron.* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hirschberg, J., and Manning, C. D. (2015). Advances in natural language processing. *Science.* 349, 261–266. doi: 10.1126/science.aaa8685

Hodgkinson, G. P., Langan-Fox, J., and Sadler-Smith, E. (2008). Intuition: a fundamental bridging construct in the behavioural sciences. *Br. J. Psychol.* 99, 1–27. doi: 10.1348/000712607X216666

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). *The curious case of neural text degeneration. arXiv [Preprint].* Available online at: http://arxiv.org/abs/1904.09751 (accessed October 2, 2022).

Hornik, K., and Grün, B. (2011). Topicmodels: an R package for fitting topic models. *J. Stat. Softw.* 40, 1–30. doi: 10.18637/jss.v040.i13

Huskey, R., Turner, B. O., and Weber, R. (2020). Individual differences in brain responses: new opportunities for tailoring health communication campaigns. *Front. Hum. Neurosci.* 14, 565973. doi: 10.3389/fnhum.2020.565973

Hutto, C., and Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Weblogs Soc. Media.* 8, 216–225. doi: 10.1609/icwsm.v8i1.14550

Karinshak, E., Liu, S. X., Park, J. S., and Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum. Comput. Interact (CSCW).* 23, 52. doi: 10.1145/3579592

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). *CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv [Preprint].* Available online at: http://arxiv.org/abs/1909.05858 (accessed August 31, 2022).

Kim, H. S. (2015). Attracting views and going viral: How message features and news-sharing channels affect health news diffusion. *J. Commun.* 65, 512–534. doi: 10.1111/jcom.12160

Kim, M., and Cappella, J. N. (2019). An efficient message evaluation protocol: two empirical analyses on positional effects and optimal sample size. *J. Health Commun.* 24, 761–769. doi: 10.1080/10810730.2019.1668090

Kreps, S. E., McCain, M., and Brundage, M. (2020). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *J. Exp. Political Sci.* 9, 104–117. doi: 10.2139/ssrn.3525002

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253. doi: 10.1017/S0140525X16001837

Lapinski, M. K., and Rimal, R. N. (2005). An explication of social norms. *Commun. Theory.* 15, 127–147. doi: 10.1111/j.1468-2885.2005.tb00329.x

Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., et al. (2022). "The bigscience roots corpus: A 1.6 tb composite multilingual dataset," in *Advances in Neural Information Processing Systems, Vol. 35.* p. 31809–31826.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature.* 521, 7553,7436. doi: 10.1038/nature14539

Lee, E. W., Zheng, H., Goh, D. H. L., Lee, C. S., and Theng, Y. L. (2023). Examining COVID-19 tweet diffusion using an integrated social amplification of risk and issue-attention cycle framework. *Health Commun.* 2, 1–14. doi: 10.1080/10410236.2023.2170201

Lin, Z., and Riedl, M. (2021). *Plug-and-blend: A framework for Controllable Story Generation With Blended Control Codes. arXiv [Preprint].* Available online at: http://arxiv.org/abs/2104.04039 (accessed October 15, 2022).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3560815

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust.* New York, NY: Pantheon Books.

Matthew, H., Montani, I., van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo.* doi: 10.5281/zenodo.1212303

McGuire, W. J. (2001). "Input and output variables currently promising for constructing persuasive communications," in *The Public Communication Campaigns,* eds. R. E. Rice and C. K. Atkin (Los Angeles, CA: Sage Publications, Inc.), 22–48.

Medawar, G., Wehbe, T., and Jaoude, E. A. (2019). Awareness and use of folic acid among women of childbearing age. *Ann. Glob. Health.* 85, 1. doi: 10.5334/aogh.2396

Miller, G., Boster, F., Roloff, M., and Seibold, D. (1977). Compliance-gaining message strategies: a typology and some findings concerning effects of situational differences. *Commun. Monograph.* 44, 37–51. doi: 10.1080/03637757709390113

Misri, I. (2021). *How to Set Sampling Temperature for GPT Models.* Available online at: https://medium.com/@imisri1/how-to-set-sampling-temperature-for-gpt-models-762887df1fac (accessed October 15, 2022).

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans.* London: Penguin UK.

Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L. (2004). Actual causes of death in the United States, 2000. *JAMA.* 291, 1238–1245. doi: 10.1001/jama.291.10.1238

Morse, A. (2022). 'Stable fertility rates 1990-2019 mask distinct variations by age', United States Census Bureau. Available online at: https://www.census.gov/library/stories/2022/04/fertility-rates-declined-for-younger-women-increased-for-older-women.html (accessed April 5, 2023).

