

# **Ensuring Accuracy and Equity: A Cross-Language Evaluation of Vaccination Information from ChatGPT and CDC**

Saubhagya Joshi, Eun Bin Ha, Andee Amaya, Melissa Mendoza, Yonaira Rivera,  
Vivek K. Singh

Submitted to: JMIR Formative Research  
on: May 26, 2024

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

*Table of Contents*

**Original Manuscript..... 5**  
**Supplementary Files..... 24**  
    Figures ..... 25  
        Figure 1..... 26

# Ensuring Accuracy and Equity: A Cross-Language Evaluation of Vaccination Information from ChatGPT and CDC

Saubhagya Joshi<sup>1</sup> MSci; Eun Bin Ha<sup>1</sup> MA; Andee Amaya<sup>1</sup>; Melissa Mendoza<sup>1</sup>; Yonaira Rivera<sup>1</sup> PhD; Vivek K. Singh<sup>1</sup> PhD

<sup>1</sup>School of Communication & Information Rutgers University New Brunswick US

## Corresponding Author:

Vivek K. Singh PhD  
School of Communication & Information  
Rutgers University  
4 Huntington Street  
New Brunswick  
US

## Abstract

**Background:** In the digital age, Large Language Models (LLMs) like ChatGPT have emerged as important sources of healthcare information. Their interactive capabilities offer promise for enhancing health access, particularly for groups facing traditional barriers such as insurance and language constraints. Despite their growing public health use, with millions of medical queries processed weekly, the quality of LLM-provided information remains inconsistent. Prior studies have predominantly assessed ChatGPT's English responses, overlooking the needs of non-English speakers in the U.S. This study addresses this gap by evaluating the quality and linguistic parity of vaccination information from ChatGPT and the CDC, emphasizing health equity.

**Objective:** This research aims to assess the quality and language equity of vaccination information provided by ChatGPT and the CDC in English and Spanish. It highlights the critical need for cross-language evaluation to ensure equitable health information access for all linguistic groups.

**Methods:** We conducted a comparative analysis of ChatGPT's and CDC's responses to frequently asked vaccination questions in both languages. The evaluation encompassed quantitative and qualitative assessments of accuracy, readability, and understandability. Accuracy was gauged by the perceived level of misinformation, readability by the Flesch-Kincaid score and grade level, and understandability by items from the NIH's PEMAT instrument.

**Results:** The study found that both ChatGPT and CDC provided mostly accurate and understandable responses. However, readability scores often exceeded the American Medical Association's recommended levels, particularly in English. CDC responses outperformed ChatGPT in readability across both languages. Notably, some Spanish responses appeared to be direct translations from English, leading to unnatural phrasing. The findings underscore the potential and challenges of utilizing ChatGPT for healthcare access.

**Conclusions:** ChatGPT holds potential as a health information resource, but requires improvements in readability and linguistic equity to be truly effective for diverse populations. Crucially, the default user experience with ChatGPT, typically encountered by those without advanced language and prompting skills, can significantly shape health perceptions. This is vital from a public health standpoint, as the majority of users will interact with LLMs in their most accessible form. Ensuring that default responses are accurate, understandable, and equitable is imperative for fostering informed health decisions across diverse communities.

(JMIR Preprints 26/05/2024:60939)

DOI: <https://doi.org/10.2196/preprints.60939>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.  
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/60939>, the full text will be available to the public.



## Original Manuscript

# Ensuring Accuracy and Equity: A Cross-Language Evaluation of Vaccination Information from ChatGPT and CDC

Saubhagya Joshi, MI<sup>1</sup>, Eunbin Ha, MA<sup>1</sup>, Andee Amaya<sup>1</sup>, Melissa Mendoza<sup>1</sup>, Yonaira M Rivera, PhD, MPH<sup>1</sup>, and Vivek K Singh, PhD<sup>1,\*</sup>

<sup>1</sup> School of Communication & Information, Rutgers University, New Brunswick, NJ, USA.

\*Corresponding Author: [v.singh@rutgers.edu](mailto:v.singh@rutgers.edu)

4 Huntington St., New Brunswick, NJ 08901, USA  
Phone: 848-932-7588

## ABSTRACT

**Background:** In the digital age, Large Language Models (LLMs) like ChatGPT have emerged as important sources of healthcare information. Their interactive capabilities offer promise for enhancing health access, particularly for groups facing traditional barriers such as insurance and language constraints. Despite their growing public health use, with millions of medical queries processed weekly, the quality of LLM-provided information remains inconsistent. Prior studies have predominantly assessed ChatGPT's English responses, overlooking the needs of non-English speakers in the U.S. This study addresses this gap by evaluating the quality and linguistic parity of vaccination information from ChatGPT and the CDC, emphasizing health equity.

