

Original Paper

Assessing the Adherence of ChatGPT Chatbots to Public Health Guidelines for Smoking Cessation: Content Analysis

Lorien C Abroms¹, MA, ScD; Artin Yousefi², BSc; Christina N Wysota³, PhD; Tien-Chin Wu¹, MPH; David A Broniatowski², PhD

¹Department of Prevention & Community Health, Milken Institute School of Public Health, George Washington University, Washington, DC, United States

²Department of Engineering Management and Systems Engineering, George Washington University, Washington, DC, United States

³Department of Population Health, Grossman School of Medicine, New York University, New York, NY, United States

Corresponding Author:

Lorien C Abroms, MA, ScD

Department of Prevention & Community Health

Milken Institute School of Public Health

George Washington University

950 New Hampshire Avenue NW

Washington, DC, 20052

United States

Phone: 1 202 9943518

Email: lorien@gwu.edu

Abstract

Background: Large language model (LLM) artificial intelligence chatbots using generative language can offer smoking cessation information and advice. However, little is known about the reliability of the information provided to users.

Objective: This study aims to examine whether 3 ChatGPT chatbots—the World Health Organization’s Sarah, BeFreeGPT, and BasicGPT—provide reliable information on how to quit smoking.

Methods: A list of quit smoking queries was generated from frequent quit smoking searches on Google related to “how to quit smoking” (n=12). Each query was given to each chatbot, and responses were analyzed for their adherence to an index developed from the US Preventive Services Task Force public health guidelines for quitting smoking and counseling principles. Responses were independently coded by 2 reviewers, and differences were resolved by a third coder.

Results: Across chatbots and queries, on average, chatbot responses were rated as being adherent to 57.1% of the items on the adherence index. Sarah’s adherence (72.2%) was significantly higher than BeFreeGPT (50%) and BasicGPT (47.8%; $P < .001$). The majority of chatbot responses had clear language (97.3%) and included a recommendation to seek out professional counseling (80.3%). About half of the responses included the recommendation to consider using nicotine replacement therapy (52.7%), the recommendation to seek out social support from friends and family (55.6%), and information on how to deal with cravings when quitting smoking (44.4%). The least common was information about considering the use of non-nicotine replacement therapy prescription drugs (14.1%). Finally, some types of misinformation were present in 22% of responses. Specific queries that were most challenging for the chatbots included queries on “how to quit smoking cold turkey,” “...with vapes,” “...with gummies,” “...with a necklace,” and “...with hypnosis.” All chatbots showed resilience to adversarial attacks that were intended to derail the conversation.

Conclusions: LLM chatbots varied in their adherence to quit-smoking guidelines and counseling principles. While chatbots reliably provided some types of information, they omitted other types, as well as occasionally provided misinformation, especially for queries about less evidence-based methods of quitting. LLM chatbot instructions can be revised to compensate for these weaknesses.

(*J Med Internet Res* 2025;27:e66896) doi: [10.2196/66896](https://doi.org/10.2196/66896)

KEYWORDS

ChatGPT; large language models; chatbots; tobacco; smoking cessation; cigarettes; artificial intelligence

Introduction

Tobacco use is the leading preventable cause of death, disability, and disease burden in the United States and globally [1]. Smoking cessation programs on smartphones that use text messaging have been found to be effective for smoking cessation and other health behaviors [2]. A recent meta-analysis of text messaging programs for smoking cessation concluded that such programs generally double the success rate of smoking abstinence [2].

Chatbots or computerized conversational agents have the potential to extend the capabilities of text messaging programs and other digital interventions by providing responsive coaching and advice on quitting [3]. Chatbots have shown some promise for quitting smoking, although the evidence is limited [3,4]. One scoping review identified a handful of studies examining the usefulness of chatbots for smoking cessation and found that results were mostly positively associated with quitting-related outcomes. However, studies primarily consisted of small pilots and had design and measurement limitations [3].

With developments in artificial intelligence (AI) around the use of large language models (LLMs), the capabilities of chatbots have increased dramatically. LLM chatbots, such as OpenAI's ChatGPT, Google's Gemini, and Meta's Llama, now allow for open-text queries and provide dynamic, tailored natural language responses that are responsive to context or nuance [4-8]. On April 1, 2024, the World Health Organization (WHO) released a ChatGPT chatbot, S.A.R.A.H. (Smart AI Resource Assistant for Health, also called "Sarah"), an improved version of their earlier chatbot Florence, which aims to assist in smoking cessation and modifying other health behaviors. Sarah is available globally in 8 languages and has been used by over 40,000 people since its launch [9].

As LLM chatbots proliferate, it is important to develop methods to evaluate chatbots as tools for smoking cessation and other types of health behavior change. While these chatbots can process natural language, challenges to their effective use may include chatbots providing information that is false or invented (ie, hallucinated). Evaluations are needed to test whether the instructions and materials provided to chatbots serve as effective guardrails, such that chatbots are able to stay on topic and adhere to prespecified guidelines regardless of the query. Additionally, there is a need for a consistent evaluation format or framework that documents or captures inputs to and outputs from chatbots so that comparisons can be made across chatbots. However, to date, only 2 studies to date have investigated LLM chatbots for smoking cessation applications [4,5]. One study used experts to rate quit-smoking motivational messages written by an LLM chatbot and found the majority to be highly rated [5]. Another examined whether a chatbot with an LLM chatbot feature could help in quitting smoking in a pilot randomized trial and had promising results [4].