Nabi, R. L., Walter, N., Oshidary, N., Endacott, C. G., Love-Nichols, J., Lew, Z. J., et al. (2020). Can emotions capture the elusive gain-loss framing effect? a meta-analysis. *Commun. Res.* 47, 1107–1130. doi: 10.1177/0093650219861256

Noar, S. M., Benac, C. N., and Harris, M. S. (2007). Does tailoring matter? meta-analytic review of tailored print health behavior change interventions. *Psychol. Bull.* 133, 673–693. doi: 10.1037/0033-2909.133.4.673

O'Keefe, D. J. (2015). *Persuasion: Theory and Research.* London: Sage Publications.

O'Keefe, D. J., and Hoeken, H. (2021). Message design choices don't make much difference to persuasiveness and can't be counted on-not even when moderating conditions are specified. *Front. Psychol.* 12, 664160. doi: 10.3389/fpsyg.2021.664160

Palan, S., and Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *J. Behav. Exp. Finance.* 17, 22–27. doi: 10.1016/j.jbef.2017.12.004

Pei, R., Schmälzle, R., Kranzler, E. C., O'Donnell, M. B., and Falk, E. B. (2019). Adolescents' neural response to tobacco prevention messages and sharing engagement. *Am. J. Prev. Med.* 56, S40–S48. doi: 10.1016/j.amepre.2018.07.044

Petty, R. E., Cacioppo, J. T., Petty, R. E., and Cacioppo, J. T. (1986). *The Elaboration Likelihood Model of Persuasion.* (New York: Springer), 1–24.

Rains, S. A. (2020). Big data, computational social science, and health communication: a review and agenda for advancing theory. *Health Commun.* 35, 26–34. doi: 10.1080/10410236.2018.1536955

Rashkin, H., Celikyilmaz, A., Choi, Y., and Gao, J. (2020). *PlotMachines: Outline-Conditioned Generation With Dynamic Plot State Tracking. arXiv [Preprint].* Available online at: http://arxiv.org/abs/2004.14967 (accessed October 15, 2022).

Reimers, N., and Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv [Preprint].* Available online at: http://arxiv.org/abs/1908.10084 (accessed September 3, 2022).

Rhodes, N., and Ewoldsen, D. R. (2013). "Outcomes of persuasion: Behavioral, cognitive, and social," in *The SAGE Handbook of Persuasion: Developments in Theory and Practice,* eds., J. P. Dillard and L. Shen (Los Angeles, CA: Sage Publications, Inc.), 53–69.

Rice, R. E., and Atkin, C. K. (2012). *Public Communication Campaigns.* Los Angeles, CA: SAGE Publications, Inc.

Rofail, D., Colligs, A., Abetz, L., Lindemann, M., and Maguire, L. (2012). Factors contributing to the success of folic acid public health campaigns. *J. Public Health.* 34, 90–99. doi: 10.1093/pubmed/fdr048

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ili,ć, S., Hesslow, D., et al. (2022). *Bloom: A 176b-parameter open-access multilingual language model. arXiv [Preprint].* Available online at: http://arxiv.org/abs/2211.05100 (accessed November 9, 2022).

Schmälzle, R., and Meshi, D. (2020). Communication neuroscience: Theory, methodology and experimental approaches. *Commun. Methods Meas.* 14, 105–124. doi: 10.1080/19312458.2019.1708283.

Schmälzle, R., Renner, B., and Schupp, H. T. (2017). Health risk perception and risk communication. *Policy Insights Behav. Brain Sci.* 4, 2. doi: 10.1177/2372732217720223

Schmälzle, R., and Wilcox, S. (2022). Harnessing artificial intelligence for health message generation: The folic acid message engine. *J. Med. Internet Res.* 24, e28858. doi: 10.2196/28858

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Scholl, T. O., and Johnson, W. G. (2000). Folic acid: Influence on the outcome of pregnancy. *Am. J. Clin. Nutr.* 71, 1295S–303S. doi: 10.1093/ajcn/71.5.1295s

Shi, J., Poorisat, T., and Salmon, C. T. (2018). The use of social networking sites (SNSs) in health communication campaigns: review and recommendations. *Health Commun.* 33, 49–56. doi: 10.1080/10410236.2016.1242035

Snscrape (2021). *snscrape: A Social Networking Service Scraper in Python.* Available online at: https://github.com/JustAnotherArchivist/snscrape (accessed May 31, 2022).