**Objective:** This research aims to assess the quality and language equity of vaccination information provided by ChatGPT and the CDC in English and Spanish. It highlights the critical need for cross-language evaluation to ensure equitable health information access for all linguistic groups.

**Methods:** We conducted a comparative analysis of ChatGPT's and CDC's responses to frequently asked vaccination questions in both languages. The evaluation encompassed quantitative and qualitative assessments of accuracy, readability, and understandability. Accuracy was gauged by the

perceived level of misinformation, readability by the Flesch-Kincaid score and grade level, and understandability by items from the NIH's PEMAT instrument.

**Results:** The study found that both ChatGPT and CDC provided mostly accurate and understandable responses. However, readability scores often exceeded the American Medical Association's recommended levels, particularly in English. CDC responses outperformed ChatGPT in readability across both languages. Notably, some Spanish responses appeared to be direct translations from English, leading to unnatural phrasing. The findings underscore the potential and challenges of utilizing ChatGPT for healthcare access.

**Conclusion:** ChatGPT holds potential as a health information resource, but requires improvements in readability and linguistic equity to be truly effective for diverse populations. Crucially, the default user experience with ChatGPT, typically encountered by those without advanced language and prompting skills, can significantly shape health perceptions. This is vital from a public health standpoint, as the majority of users will interact with LLMs in their most accessible form. Ensuring that default responses are accurate, understandable, and equitable is imperative for fostering informed health decisions across diverse communities.

**Key words:** Vaccination, Health Equity, Language, Multilingualism, Algorithms

## INTRODUCTION

There is a growing recognition of the role of information as a crucial determinant of health [1]. Globally, Google witnesses more than 100 million daily health-related searches. Similarly, Open AI's ChatGPT experiences over a billion monthly visits and is increasingly utilized in medical contexts [2]. One survey reports more than 80% of US respondents had used a chatbot in 2023, and another suggests that despite prohibitions on medical use by vendors, millions of medical queries are

submitted on a weekly basis by users of OpenAI alone [3,4]. Such publicly available large language models (LLMs) like ChatGPT can be a promising source of healthcare information. Individuals may readily derive benefits from straightforward queries and interactive dialogue when seeking medical advice or making health-related decisions. However, evaluations on the quality of LLM responses still show conflicting results. Recent studies reveal that human experts perceived ChatGPT's responses to be accurate, relevant, and comprehensive [5,6]. Despite the potential utility of LLMs, scholars pose concerns about the substitution of ChatGPT and other LLMs for professionals' medical advice [7,8]. Empirical evidence still exists for plausible-sounding yet inaccurate or fraudulent outcomes, as well as limited readability, semantic repetition, or coherence loss in lengthy passages[7,9-12].

Notably, there is a critical need to examine how LLMs respond to controversial topics like vaccination. Communication and media environments can be regarded as determinants of hesitant vaccine attitudes [13]. Given the emergence of LLMs such as ChatGPT as channels of health information, their responses may shape users' perceptions of the vaccine and healthcare decision-making. Recent work suggests that ChatGPT exhibits a notably precise, clear, easy-to-understand, and unbiased tone in its delivery of vaccination [8,14,15]. Yet, most research has solely focused on the quality of ChatGPT responses in English, limiting considerations of linguistic equity.

Despite scholars paying relatively little attention to multilingual ChatGPT responses, there are remarkable disparities in vaccination coverage and attitudes by racial and language groups in the United States. For example, Latinos and Blacks were more hesitant to COVID-19 vaccines compared to White populations [16]. Latino parents have also shown a high rate of COVID-19 vaccine resistance and uncertainty [17,18]. Furthermore, Latino adults have reported lower HPV vaccination rates (40.6%) when compared to whites and Blacks, respectively (49.7% and 45.8%) [19]. Given the need to increase information access and equity surrounding vaccines in non-English languages for those with low English preference, we argue that we should pay attention to the different linguistic



features of ChatGPT responses - particularly as it relates to Spanish, the most spoken non-English language in the U.S. [20]. It is necessary to comprehensively evaluate LLM's multilingual outcomes with consideration to both response quality and equity. However, this area is severely understudied. A notable exception is our previous work, which found the disparity in vaccine hesitancy-related responses across different languages [21]. Though we found that vaccine-hesitancy was the most in English responses and the least in Spanish responses, the study was limited to comparing single-word numerical responses to vaccination survey questions in English, Spanish, and French. To ensure qualified and equitable health information in multilingual LLMs, we need more research that examines the cross-language health content across diverse dimensions.