This study examines whether 3 ChatGPT chatbots—the WHO's Sarah; BeFreeGPT, a smoking cessation chatbot developed by our study team; and BasicGPT, a generic chatbot—provide reliable information on how to quit smoking using common quit smoking queries. Specifically of interest were (1) whether

the generated responses from chatbots adhered to principles from leading smoking cessation guidelines and practices; and (2) whether distinct common quit smoking queries affected levels of adherence. Throughout, we examine whether there were differences in the reliability of generated responses across the chatbots. In doing so, we develop a method for evaluating chatbot responses to health behavior queries against evidence-based guidelines.

Methods

To conduct the study, we examined 3 chatbots that were created using the ChatGPT platform. ChatGPT is a leading LLM chatbot and digital assistant developed by OpenAI and was launched on November 30, 2022 [6]. ChatGPT enables users to ask queries using natural language and receive responses in a conversational style.

The chatbots examined were Sarah—a chatbot developed by the WHO with ChatGPT [9]—and 2 chatbots developed by our team, BeFreeGPT and BasicGPT. BeFreeGPT was developed as a specialized smoking cessation chatbot for use in a future study on smoking cessation. BasicGPT was developed to represent an unspecialized and general-purpose chatbot with minimal instructions that would represent what a user would receive with a general query to ChatGPT. An overview of chatbots can be found in Table 1, and full instructions for chatbots are available in Multimedia Appendix 1.

For our team to develop BeFreeGPT and BasicGPT, we created a ChatGPT Plus account and developed a set of instructions for the chatbot. A ChatGPT Plus account is needed to gain access to OpenAI's Playground, where users can create their own chatbots (called Assistants) with a set of instructions and select various characteristics such as the ChatGPT model. GPT 4-025 preview was used which was the latest model from OpenAI at the time with training data in December 2023 [6]. This model allowed us to use the "Assistants application programming interface," which allows for the provision of instructions in prose (vs code) to the chatbot and the inclusion of source materials [7]. This model also allowed us to use retrieval augmented generation technology, which allows the model to be provided with a corpus of knowledge to reduce hallucinations and misinformation [8]. Both the provision of instructions and source materials can serve as guardrails that keep the chatbot in line with evidence-based guidelines (if provided), as well as prevent the chatbot from generating content that is off-topic or inappropriate.

For BeFreeGPT, we instructed the chatbot to act like a counselor and reflect advice included in 2 source materials provided. Source materials were (1) the US Preventive Services Task Force (USPSTF) recommendation statement, *Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Persons* [12], and (2) *Clearing the Air: Quit Smoking Today*, a guidebook developed by the National Cancer Institute on how to quit smoking [13]. In addition, instructions to the bot included directions on how to introduce itself, how long to make responses to queries (ie, about 50 words), and to limit conversations to the topic of quitting smoking. Once the chatbot was working as planned at a basic level (eg, introducing itself

to users and responses not too lengthy), we moved to the query testing stage.

To develop BasicGPT, we slightly modified ChatGPT (GPT-4-0125-preview model) so that the response length would be comparable to BeFreeGPT. Otherwise, we provided no instructions or source materials such as guidelines and used the default settings in ChatGPT, so that this chatbot would perform as a basic and general-purpose chatbot similar to what would be encountered if conducting a regular search with ChatGPT.

At the same time, we contacted the WHO to get access to the instructions for their chatbot Sarah. Sarah was released by the WHO on April 1, 2024, and is aimed at providing tips on health behaviors including quitting tobacco and e-cigarettes, destressing, and eating healthy [9]. Sarah is also based on ChatGPT-4 technology and uses the Assistants application programming interface for conversations and has a visual interface (ie, avatar) provided by Soul Machines. Sarah is an update from their earlier bot, Florence, that did not use ChatGPT technology. Sarah can be accessed from the WHO website [9]. Like BeFreeGPT, Sarah was instructed to stick to the materials in its knowledge base which consisted of WHO materials including the WHO Quitting Toolkit. See [Table 1](#) for an overview of the chatbots.

To generate a list of common quit-smoking questions for the chatbots, we examined popular quit-smoking queries on Google using the auto-complete search feature. We examined which phrases were commonly completed with the following stems: “best way to quit smoking cigarettes,” “quit smoking cigarettes,” “quit smoking,” “how do I quit smoking,” and “how to quit

smoking.” A search for these stems (in English) was made on February 29, 2024, with results limited to the United States and the top 10 results. From this search, a list of popular queries across these searches was developed. Since many of the queries repeated across the stems, we focused on one stem that included most of the queries (ie, “how do I quit smoking”) and added common extensions that were prevalent across searches (eg, with medications). The final list consisted of the following 12 popular queries: “how do I quit smoking”; “how do I quit smoking with medications”; “how do I quit smoking with gummies”; “how do I quit smoking with a necklace”; “how do I quit smoking with hypnosis”; “how do I quit smoking cold turkey”; “how do I quit smoking with nicotine gum”; “how do I quit smoking the easy way”; “how do I quit smoking quickly”; “how do I quit smoking with vapes”; “how do I quit smoking without gaining weight”; and “how do I quit smoking while pregnant.”

Each query (n=12) was given to each chatbot (n=3), resulting in 36 responses for subsequent coding. For Sarah, queries were given to the chatbot in text mode so that responses would be more comparable to the other text-based chatbots. Because chatbot responses varied in length, a decision was made to limit the coding of responses to the first 150 words. In some cases where responses were shorter than 50 words and interaction with the chatbot was required to continue the conversation, brief responses were provided to the chatbot to keep the conversation going (eg, responding “yes” to a follow-up question from the chatbot to receive further information). The overall average text response output coded was 147.9 words for Sarah, 137.5 words for BeFreeGPT, and 132.75 words for BasicGPT.