Snyder, L. B. (2007). Health communication campaigns and their impact on behavior. *J. Nutr. Educ. Behav.* 39, S32–S40. doi: 10.1016/j.jneb.2006.09.004

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., et al. (2019). *Release Strategies and the Social Impacts of Language Models.* arXiv [Preprint]. Available online at: http://arxiv.org/abs/1908.09203 (accessed October 13, 2022).

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. *Proc. Int. Conf. WWW.* 25, 613–624. doi: 10.1145/2872427.2883081

Thackeray, R., Neiger, B. L., Hanson, C. L., and McKenzie, J. F. (2008). Enhancing promotional strategies within social marketing programs: use of Web 2.0 social media. *Health Promot. Pract.* 9, 338–343. doi: 10.1177/1524839908325335

Theron, D. (2022). *Getting Started With Bloom.* Available online at: https://towardsdatascience.com/getting-started-with-bloom-9e3295459b65 (accessed July 28, 2022).

Thompson, T. L., and Harrington, N. G. (2021). The *Routledge handbook of health communication.* New York, NY: Routledge.

Tunstall, L., von Werra, L., and Wolf, T. (2022). *Natural Language Processing With Transformers.* Sebastopol: O'Reilly Media, Inc.

Turner, M., and Rains, S. (2021). Guilt appeals in persuasive communication: a meta-analytic review. *Commun. Stud.* 72, 684–700. doi: 10.1080/10510974.2021.1953094

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process Syst.* 30, 25. doi: 10.48550/arXiv.1706.03762

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods.* 17, 261–272. doi: 10.1038/s41592-020-0772-5

von Platen, P. (2020). *How to Generate Text: Using Different Decoding Methods for Language Generation With Transformers.* Available online at: https://huggingface.co/blog/how-to-generate (accessed October 2, 2022).

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., and Lambert, N. (2020). *TRL: Transformer Reinforcement Learning.* GitHub. Available online at: https://github.com/lvwerra/trl

Wang, X., Chen, L., Shi, J., and Peng, T.-Q. (2019). What makes cancer information viral on social media? *Comput. Hum. Behav.* 93, 149–156. doi: 10.1016/j.chb.2018.12.024

Wang, X., Qin, X., Demirtas, H., Li, J., Mao, G., Huo, Y., et al. (2007). Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. *Lancet.* 369, 1876–1882. doi: 10.1016/S0140-6736(07)60854-X

Weber, R., Fisher, J. T., Hopp, F. R., and Lonergan, C. (2018). Taking messages into the magnet: method–theory synergy in communication neuroscience. *Commun. Monogr.* 85, 81–102. doi: 10.1080/03637751.2017.1395059

Wicks, P. (2014). The ALS ice bucket challenge—can a splash of water reinvigorate a field? *Amyotroph. Lateral Scler. Frontotemporal Degener.* 15, 479–480. doi: 10.3109/21678421.2014.984725

Willoughby, J. F., and Noar, S. M. (2022). Fifteen years after a 10-year retrospective: the state of health mass mediated campaigns. *J. Health Commun.* 2, 1–13. doi: 10.1080/10810730.2022.2110627

Witte, K., Meyer, G., and Martell, D. (2001). *Effective Health Risk Messages: A step-by-step Guide.* Thousand Oaks: Sage Publications, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: state-of-the-Art natural language processing. *Proc. Conf. EMNLP.* 2, 38–45. doi: 10.18653/v1/2020.emnlp-demos.6

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). Explainable AI: a brief survey on history, research areas, approaches and challenges. *NLPCC.* 1, 563–574. doi: 10.1007/978-3-030-32236-6_51

Yang, J., and Mackert, M. (2021). The effectiveness of CDC's Rx awareness campaigns on reducing opioid stigma: Implications for health communication. *Health Commun.* 21, 1–10. doi: 10.1080/10410236.2021.1982561

Yeomans, M., Kantor, A., and Tingley, D. (2019). The politeness package: detecting politeness in natural language. *R. J.* 10, 489. doi: 10.32614/RJ-2018-079

Zhou, S., Silvasstar, J., Clark, C., Salyers, A. J., Chavez, C., and Bull, S. S. (2023). An artificially intelligent, natural language processing chatbot designed to promote COVID-19 vaccination: a proof-of-concept pilot study. *Digit. Health.* 9, 20552076231155679. doi: 10.1177/20552076231155679