The current study aims to expand our prior work by exploring whether popular LLMs, such as ChatGPT, provide reliable health information in multiple languages. We specifically aim to compare childhood vaccination FAQ responses from the Centers for Disease Control and Prevention (CDC) and ChatGPT across accuracy, understandability, and readability dimensions in both English and Spanish using both quantitative and qualitative approaches. By doing so, we hope to identify and address systemic inequities that may exist in LLMs, so as to better promote equity and inclusion in the development of technologies that impact access to quality health information and communication.

## **METHODS**

### **Data collection**

In order to examine the variation in ChatGPT response, we used a set of 16 FAQ questions on the CDC website in both English and Spanish [22,23]. Chat completion API from OpenAI was used to interface with the GPT-4 engine to obtain textual responses to each of the questions [24]. These questions yielded short free-form responses. Figure 1 shows an example of a single response. This

process was repeated three times to avoid extreme single responses impacting the results dramatically. Overall, we had a total of 96 responses from ChatGPT (3 responses for each of the 16 questions in both languages) and 32 responses from CDC's FAQs (16 responses each in both languages) (128 responses in total).

Question 5		Answer
English	Can vaccines overload my baby's immune system?	CDC Vaccines do not overload the immune system. Every day, a healthy baby's immune system successfully fights off thousands of germs. Antigens are parts of germs that cause the body's immune system to go to work to build antibodies, which fight off diseases. \nThe antigens in vaccines come from the germs themselves, but the germs are weakened or killed so they cannot cause serious illness. Even if babies receive several vaccinations in one day, vaccines contain only a tiny fraction of the antigens they encounter every day in their environment. Vaccines give your child the antibodies they need to fight off serious vaccine-preventable diseases.
		GPT No, vaccines do not overload your baby's immune system. Even newborns have the ability to respond to a large number of vaccines concurrently. Your baby's immune system is capable of recognizing millions of different viruses and bacteria every day. The tiny amount of dead or inactivated bacteria or virus, or pieces of them in each vaccine is just a tiny fraction of what their system manages regularly. So, the vaccines are not a burden for your child's immune system. In fact, they help to build it stronger.
Spanish	¿Pueden las vacunas sobrecargar el sistema inmunitario de mi bebé?	CDC Las vacunas no sobrecargan el sistema inmunitario. Cada día, el sistema inmunitario de los bebés sanos combate con éxito miles de microbios. Los antígenos son partes de los microbios que activan al sistema inmunitario para que cree anticuerpos, los cuales luchan contra las enfermedades. \nLos antígenos en las vacunas provienen de los microbios mismos; sin embargo, están debilitados o muertos, por lo cual no pueden hacer que se enfermen gravemente. Incluso si los bebés reciben varias vacunas en un día, las vacunas contienen solo una muy pequeña fracción de la cantidad de antígenos que se encuentran en su ambiente diario. Las vacunas le proveen a su hijo los anticuerpos que necesita para combatir enfermedades graves, prevenibles a través de ellas.
		GPT No, las vacunas no sobrecargan el sistema inmunitario de tu bebé. El sistema inmunológico de los bebés es capaz de manejar una gran cantidad de antígenos (sustancias que provocan una respuesta inmunitaria). Desde el nacimiento, los bebés entran en contacto con numerosos antígenos todos los días a través de la comida, la suciedad, y las infecciones, y su sistema inmunológico es capaz de manejarlo. De hecho, las vacunas representan solo una pequeña fracción de los antígenos a los que un bebé se expone. Las vacunas ayudan a "entrenar" el sistema inmunológico de tu bebé para reconocer y combatir virus y bacterias específicos, lo que les proporciona una protección muy necesaria contra enfermedades graves.

**Figure 1.** A sample question from the FAQs and sample answers as obtained from CDC and ChatGPT.

Responses were evaluated on three dimensions: (1) accuracy, (2) understandability, and (3) readability. These dimensions were identified based on their importance in literature and the potential impact that they can have in contentious health contexts [5-10, 25-28].

*Accuracy* was assessed using a 3-point item to measure the level of misinformation in responses (1=no misinformation, 2=some misinformation, 3=high misinformation).

*Understandability* was assessed using the National Library of Medicine's Health Education Materials

Assessment Tool [25], which consists of 8 yes/no items adapted from the Patient Education Materials Assessment Tool's understandability domain [26]. A final score was derived as an average of the eight items and represented as a percentage, with higher scores meaning higher understandability. This instrument is available in English and Spanish (see Supplementary Material for more information). Two bilingual coders (A.A. & M.M.) independently scored blinded responses for accuracy and understandability. Interrater reliability was high for both domains (97.66% agreement in accuracy,  $\kappa=0.01$ ; 98.92% agreement in understandability,  $\kappa=0.86$ ). *Readability* was assessed using Flesch-Kincaid (FK) readability scores for English and Flesch-Huerta index for Spanish [27,28] where scores between 0 and 100 are scaled to grade levels from fifth grade (90-100) to professional (0-10). Data was extracted and summarized using Python and transformed into spreadsheets; basic statistical tests were conducted using Microsoft Excel (Version 2312).