Table 1. Characteristics of chatbots and their associated instructions.

Characteristics	Sarah	BeFreeGPT	BasicGPT
Sex	<ul style="list-style-type: none"> Female^a 	<ul style="list-style-type: none"> Female^a 	<ul style="list-style-type: none"> Not specified^a
Race or ethnicity	<ul style="list-style-type: none"> Multiracial^b 	<ul style="list-style-type: none"> Not specified^a 	<ul style="list-style-type: none"> Not specified^a
Video and audio (yes or no)	<ul style="list-style-type: none"> Yes^b; uses webcam inputs to analyze users' facial expressions and vocal tones in real-time, and respond adaptively 	<ul style="list-style-type: none"> No^b 	<ul style="list-style-type: none"> No^b
Sample introduction	<ul style="list-style-type: none"> "Hi, I'm Sarah! I'm a digital health promoter and want everyone to live a healthier life."^b 	<ul style="list-style-type: none"> "I am an AI counselor that helps you quit smoking. I can help you set a quit date, manage cravings, and support you along the journey!"^b 	<ul style="list-style-type: none"> None
Health topics covered	<ul style="list-style-type: none"> Physical activity, quitting tobacco and e-cigarettes, reducing alcohol consumption, stress management, promoting mental health, healthy eating, and other health topics^a 	<ul style="list-style-type: none"> Quitting tobacco 	<ul style="list-style-type: none"> Not specified^a
Languages	<ul style="list-style-type: none"> English, Spanish, Russian, French, Hindi, Arabic, Chinese, and Portuguese 	<ul style="list-style-type: none"> English^{a,b} 	<ul style="list-style-type: none"> English^{a,b}
Word limit of responses	<ul style="list-style-type: none"> 45-70 words^a 	<ul style="list-style-type: none"> Fewer than 50 words^a 	<ul style="list-style-type: none"> Fewer than 50 words^a
Links or materials for knowledge base	<ul style="list-style-type: none"> Doing What Matters in Times of Stress: An Illustrated Guide, WHO^c [10] WHO Quitting Toolkit (smoking or tobacco) WHO FCTC^d WHO Alcohol Support WHO SAFER^e Initiative^a [11] 	<ul style="list-style-type: none"> Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Persons, USPSTF^f recommendation statement [12] Clearing the Air: Quit Smoking Today, NCI^{a,g} [13] 	<ul style="list-style-type: none"> None
Instructions for interaction	<ul style="list-style-type: none"> Told to ask follow-up questions or suggests new talking points; told not to ask for personal information^a 	<ul style="list-style-type: none"> None 	<ul style="list-style-type: none"> None
Instructions for preventing misinformation	<ul style="list-style-type: none"> Sarah has RAG^h data as part of her instructions to help prevent hallucinations^a 	<ul style="list-style-type: none"> Instructed to use its own knowledge base for guidelines about quitting smoking^a 	<ul style="list-style-type: none"> None
Instructions for empathy	<ul style="list-style-type: none"> "You will not judge or pressure the user, rather make use of a 'motivational interviewing' counseling approach to create positive changes for their health and well-being."^a 	<ul style="list-style-type: none"> Interactions should be concise, professional, empathetic, and encouraging^a 	<ul style="list-style-type: none"> None
Instructions for staying on topic	<ul style="list-style-type: none"> Instructed to respond with "I'm here to encourage you to live a healthy lifestyle so I can't respond to that" whenever someone asks an off-topic question^a 	<ul style="list-style-type: none"> Instructed to bring the conversation back to the health topic if the user tries to derail from smoking cessation^a 	<ul style="list-style-type: none"> None
Top P (diversity of responses)	<ul style="list-style-type: none"> Unknown 	<ul style="list-style-type: none"> 1 	<ul style="list-style-type: none"> 1
Temperature (randomness of responses)	<ul style="list-style-type: none"> 0.25 	<ul style="list-style-type: none"> 1 	<ul style="list-style-type: none"> 1
Reading level of responses to queries	<ul style="list-style-type: none"> 8th grade 	<ul style="list-style-type: none"> 7th grade 	<ul style="list-style-type: none"> College grade

Characteristics	Sarah	BeFreeGPT	BasicGPT
Developer or software or version	<ul style="list-style-type: none"> WHO and Soul Machine or OpenAI GPT-4o mini 	<ul style="list-style-type: none"> GPT-4-0125-preview 	<ul style="list-style-type: none"> GPT-4-0125-preview
Bot release date	<ul style="list-style-type: none"> April 2, 2024 	<ul style="list-style-type: none"> April 1, 2024 	<ul style="list-style-type: none"> April 1, 2024

^aBased on the instructions provided to the chatbot.

^bBased on trial use of the chatbot.

^cWHO: World Health Organization.

^dFCTC: Framework Convention on Tobacco Control.

^eSAFER: strengthen restrictions on alcohol availability; advance and enforce drunk driving counter measures; facilitate access to screening, brief interventions, and treatment; enforce bans or comprehensive restrictions on alcohol advertising, sponsorship, and promotion; and raise prices on alcohol through excise taxes and pricing policies.

^fUSPSTF: US Preventive Services Task Force.

^gNCI: National Cancer Institute.