We also conducted a qualitative analysis of the responses to explore nuances within and between languages. For each question, coders provided feedback on the overall tone of the responses for each language separately and as a whole. These were discussed with a third bilingual study team member (Y.R.) to assess any similarities and differences in responses between languages.

## RESULTS

Table 1 shows average word, sentence, and syllable counts. On average, Spanish used more words, sentences, and syllables per response. ChatGPT responses were generally more verbose than CDC responses. Additionally, ChatGPT sentence count ranges were more variable than those of CDC responses for both English (ChatGPT: 1-22 and CDC: 2-7) and Spanish (ChatGPT: 2-24 and CDC: 2-8).

**Table 1.** Mean[range] of verbosity measures

	English		Spanish		Total	
	CDC	ChatGPT	CDC	ChatGPT	CDC	ChatGPT
Sentence count	4.06 [2 - 7]	7 [1 - 22]	4.38 [2 - 8]	7.15 [2 - 24]	4.22 [2 - 8]	7.07 as [1 - 24]
Syllable count	116.75 [53 - 241]	172.79 [63 - 515]	189.94 [82 - 371]	270.5 [100 - 672]	153.34 [53 - 371]	221.65 [63 - 672]

## Accuracy

We found that all responses had high accuracy: CDC responses in both languages had no misinformation, while only three responses were coded as having some misinformation by one coder each (due to nuanced responses lacking clarifying context). None of the responses were rated as having high misinformation.

## Understandability

Responses also rated high in understandability (Table 2). There were no significant differences in understandability between CDC and ChatGPT responses within or between languages, suggesting ChatGPT responses were in alignment with CDC messaging.

**Table 2.** Mean [range] of different attributes in English, Spanish, and Total for CDC and ChatGPT

	English		Spanish		Total	
	CDC	ChatGPT	CDC	ChatGPT	CDC	ChatGPT
N	16	16 x 3	16	16 x 3	2 x 16	2 x 16 x 3
Accuracy	1 [1 - 1]	1.02 [1 - 2]	1 [1 - 1]	1.01 [1 - 2]	1 [1 - 1]	1.02 [1 - 2]
Understandability	95.65 [85.7 - 100]	95.87 [57.1 - 100]	98.83 [75 - 100]	95.7 [71.43 - 100]	97.24 [75 - 100]	95.79 [57.14 - 100]
Readability score	48.1 [26.61-65.93]	42.52 [22.36-62.1]	74.92 [53.65-89.02]	69.63 [54.56-80.72]	61.51 [26.61-89.02]	56.08 [22.36-80.72]
Grade level	12.13 [8.5 - 16]	12.84 [8.5 - 16]	7.19 [6 - 11]	7.93 [6 - 11]	9.66 [6 - 16]	10.39 [6 - 16]

## Readability

There was significant variation in readability scores between the responses within languages and between CDC and ChatGPT (Table 2). On average, ChatGPT responses had lower readability scores than the CDC responses, regardless of language (56.08 vs. 61.51,  $P=.03$ ,  $df=23$ ). Meanwhile, when comparing responses by language, English responses had lower average readability scores than Spanish responses for both CDC and ChatGPT (CDC: 48.10 vs. 74.92,  $P<.001$ ,  $df=29$ ; ChatGPT: 42.52 vs. 69.63,  $P<.001$ ,  $df=24$ ) (Table 2 and Table 3). When comparing grade levels, English responses for both CDC and ChatGPT were significantly higher than those in Spanish (CDC: 12th grade English vs. 7th grade Spanish,  $P<.001$ ,  $df=26$ ; ChatGPT: 13th grade English vs. 8th grade Spanish,  $P<.001$ ,  $df=26$ ). Given the AMA's recommendation that patient materials be written at the sixth-grade level [29], we assessed the odds of each platform in satisfying this requirement. Overall, CDC responses were 13.57 times more likely to satisfy the sixth-grade level than ChatGPT responses ( $Q=5.62$ ,  $df=1$ ,  $P=.018$ ; Fisher's Exact  $P=.014$ ). This was similar among Spanish language responses (CDC 15.64 times higher than ChatGPT;  $Q=5.86$ ,  $df=1$ ,  $P=.016$ ; Fisher's Exact  $P=.012$ ). We did not observe any significant differences between CDC and ChatGPT English responses. (Details on post-hoc pairwise across different groups are available in Supplementary Information.) In order to verify the effect of metric across different groups, post-hoc two-tail pairwise t-tests at 95% significance were conducted as shown in Table 3.

**Table 3.** Significance of difference between groups using t-tests

	EN & ES				GPT				CDC			
	CDC	GPT	t-stat	P	EN	ES	t-stat	P	EN	ES	t-stat	P
Readability score	61.51	56.08	2.32	0.03	42.52	69.63	-13.03	<.001	48.10	74.92	-7.87	<.001
Grade level	9.66	10.39	-1.96	0.06	12.84	7.93	16.19	<.001	12.13	7.19	8.71	<.001

Note: 2 tail significance at 95%

## Qualitative observations

When qualitatively comparing responses in both languages, several differences were observed. ChatGPT would oftentimes respond in list format, making it somewhat easier to read responses comparing risks and benefits, side effects, or other reasons to vaccinate. ChatGPT would also provide additional information and examples to questions. When specifically looking at Spanish responses, we observed some Spanish text using English words in quotations (e.g., “herd immunity” and “fake”). We also noticed that, despite better readability scores than English, some Spanish responses would use less colloquial words (e.g., “proporcionar” instead of “proveer” or “patógeno” instead of “infección”), while others had sentence structures that resembled a word-by-word English translation (e.g., “Retrasar las vacunas puede poner en riesgo a su hijo (y a otros) de contraer enfermedades que podrían haberse prevenido” rather than “Al retrasar vacunas, su hijo y otros pueden contraer enfermedades que podrían prevenirse”, which might be more commonly used by a native Spanish speaker).

## DISCUSSION

This study evaluated the quality and equity of LLM’s outcomes. Our findings show that ChatGPT provided adequate levels of accuracy and understandability to vaccine-related questions in both English and Spanish. Past results on the accuracy of ChatGPT have been mixed. While some recent work exploring ChatGPT’s responses to health-related content also suggests little to no misinformation is being shared [6,7,21], others suggest significant levels of misinformation [10]. Our results suggest that in the context of vaccine FAQs, ChatGPT provides information with high accuracy. Furthermore, ChatGPT’s easy-to-understand responses could be an accessible resource for users with limited health literacy or with limited access to healthcare services, thereby contributing to efforts to address health disparities and inequities. This may be particularly useful to Spanish-speaking individuals in areas where there is limited access to language-concordant health education.

However, our study also found some challenges in the quality and equity of LLM’s outcomes.

Firstly, there is a need to moderate ChatGPT's responses, particularly in English, to adhere to recommended reading levels. The AMA recommended reading levels for healthcare material are at 6th grade or below. However, ChatGPT's English responses to childhood vaccination questions often necessitated reading skills well above that of a 6th-grade level. This was also the case with CDC. Both scenarios merit attention since failure to adhere to acceptable readability standards could act as a potential barrier to health information. Ease of reading may lead to enhanced knowledge of health, thereby playing a crucial role in taking functional health literacy [30].

Secondly, we observed that the representations of words in ChatGPT occasionally exhibited the linguistic patterns of English in the Spanish responses. While these were not incorrectly written, some Spanish responses seem to be translated directly from English text or used less common Spanish vocabulary. There were also several instances where the Spanish response had English words in quotation marks, even though a Spanish equivalent exists. Although it may not merely translate word-for-word between English and other languages, recent evidence found that the multilingual language model, Llama-2, is primarily dependent on English to understand the meanings of ideas across different languages [31]. While LLMs use multilingual training data, English is still the most dominant language in their training dataset [32]. Indeed, LLMs are mostly skilled in English-based tasks and are also proficient in translating from English to non-English languages. However, such verbatim translations of English could fail to capture adequate domain-specific jargon and nuances of cultural context [33] and lead to a lack of information support for those with preferences for non-English languages to obtain public health information. Therefore, English dependency in the training data of LLMs could be a potential risk to healthcare equity. In the future, more inclusion of more diverse datasets from other languages including minority dialects should be considered in training data.

We note that the results presented in this work focus on those obtained without any prompt engineering. For instance, carefully crafted prompt engineering could impact the readability of

ChatGPT-generated responses. Our study centers on the natural querying behavior exhibited by the majority of ChatGPT users, who typically engage with the system in a conversational manner, similar to their interactions with traditional search engines like Google. This is particularly true for vulnerable populations seeking health information, who may not be aware of or utilize prompt engineering techniques. While ChatGPT has 100 million active weekly users [34], there is no clear data on how many of these users employ prompt engineering. However, it is reasonable to assume that a significant portion of these users, especially those from non-technical backgrounds, with limited English proficiency, and those under medical duress, interact with ChatGPT without advanced prompting strategies. Our research illuminates the natural user experience and the inherent readability of ChatGPT's responses, which holds significant implications for public health informatics. The differences in responses under typical user conditions are noteworthy and warrant further examination, particularly in light of multilingual users who may be at higher risk of health inequities.