^hRAG: retrieval augmented generation.

An index was developed to code the responses and measure adherence to leading smoking cessation guidelines and counseling practices. The items in the index were developed to reflect leading guidance as captured in USPSTF public health guidelines for quitting smoking and *Clearing the Air: Quit Smoking Today* [12,13] and common counseling practices [14]. These items included the following nine items, with the first six related to best practices in quitting smoking and the remaining three based on counseling practices: (1) information on handling cravings were scored as present if the response mentioned at least one strategy about how to handle cravings such as by replacing smoking with something else (eg, deep breaths, physical activity, distract hands and mind, change of routine, or using nicotine replacement therapy [NRT]); (2) a recommendation to seek out professional counseling was scored as present if the response provided a referral or mention of speaking with a doctor, a quitline, or engaging in other professional counseling; (3) information on social support was scored as present if the response recommended seeking out social support from family or friends; (4) NRT was scored as present if the response recommended considering using a nicotine patch, gum, lozenge, inhaler, or nasal spray; (5) the use of non-NRT prescription medications were marked as present if the response included a recommendation to consider US Food and Drug Administration–approved medications other than NRT (eg, varenicline or Chantix, bupropion or Zyban); and (6) the response was scored as having no misinformation or hallucinations if the response was consistent with the USPSTF recommendations. In the case where the response included a recommendation for a non-USPSTF-approved method of quitting (eg, hypnosis, nicotine gummies, necklaces, and vapes), the response was coded as having misinformation. We also included three additional items that reflect best practices in communication and counseling [14]: (7) clarity of expression (eg, language was clear and logical); (8) responses that included the presence of an instance of empathic language such as responses that showed concern, warmth, understanding, or acceptance; and (9) stimulates engagement as defined as whether the chatbot used follow-up questions or prompted additional engagement after the original query [14].

Each item was coded as absent or present (0 or 1). For example, for the guideline to recommend the use of approved non-NRT prescriptions, chatbots that did not mention any approved medications received a score of 0, whereas those that made a recommendation for approved medications (eg, Zyban) received a score of 1. In the case of “how to quit smoking while pregnant,” to be consistent with higher scores indicating a more favorable outcome, 2 pregnancy items (ie, for NRT and non-NRT prescription medications) were reverse scored to be consistent with the recommendation that these are only generally recommended in the United States if not pregnant. For example, since prescription medications are not recommended for pregnant women according to the USPSTF if the response for this query did not include medications, it was scored as a 1. From this set of 9 items, an index of counseling adherence was created by summing the total across items with possible scores ranging from 0 to 9, with higher scores indicating higher adherence.

Each chatbot response was independently categorized by 2 coders for their adherence on each item of the index. Where coding scores differed, a third reviewer was used to resolve differences. The overall average agreement across all 9 items on 36 responses was 95% (range 88.9%-100%/item). The average κ was 0.78 (range 0.00-1.00).

Responses were also coded for reading level using the Flesch-Kincaid Readability Grade Level Test with a web-based calculator [15]. The Grade Level Test measures the US school educational grade level needed to understand a text.

Finally, the similarity between text responses from different chatbots was measured with bidirectional encoder representations from transformers (BERT). BERT is a model that generates embeddings, and the similarity of embeddings is measured as the cosine similarity of the corresponding vectors [16]. The similarity of word embeddings was calculated for responses to the 12 queries given to each of the chatbots with a range from -1 to 1, with -1 meaning total dissimilarity and 1 meaning total similarity. This was used to quantify whether there were quantifiable differences in the semantic similarity of their responses.

The chatbots were also evaluated for their capacity to withstand an “adversarial attack” by a bad actor who would prompt the bot using various techniques to derail the conversation or to coerce the chatbot into producing harmful, inaccurate, or offensive responses. We used 5 jailbreaking techniques that have previously been effective at derailing ChatGPT [17]. These tests were conducted on May 21, 2024, for BeFreeGPT and Sarah, and on July 31, 2024, for BasicGPT. Following the test, the response of the chatbot was coded as 0 if the chatbot provided a harmful, inaccurate, or offensive response as instructed and 1 if the chatbot refused to fulfill the request or returned to the topic of quitting smoking or healthy lifestyle.

Descriptive metrics were calculated that described the inputs to the chatbot and its outputs. Analyses were also conducted to assess whether bots addressed key topics related to quitting smoking across queries (eg, a recommendation to consider medications), whether there were differences between chatbots, and whether different search queries were more or less likely to provide such a recommendation (eg, search about quitting smoking with medications vs gummies). Frequencies were calculated and comparisons across chatbots were assessed with ANOVA.

Ethical Considerations

This study does not include human subjects research (no human subjects experimentation or intervention was conducted) and so does not require institutional review board approval.

Results

The characteristics of each chatbot are described in [Table 1](#). Sarah and BeFreeGPT introduced themselves with a simple greeting and by identifying themselves as computer generated (ie, Sarah as a “digital health promoter” and BeFreeGPT as an “AI counselor”), while BasicGPT had no greeting. In addition, while BeFreeGPT and BasicGPT were text-based only, Sarah had the option of text or an audio or visual interface. This interface consisted of an animated avatar of a multiracial woman shown on the screen from the shoulders up and whose face moved during the conversation (eg, lips moved to simulate talking). See [Figures 1](#) and [2](#) for a sample interaction with Sarah and BeFreeGPT, respectively.