The work also intersects with recent legislation and policy discussions around guardrails needed for automated AI systems. According to the Executive Order [35], "irresponsible use [of AI] could exacerbate societal harms such as fraud, discrimination, bias, and disinformation..." LLM implementations like ChatGPT are classified as "automated systems" that have a direct effect on decision-making for communities due to continuous data exchange, as opposed to "passive computing infrastructure" [36]. Therefore, it is imperative that proper guardrails are put in place to maintain fairness and equity of health information by continuously monitoring the metrics like accuracy and quality of, and access to, health information produced by LLMs like ChatGPT for everybody including underserved communities. Similarly, one of the findings of the recent report [37(p54)] from PCAST for the President states that "Without proper benchmark metrics, validation procedures, and responsible practices, AI systems can give unreliable outputs whose quality is difficult to evaluate, and which could be harmful for a scientific field and its applications." Since



there is demonstrated disparity in the grade level of ChatGPT responses in different languages, it is imperative that thorough study of its impact in health information equity is conducted. In fact, the OMB issued a memo recommending “Minimum Practices for Rights-Impacting AI” [38(p21)] that involves identifying and assessing “AI’s impact on equity and fairness” and mitigating “algorithmic discrimination when it is present” by December 2024 and studies like ours are important in identifying as yet understudied dimensions of health equity, i.e., cross-language comparison of LLM responses in the health context.

The paper also has some limitations. It focuses on a single set of FAQs sourced from one agency (CDC) on a particular topic (vaccination). The results have only been evaluated on a single LLM technology (ChatGPT) at one time. As ChatGPT responses can vary over iterations, we have averaged them over three iterations. Our focus is limited to comparing two languages (English and Spanish) and future studies should consider more variations in languages, questionnaires, and information systems. However, beyond the results with a specific LLM or languages, this work aims to motivate an important area of research - equity audits across languages in different languages for health-centric conversations with automated agents.

## CONCLUSION

This study compared ChatGPT and CDC vaccination information in English and Spanish. We found that both sources were accurate and understandable, but ChatGPT had lower readability (higher grade level) than CDC in both languages. Further, some Spanish responses often appeared to be translations of the English ones, rather than independently generated, which could hinder information access for Spanish speakers. These findings suggest that ChatGPT is a promising tool for providing health information, but it needs to improve its readability and cultural sensitivity to ensure quality and equity. We recommend further research on the impact of natural language generation systems on public health outcomes and behaviors.

## FUNDING

This work was funded in part by a Rutgers School of Communication & Information Scholarly Futures grant and a New Jersey State Policy Lab grant.

## AUTHOR CONTRIBUTIONS

Conception and Development were done by S.J., E.H, Y.M.R., and V.K.S. Theory and Computations were led by S.J. with support from E.H., Y.M.R., and V.K.S. Qualitative Analysis was undertaken by A.A. and M.M. under guidance from Y.M.R. Y.M.R. and V.K.S. supervised and coordinated the work. All authors discussed the results and contributed to the final manuscript.

## CONFLICT OF INTEREST STATEMENT

The authors do not declare any Conflict of Interest.

## CODE AVAILABILITY STATEMENT

- Code to be made public upon acceptance.
- Survey coding:
  - [https://rutgers.ca1.qualtrics.com/jfe/form/SV\\_9Ep3HcaBWx2HRwq](https://rutgers.ca1.qualtrics.com/jfe/form/SV_9Ep3HcaBWx2HRwq)

## DATA AVAILABILITY

Attached and to be made public upon acceptance.

- data from ChatGPT as text in the zipped file: **gptData-16Q-EN\_ES.zip**
- coding and readability data as MS XLSX file: **gptCDCdata.xlsx**

## ABBREVIATIONS

<b>CDC</b>	Centers for Disease Control and Prevention
<b>EHR</b>	Electronic Health Record
<b>LLM</b>	Large Language Model
<b>NIH</b>	National Institutes of Health
<b>OMB</b>	Office of Management and Budget
<b>OSTP</b>	Office of Science and Technology Policy
<b>PCAST</b>	President's Council of Advisors on Science and Technology

## REFERENCES

1. Horn I. Our work toward health equity. Google. Sep 12, 2022. URL: <https://blog.google/technology/health/health-equity-summit-2022> [accessed 2024-02-22]
2. Duarte F. Number of ChatGPT users. Exploring topics. Apr 30, 2024. URL: <https://explodingtopics.com/blog/chatgpt-users> [accessed 2024-05-12]
3. Porter J. ChatGPT continues to be one of the fastest-growing services ever. The verge. Nov 6, 2023. URL: <https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference> [accessed 2024-05-12]
4. Leah L. What Do Your Customers Actually Think About Chatbots? [Research Study]. Userlike. Dec 15, 2022. URL: <https://www.userlike.com/en/blog/consumer-chatbot-perceptions> [accessed 2024-05-12]
5. Hamidi A, Roberts K. Evaluation of AI chatbots for patient-specific EHR questions. arXiv preprint arXiv:2306.02549; Jun 5, 2023. [doi:10.48550/arxiv.2306.02549]
6. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information.

- JNCI cancer spectrum. 2023;7(2). [doi:10.1093/jncics/pkad015]
7. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA oncology*. 2023;9(10):1437-1440. [doi:10.1001/jamaoncol.2023.2947]
  8. Sallam M, Salim NA, Al-Tammemi AB, et al. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. *Cureus*. 2023;15(2):e35029. [doi:10.7759/cureus.35029]
  9. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. [doi:10.7759/cureus.35179]
  10. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15(5):e39238. [doi:10.7759/cureus.39238]
  11. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*; May 28, 2020. [doi: 10.48550/arxiv.2005.14165]
  12. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of Medical Internet Research*. 2023;25:e46924. [doi:10.2196/46924]
  13. Rodrigues F, Block S, Sood S. What determines vaccine hesitancy: recommendations from childhood vaccine hesitancy to address COVID-19 vaccine hesitancy. *Vaccines*. 2022;10(1):80. [doi:10.3390/vaccines10010080]
  14. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*. 2023;11(7):1217. [doi:10.3390/vaccines11071217]
  15. Torun C, Sarmis A, Oguz A. Is ChatGPT an accurate and reliable source of information for

- patients with vaccine and statin hesitancy? *Medeniyet Medical Journal*. 2024;39(1):1–7. [doi:10.4274/MMJ.galenos.2024.03154]
16. Ndugga N, Hill L, Artiga S, Haldar S. Latest data on COVID-19 vaccinations by race/ethnicity. KFF. Jul 14, 2022. URL: <https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-by-race-ethnicity> [accessed 2024-02-11]
17. Fisher C, Bragard E, Madhivanan P. COVID-19 vaccine hesitancy among economically marginalized Hispanic parents of children under five years in the United States. *Vaccines*. 2023;11(3):599. [doi:10.3390/vaccines11030599]
18. Valier MR, Elam-Evans LD, Mu Y, et al. Racial and ethnic differences in COVID-19 vaccination coverage among children and adolescents aged 5–17 years and parental intent to vaccinate their children—national immunization survey—child COVID module, United States, December 2020–September 2022. *Morbidity and Mortality Weekly Report*. 2023;72(1):1-8. [doi:10.15585/mmwr.mm7201a1]
19. Black L, Boersma P. QuickStats: percentage of adults aged 18–26 years who ever received a human papillomavirus vaccine, by race and Hispanic origin and sex — national health interview survey, United States, 2019. *Morbidity and Mortality Weekly Report*. 2021;70(21):797. [doi:10.15585/mmwr.mm7021a5]
20. Dietrich S, Hernandez E. What languages do we speak in the United States?. United States Census Bureau. Dec 6, 2022. URL: <https://www.census.gov/library/stories/2022/12/languages-we-speak-in-united-states.html> [accessed 2024-02-22]
21. Joshi S, Ha E, Rivera YM, Singh VK. ChatGPT and vaccine hesitancy: a comparison of English, Spanish, and French responses. *Proceedings of the AMIA Informatics Summit*; 2024 Mar 18-21; Boston, Massachusetts. American Medical Informatics Association; 2024.
22. Centers for Disease Control and Prevention (CDC). Common questions about vaccines. URL:

- <https://www.cdc.gov/vaccines/parents/FAQs.html> [accessed 2023-08-07]
23. Centers for Disease Control and Prevention (CDC). Preguntas frecuentes sobre la vacunación infantil. URL: <https://www.cdc.gov/vaccines/parents/FAQs-sp.html> [accessed 2023-08-07]
24. OpenAI. Chat Completions API. URL: <https://platform.openai.com/docs/guides/text-generation/chat-completions-api> [accessed 2023-08-07]
25. National Library of Medicine (NIH). Health education materials assessment tool. URL: <https://medlineplus.gov/pdf/health-education-materials-assessment-tool.pdf> [accessed 2023-02-22]
26. Agency for Healthcare Research and Quality (AHRQ). The patient education materials assessment tool (PEMAT) and user's guide. URL: <https://www.ahrq.gov/health-literacy/patient-education/pemat.html> [accessed 2023-08-07]
27. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. URL: <https://apps.dtic.mil/sti/citations/ADA006655> [accessed 2023-01-24]
28. Separar en sílabas, contador de palabras y analizador en línea. URL: <https://www.separarensilabas.com/index.php> [accessed 2024-03-30]
29. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *Journal of Patient Experience*. 2021;8:2374373521998847. [doi:10.1177/2374373521998847]
30. Millar BC, Moore JE. Improving vaccine-related health literacy in parents: comparison on the readability of CDC Vaccine Information Statements (VIS) and Health and Human Services (HHS) patient-facing vaccine literature. *Therapeutic Advances in Vaccines and Immunotherapy*. 2021;9:25151355211047521. [doi: 10.1177/25151355211047521]
31. Wendler C, Veselovsky V, Monea G, West R. Do Llamas work in English? on the latent language of multilingual transformers. arXiv preprint arXiv:2402.10588; Feb 16, 2024.