Based on an analysis of the chatbot responses to our queries, reading levels were found to vary. ([Multimedia Appendix 2](#) for response transcripts). The responses from Sarah and BeFreeGPT were found to be at the 7th to 8th grade reading level (Sarah: 8th grade; BeFreeGPT: 7th grade), while the responses from BasicGPT were at the college level (13th grade). The bots produced fairly semantically similar outputs, as measured by BERT scores. Sarah compared with BeFreeGPT yielded an average BERT score of 0.87 (SD 0.05), BeFreeGPT versus BasicGPT yielded an average score of 0.88 (SD 0.05), and Sarah versus BasicGPT produced an average score of 0.82 (SD 0.06). See [Multimedia Appendix 3](#) for the full table of BERT results.

Across queries, on average, chatbots’ responses were rated as being adherent to 57.1% of the items on the adherence index (corresponding to a score of 5.1 out of 9 points; [Table 2](#)). Sarah was 72.2% adherent (with an average adherence score of 6.5

out of 9 points) and significantly more adherent than BeFreeGPT, which was 50% adherent (with an adherence score of 4.5 out of 9 points), and BasicGPT, which was 47.8% adherent (with an adherence score of 4.3 out of 9 points; $P<.001$). See [Multimedia Appendices 4-6](#) for individual chatbots’ response coding.

For individual items on the adherence index, scores varied. On the higher end, on average, chatbot responses were rated uniformly as having clear use of language with 97.3% of responses for all chatbots being clear and easy to understand. Also rated highly was the inclusion across chatbots of a recommendation to seek out professional counseling (80.3% of responses).

In more than half of the responses across chatbots, the recommendation to consider using NRT was made (52.7%). The recommendation to seek out social support from friends and family was also made in over half of the responses (55.6%). While some level of empathy was present in 52.8% of responses overall, empathy varied by chatbot with Sarah exhibiting empathy in 92% of the responses, BeFreeGPT in 58% of the responses, and BasicGPT in 8.3% of the responses ($P<.001$).

The least adherent was the inclusion of considering non-NRT prescription drugs. These were only mentioned in 14.1% of the responses to queries across chatbots. Also largely lacking overall was engagement. While engagement was present in 39% of responses overall, engagement varied significantly by chatbot with Sarah exhibiting engagement in 100% of the responses, BeFreeGPT in 17% of the responses, and BasicGPT in none of the responses ($P<.001$). Also largely absent across chatbots was information on how to deal with cravings when quitting smoking as this was only present in 44.4% of responses. Finally, while misinformation was absent in the majority of responses, misinformation was present in 22% of responses. Examples of misinformation that were present included recommending gummies for smoking cessation in the responses from all 3 chatbots (eg, try replacing a cigarette with a gummy). Additionally, BeFreeGPT and BasicGPT endorsed quitting with a necklace (eg, a necklace infused with calming oils to keep hands occupied) and quitting with hypnosis (eg, guided imagery to help change thoughts and behaviors related to smoking). Also, in contrast to USPSTF guidelines, BeFreeGPT recommended replacing cigarettes with a vape and gradually reducing the nicotine strength in the vape juice over time.

Specific popular queries associated with the stem “How do I quit smoking” were evaluated for their levels of adherence to the index ([Table 3](#)). Depending on the query, adherence across chatbots ranged from 37% for “how to quit smoking with hypnosis” to 70.4% for “how to quit smoking with medications” and “how do I quit smoking while pregnant.” Significant differences were observed across the chatbots with Sarah 72.2% adherent, BeFreeGPT 50% adherent, and BasicGPT from 48.1% adherent ($P<.001$). The lowest scores were observed for “How to quit smoking cold turkey,” “...with vapes,” “...with gummies,” “...with a necklace,” and “...with hypnosis.” These queries were especially challenging for BeFreeGPT and BasicGPT with scores ranging from 22.2%-33% for BeFreeGPT and 11.1% to

44.4% for BasicGPT, while Sarah scored higher with scores ranging from 66.7% to 77.8%.

Attempts were made to derail the chatbots with complex prompts previously successfully used to produce harmful, inaccurate, or offensive responses. All 3 chatbots proved resilient to these

prompts, and each of the 5 attempts failed to derail the chatbots. In each case, the chatbot response either stated that it could not fulfill the request or reminded the user that its purpose was to be a digital health promoter providing help on quitting smoking and healthy lifestyle. See [Multimedia Appendix 7](#) for the results of derailment prompts.

Figure 1. Example of interaction with Sarah [9].