[doi:10.48550/arxiv.2402.10588]

32. OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774; Mar 15, 2023. [doi:10.48550/arxiv.2303.08774]
33. Lai VD, Ngo NT, Veyseh AP, et al. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613; Apr 12, 2023. [doi:10.48550/arxiv.2304.05613]
34. Nerdynav. 107 Up-to-Date ChatGPT Statistics & User Numbers. Dec 13, 2022. URL: <https://nerdynav.com/chatgpt-statistics> [accessed 2024-03-14]
35. Executive Order 14110. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House. Oct 30, 2023. URL: <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence> [accessed 2024-05-12]
36. OSTP. Blueprint for an AI Bill of Rights. The White House. Oct, 2022. URL: <https://www.whitehouse.gov/ostp/ai-bill-of-rights> [accessed 2024-05-12]
37. PCAST. Supercharging Research: Harnessing Artificial Intelligence to Meet Global Challenges [Report to the President]. The White House. Apr, 2024. URL: [https://www.whitehouse.gov/wp-content/uploads/2024/04/AI-Report\\_Upload\\_29APRIL2024\\_SEND-2.pdf](https://www.whitehouse.gov/wp-content/uploads/2024/04/AI-Report_Upload_29APRIL2024_SEND-2.pdf) [accessed 2024-05-12]
38. OMB. Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence. The White House. Mar 28, 2024. URL: <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf> [accessed 2024-05-12]

## Supplementary Files



## Figures

A sample question from the FAQs and sample answers as obtained from CDC and ChatGPT.

Question 5		Answer
English	Can vaccines overload my baby's immune system?	<p><b>CDC</b></p> <p>Vaccines do not overload the immune system. Every day, a healthy baby's immune system successfully fights off thousands of germs. Antigens are parts of germs that cause the body's immune system to go to work to build antibodies, which fight off diseases. The antigens in vaccines come from the germs themselves, but the germs are weakened or killed so they cannot cause serious illness. Even if babies receive several vaccinations in one day, vaccines contain only a tiny fraction of the antigens they encounter every day in their environment. Vaccines give your child the antibodies they need to fight off serious vaccine-preventable diseases.</p>
		<p><b>GPT</b></p> <p>No, vaccines do not overload your baby's immune system. Even newborns have the ability to respond to a large number of vaccines concurrently. Your baby's immune system is capable of recognizing millions of different viruses and bacteria every day. The tiny amount of dead or inactivated bacteria or virus, or pieces of them in each vaccine is just a tiny fraction of what their system manages regularly. So, the vaccines are not a burden for your child's immune system. In fact, they help to build it stronger.</p>
Spanish	¿Pueden las vacunas sobrecargar el sistema inmunitario de mi bebé?	<p><b>CDC</b></p> <p>Las vacunas no sobrecargan el sistema inmunitario. Cada día, el sistema inmunitario de los bebés sanos combate con éxito miles de microbios. Los antígenos son partes de los microbios que activan al sistema inmunitario para que cree anticuerpos, los cuales luchan contra las enfermedades. Los antígenos en las vacunas provienen de los microbios mismos; sin embargo, están debilitados o muertos, por lo cual no pueden hacer que se enfermen gravemente. Incluso si los bebés reciben varias vacunas en un día, las vacunas contienen solo una muy pequeña fracción de la cantidad de antígenos que se encuentran en su ambiente diario. Las vacunas le proveen a su hijo los anticuerpos que necesita para combatir enfermedades graves, prevenibles a través de ellas.</p>
		<p><b>GPT</b></p> <p>No, las vacunas no sobrecargan el sistema inmunitario de tu bebé. El sistema inmunológico de los bebés es capaz de manejar una gran cantidad de antígenos (sustancias que provocan una respuesta inmunitaria). Desde el nacimiento, los bebés entran en contacto con numerosos antígenos todos los días a través de la comida, la suciedad, y las infecciones, y su sistema inmunológico es capaz de manejarlo. De hecho, las vacunas representan solo una pequeña fracción de los antígenos a los que un bebé se expone. Las vacunas ayudan a "entrenar" el sistema inmunológico de tu bebé para reconocer y combatir virus y bacterias específicos, lo que les proporciona una protección muy necesaria contra enfermedades graves.</p>