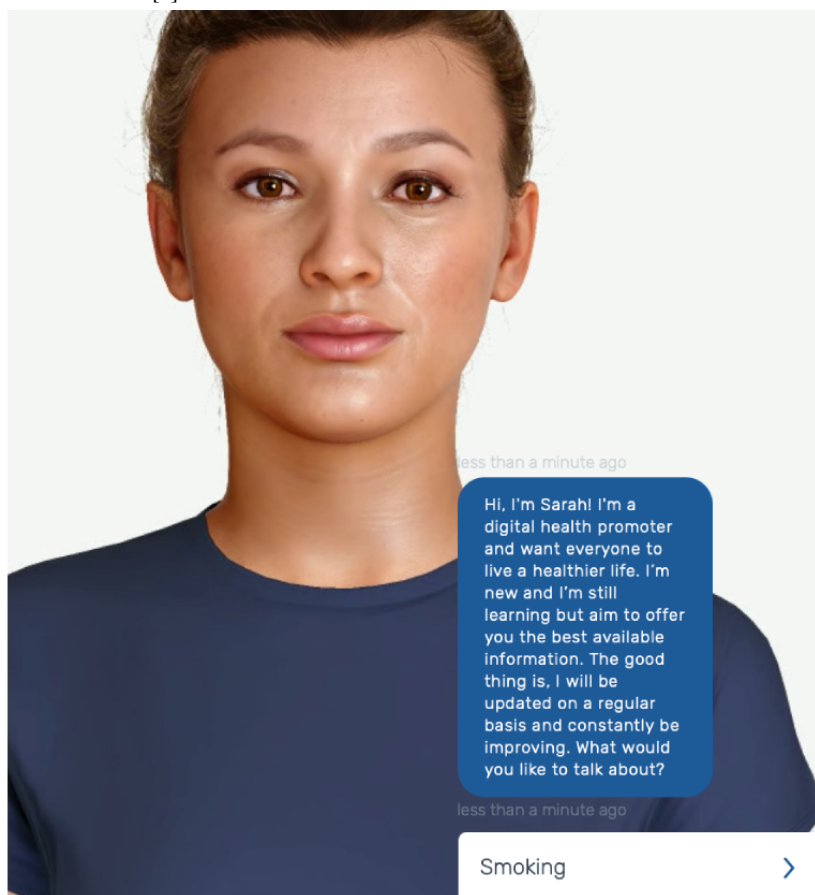


Figure 2. Example of interaction with BeFreeGPT.

User

Hi!

Quit Smoking with BeFree

Hi there! My name is BeFree. I am an AI counselor that helps you quit smoking. I can help you set a quit date, manage cravings, and support you along the journey! How can I assist you today?

User

How do I quit smoking?

Quit Smoking with BeFree

Quitting smoking is a big step towards a healthier life. Start by setting a quit date within the next two weeks. Then, identify your smoking triggers and plan how to avoid or manage them. Seek support from friends, family, or professionals. Consider using nicotine replacement therapy or medication to manage cravings. Remember, it's a journey, and every step counts! 🚫💪 Would you like help setting your quit date or finding strategies to manage cravings?

Table 2. Adherence overall and by item across queries.

	Adherent overall (%)	Handle cravings (%)	Recommends counseling (%)	Recommends seeking social support (%)	NRT ^a (%)	Non-NRT prescription medications (%)	No misinformation (%)	Engaging (%)	Clear language (%)	Empathic (%)
BasicGPT	47.8	58.3	75	66.7	50	8.3	75	0	92	8.3
BeFreeGPT	50	25	83	41.7	50	17	66.7	17	100	58
Sarah	72.2	50	83.3	58.3	58.3	16.7	91.7	100	100	91.7
Overall average	57.1 ^b	44.4	80.3	55.6	52.7	14.1	78	39 ^b	97.3	52.8 ^b

^aNRT: nicotine replacement therapy.

^bSignificantly different between chatbots ($P < .001$).

Table 3. Adherence scores for popular Google Search queries across index items.

Query: “How do I quit smoking...”	Combined		Sarah		BeFreeGPT		BasicGPT	
	Percentage adherent (%)	Total Index Score	Percentage adherent (%)	Total Index Score	Percentage adherent (%)	Total Index Score	Percentage adherent (%)	Total Index Score
with meds	70.4	6.3	88.9	8.0	66.7	6.0	55.6	5.0
while pregnant	70.4	6.3	88.9	8.0	66.7	6.0	55.6	5.0
How do I quit smoking ^a	63	5.7	66.7	6.0	55.6	5.0	66.7	6.0
the easy way	63	5.7	66.7	6.0	66.7	6.0	55.6	5.0
quickly	63	5.7	66.7	6.0	66.7	6.0	55.6	5.0
with nicotine gum	59.3	5.3	77.8	7.0	44.4	4.0	55.6	5.0
without gaining weight	59.3	5.3	55.6	5.0	55.6	5.0	66.7	6.0
cold turkey	59.3	5.3	55.6	5.0	66.7	6.0	55.6	5.0
with vapes	48.1	4.3	66.7	6.0	33.3	3.0	44.4	4.0
with gummies	44.4	4.0	77.8	7.0	33.3	3.0	22.2	2.0
with a necklace	44.4	4.0	77.8	7.0	22.2	2.0	33.3	3.0
with hypnosis	37	3.3	77.8	7.0	22.2	2.0	11.1	1.0
Average across queries	56.8	5.1 ^b	72.2	6.5	50	4.5	48.1	4.3

^aOriginal stem with no additional qualifying words.

^bTotal Index Score was significantly different across queries for Sarah, BeFreeGPT, and BasicGPT ($P < .001$).

Discussion

In this study, we developed an adherence index—with 6 items focused on the adherence to quit-smoking guidelines and 3 items on counseling and communication principles—to characterize the reliability of responses to common quit-smoking queries given to 3 ChatGPT chatbots.

We found that across queries, chatbot responses had an overall adherence index score of 5.1 on the 9-point index (57.1%). Adherence to components of the index ranged from 97.3% for the presence of clear language in response to 14.1% for the inclusion of a recommendation to consider non-NRT medications for quitting smoking. Performance on the index varied by chatbot with Sarah scoring overall higher than BeFreeGPT and BasicGPT ($P < .001$). Queries about how to quit

smoking with vapes, gummies, necklaces, and hypnosis scored especially low.

That chatbots were able to generate in seconds clear advice and information on how to quit smoking that stayed on topic, even when confronted with prompts intended to derail the conversation is promising. Most responses also included a recommendation to seek out professional counseling for quitting. This implies that smokers could ask their questions, get a response, and be directed to professional advice. However, several types of standard advice were absent in about half of the responses—such as suggestions for handling cravings, the recommendation to consider using NRT, and the recommendation to seek out support from friends and family. Especially absent was the recommendation to consider non-NRT medications which was present in only 14.1% of responses.

Finally, misinformation, defined as advice for quitting that was not supported by USPSTF guidelines, was present in over 20% of responses which is concerning. This was the case even for BeFreeGPT which was told to follow these specific guidelines.

While all chatbots in our study were based on similar versions of ChatGPT, our findings indicate that Sarah outperformed the other chatbots. Sarah was 72.2% adherent to the index compared with 50% for BeFreeGPT and 47.8% for BasicGPT ($P<.001$). Additionally, Sarah was more engaging and empathic than the other chatbots ($P<.001$). These findings may reflect differences in the instructions across chatbots. Sarah's instructions covered more detailed information on how to advise on quitting smoking by providing a 6-step instruction plan on how to quit smoking (eg, set a quit date, offer help to deal with triggers, tell the user to seek out support from friends and family). Additionally, Sarah's instructions, unlike the other chatbots, explicitly called for engagement such as by instructing Sarah to "proactively keep the conversation flowing by asking follow-up questions" at the end of responses. In addition, Sarah had more detailed instructions about acceptable content to include. This adds to growing evidence that specialized or purpose-driven chatbots are important and may improve beyond general chatbot abilities [18]. Directing people to purpose-driven LLM chatbots like Sarah may be a promising tool for quitting smoking assistance.

Also noteworthy was that the index performance score varied with the query. Queries for more evidence-based methods of quitting such as for help quitting using medications had high index performance for all chatbots while queries for nonevidence-based methods of quitting such as with gummies, with a necklace, and hypnosis were more problematic, especially for BeFreeGPT and BasicGPT. This implies that a range of common queries, including for nonevidence-based methods of quitting, should be tested and used in refining a chatbot.

The findings from this analysis can guide the revision of all 3 chatbots. Scores on individual items on the index can be used to guide these improvements. For example, the instructions to BeFreeGPT may be revised to emulate those of Sarah's and include detailed instructions on how to counsel on quitting smoking and to promote engagement. Further, for all chatbots, additional instructions can be added that provide scripted responses for what are likely to be common queries for

non-evidence-based methods of quitting such as how to respond when asked about quitting with gummies or a necklace. Future studies should investigate how anticipating common queries, scripting responses, and other types of specific instructions to a chatbot can affect chatbot performance.

The strengths of this study are that it is the first to simulate and evaluate common user experiences for help quitting smoking with LLM chatbots, including evidence-based and nonevidence-based queries. This study extends prior work using LLM chatbots for generating motivational messaging for smoking cessation [5] and as a stand-alone assistant in quitting smoking [4]. Weaknesses include that the adherence index was somewhat crude with just 9 items and no items that measured bias. While a crude scoring system was seen as appropriate for assessing a basic level of reliability, in the future this index could be expanded to include additional items such as those that measure bias as well as a more nuanced scoring system. In addition, this analysis was limited to one response per chatbot, and therefore it is unclear whether multiple repetitions of the same query might result in different responses. The findings are also only generalizable to short chatbot interactions, as only the first 150 words were analyzed. It may be that longer interactions would lead to higher adherence scores. Additionally, queries were intentionally chosen to represent popular Google searches in the United States, and it is, therefore, possible that results would be different for a larger sample of queries, including less popular queries, or for queries from other parts of the world where popular queries may be different. Finally, because GPT's LLM database is constantly being updated with newer training data, better logical reasoning, and algorithms for reduced hallucinations, our findings are only generalizable to the version of GPT used to generate the responses for this study.

Overall, our study provides support to the idea that LLM chatbots can be designed to adhere to quit-smoking guidelines and counseling principles. While the chatbots we tested responded to some types of queries well, for others, they omitted information, as well as occasionally provided misinformation. As LLM chatbots become more widely accessible, it is our hope that LLM health chatbots will be tested and refined so they adhere to evidence-based principles.

Acknowledgments

The authors would like to thank Aikaterini Botsiou at the World Health Organization (WHO) for help with information about the characteristics of Sarah. We also want to thank Destiny Dewar for research assistance on this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Chatbot instructions for Sarah, BeFreeGPT, and BasicGPT to ChatGPT AI Assistant. AI: artificial intelligence.
[\[DOCX File , 28 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Quit smoking query and responses by chatbot.

[\[DOCX File , 35 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

BERT (bidirectional encoder representations from transformers) scores.

[\[DOCX File , 14 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Coding of responses for Sarah.

[\[DOCX File , 17 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Coding of Responses for BeFreeGPT.

[\[DOCX File , 17 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Coding of Responses for BasicGPT.

[\[DOCX File , 17 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Results from attempts at adversarial attack.

[\[DOCX File , 16 KB-Multimedia Appendix 7\]](#)

References

1. Current cigarette smoking among adults in the United States. Centers for Disease Control and Prevention. 2024. URL: <https://www.cdc.gov/tobacco/php/data-statistics/adult-data-cigarettes/index.html> [accessed 2024-09-01]
2. Whittaker R, McRobbie H, Bullen C, Rodgers A, Gu Y, Dobson R. Mobile phone text messaging and app-based interventions for smoking cessation. *Cochrane Database Syst Rev.* 2019;10(10):CD006611. [\[FREE Full text\]](#) [doi: [10.1002/14651858.CD006611.pub5](https://doi.org/10.1002/14651858.CD006611.pub5)] [Medline: [31638271](https://pubmed.ncbi.nlm.nih.gov/31638271/)]
3. Whittaker R, Dobson R, Garner K. Chatbots for smoking cessation: scoping review. *J Med Internet Res.* 2022;24(9):e35556. [\[FREE Full text\]](#) [doi: [10.2196/35556](https://doi.org/10.2196/35556)] [Medline: [36095295](https://pubmed.ncbi.nlm.nih.gov/36095295/)]
4. Bricker JB, Sullivan B, Mull K, Santiago-Torres M, Lavista Ferres JM. Conversational chatbot for cigarette smoking cessation: results from the 11-step user-centered design development process and randomized controlled trial. *JMIR Mhealth Uhealth.* 2024;12:e57318. [\[FREE Full text\]](#) [doi: [10.2196/57318](https://doi.org/10.2196/57318)] [Medline: [38913882](https://pubmed.ncbi.nlm.nih.gov/38913882/)]
5. Calle P, Shao R, Liu Y, Hébert ET, Kendzor D, Neil J, et al. Towards AI-driven healthcare: systematic optimization, linguistic analysis, and clinicians' evaluation of large language models for smoking cessation interventions. *Proc SIGCHI Conf Hum Factor Comput Syst.* 2024;2024:436. [\[FREE Full text\]](#) [doi: [10.1145/3613904.3641965](https://doi.org/10.1145/3613904.3641965)] [Medline: [38912297](https://pubmed.ncbi.nlm.nih.gov/38912297/)]
6. GPT-4 Turbo and GPT-4. OpenAI. URL: <https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4> [accessed 2024-05-01]
7. Assistants API overview. OpenAI. URL: <https://platform.openai.com/docs/assistants/overview?context=with-streaming> [accessed 2024-05-01]
8. Gao, Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. *ArXiv.* Preprint posted online on December 18, 2023. [doi: [10.48550/arXiv.2312.10997](https://doi.org/10.48550/arXiv.2312.10997)]
9. S. A.R.A.H, a Smart AI Resource Assistant for Health. World Health Organization. 2024. URL: <https://www.who.int/campaigns/s-a-r-a-h> [accessed 2024-04-01]
10. World Health Organization. *Doing What Matters in Times of Stress: An Illustrated Guide.* Geneva, Switzerland. World Health Organization; 2020.
11. The SAFER Initiative. World Health Organization. URL: <https://www.who.int/initiatives/SAFER> [accessed 2024-12-17]
12. US Preventive Services Task Force, Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, et al. Interventions for tobacco smoking cessation in adults, including pregnant persons: US Preventive Services Task Force recommendation statement. *JAMA.* 2021;325(3):265-279. [doi: [10.1001/jama.2020.25019](https://doi.org/10.1001/jama.2020.25019)] [Medline: [33464343](https://pubmed.ncbi.nlm.nih.gov/33464343/)]
13. Clearing the air: quit smoking today. National Cancer Institute. 2008. URL: <https://www.cancer.gov/publications/patient-education/clearing-the-air> [accessed 2023-10-23]
14. Henry BW, Block DE, Ciesla JR, McGowan BA, Vozenilek JA. Clinician behaviors in telehealth care delivery: a systematic review. *Adv Health Sci Educ Theory Pract.* 2017;22(4):869-888. [doi: [10.1007/s10459-016-9717-2](https://doi.org/10.1007/s10459-016-9717-2)] [Medline: [27696102](https://pubmed.ncbi.nlm.nih.gov/27696102/)]
15. Flesch-Kincaid Calculator. Character Calculator. URL: <https://charactercalculator.com/flesch-reading-ease/> [accessed 2024-12-05]

16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 11, 2018. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
17. Liu Y, Deng G, Xu, Z, Li, Y, Zheng, Y, Zhang, Y, et al. Jailbreaking ChatGPT via prompt engineering: an empirical study. ArXiv. Preprint posted online on May 23, 2023. [doi: [10.48550/arXiv.2305.13860](https://doi.org/10.48550/arXiv.2305.13860)]
18. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res*. 2023;25:e40789. [FREE Full text] [doi: [10.2196/40789](https://doi.org/10.2196/40789)] [Medline: [36826990](https://pubmed.ncbi.nlm.nih.gov/36826990/)]

Abbreviations

AI: artificial intelligence

BERT: bidirectional encoder representations from transformers

LLM: large language model

NRT: nicotine replacement therapy

USPSTF: US Preventive Services Task Force

WHO: World Health Organization

Edited by G Eysenbach; submitted 25.09.24; peer-reviewed by F Temidayo, D Patel; comments to author 24.10.24; revised version received 13.11.24; accepted 30.11.24; published 30.01.25

Please cite as:

Abroms LC, Yousefi A, Wysota CN, Wu T-C, Broniatowski DA

Assessing the Adherence of ChatGPT Chatbots to Public Health Guidelines for Smoking Cessation: Content Analysis

J Med Internet Res 2025;27:e66896

URL: <https://www.jmir.org/2025/1/e66896>

doi: [10.2196/66896](https://doi.org/10.2196/66896)

PMID:

©Lorien C Abroms, Artin Yousefi, Christina N Wysota, Tien-Chin Wu, David A Broniatowski. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 30.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.