

# **Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study**

Michael S. Deiner, Vlad Honcharov, Jiawei Li, Tim K. Mackey, Travis C. Porco, Urmimala Sarkar

Submitted to: JMIR Infodemiology  
on: April 24, 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..... 32**  
    Figures ..... 33  
        Figure 1..... 34  
Multimedia Appendixes ..... 35  
    Multimedia Appendix 1..... 36

# Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study

Michael S. Deiner<sup>1\*</sup> PhD; Vlad Honcharov<sup>2\*</sup> MPH; Jiawei Li<sup>3</sup> MS; Tim K. Mackey<sup>4</sup> MAS, PhD; Travis C. Porco<sup>5</sup> MPH, PhD; Urmimala Sarkar<sup>2</sup> MPH, MD

<sup>1</sup>Department of Ophthalmology and Francis I Proctor Foundation, University of California San Francisco San Francisco US

<sup>2</sup>Division of General Internal Medicine at Zuckerberg San Francisco General Hospital and Trauma Center, and Center for Vulnerable Populations University of California San Francisco San Francisco US

<sup>3</sup>S-3 Research LLC San Deigo US

<sup>4</sup>S-3 Research LLC, Global Health Policy and Data Institute, and University of California San Diego Global Health Program Department of Anthropology San Diego US

<sup>5</sup>Departments of Ophthalmology, Epidemiology and Biostatistics, Global Health Sciences, and Francis I Proctor Foundation University of California San Francisco San Francisco US

\*these authors contributed equally

## Corresponding Author:

Travis C. Porco MPH, PhD

Departments of Ophthalmology, Epidemiology and Biostatistics, Global Health Sciences, and Francis I Proctor Foundation

University of California San Francisco

490 Illinois St, 2nd Floor

San Francisco

US

## Abstract

**Background:** Manually analyzing public health-related content from social media provides valuable insights into beliefs, attitudes, and behaviors of individuals, shedding light on trends and patterns that can inform public understanding, policy decisions, targeted interventions, and communication strategies. Unfortunately, the time and effort needed from well-trained human subject matter experts makes extensive manual social media listening unfeasible. Generative large language models (LLMs) have potential for not only summarizing but also interpreting large amounts of text. It is not clear to what extent LLMs can analyze a large set of social media posts at once to glean subtleties of health-related meaning and reasonably report on health-related themes.

**Objective:** Assess feasibility of using LLMs for topic model selection or inductive thematic analysis of large contents of social media posts: Can LLMs conduct topic model selection and inductive thematic analysis as effectively as humans did in a prior manual study, or at least reasonably, as judged by subject matter experts?

**Methods:** We asked the same research question and used the same set of social media content for both the LLM selection of relevant topics and the LLM analysis of themes, as was conducted manually in a prior published study about vaccine rhetoric. We compared the results from the prior manual human analyses to results from analyses by the LLMs GPT4-32K, Claude-instant-100K, and Claude-2-100K. We also assessed if multiple LLMs had equivalent ability and assessed consistency of repeated analysis from each single LLM.

**Results:** Overall, all three LLMs could assess the large corpus of social media posts and summarize content. LLMs generally gave high rankings to the topics chosen previously by humans as most relevant. We reject a null hypothesis ( $P < 0.001$ , overall comparison) and conclude that these LLMs are more likely to include the human-rated top 5 content areas in their own top rankings than would occur by chance. Regarding theme identification, LLMs identified several themes similar to those identified by humans, with very low hallucination rates. Variability occurred between LLMs and between test runs of an individual LLM. Despite not consistently matching the human-generated themes, subject matter experts found themes generated by the LLMs were still reasonable and relevant.

**Conclusions:** LLMs can effectively and efficiently process large social media health-related datasets. LLMs can extract themes from such data that human subject matter experts deem reasonable. However, we were unable to show that the LLMs we tested can replicate the depth of analysis from human subject matter experts by consistently extracting the same themes from the same

data. There is vast potential, once better validated, for automated LLM-based real-time social listening for common and rare health conditions, informing public health understanding of the public's interests and concerns and determining the public's ideas to address them. Clinical Trial: not applicable

(JMIR Preprints 24/04/2024:59641)

DOI: <https://doi.org/10.2196/preprints.59641>

## 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 all users.

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>

## Original Manuscript

## Title

# Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study

## Abstract

**Background:** Manually analyzing public health-related content from social media provides valuable insights into beliefs, attitudes, and behaviors of individuals, shedding light on trends and patterns that can inform public understanding, policy decisions, targeted interventions, and communication strategies. Unfortunately, the time and effort needed from well-trained human subject matter experts makes extensive manual social media listening unfeasible. Generative large language models (LLMs) have potential for not only summarizing but also interpreting large amounts of text. It is not clear to what extent LLMs can analyze a large set of social media posts at once to glean subtleties of health-related meaning and reasonably report on health-related themes.

**Objective:** Assess feasibility of using LLMs for topic model selection or inductive thematic analysis of large contents of social media posts: Can LLMs conduct topic model selection and inductive thematic analysis as effectively as humans did in a prior manual study, or at least reasonably, as judged by subject matter experts?

**Methods:** We asked the same research question and used the same set of social media content for both the LLM selection of relevant topics and the LLM analysis of themes, as was conducted manually in a prior published study about vaccine rhetoric. We used the results from that study as background for our current LLM experiment by comparing the results from the prior manual human analyses to results from analyses by the LLMs GPT4-32K, Claude-instant-100K, and Claude-2-100K. We also assessed if multiple LLMs had equivalent ability and assessed consistency of repeated analysis from each single LLM.

**Results:** LLMs generally gave high rankings to the topics chosen previously by humans as most relevant. We reject a null hypothesis ( $P < 0.001$ , overall comparison) and conclude that these LLMs are more likely to include the human-rated top 5 content areas in their own top rankings than would occur by chance. Regarding theme identification, LLMs identified several themes similar to those identified by humans, with very low hallucination rates. Variability occurred between LLMs and between test runs of an individual LLM. Despite not consistently matching the human-generated themes, subject matter experts found themes generated by the LLMs were still reasonable and relevant.

**Conclusions:** LLMs can effectively and efficiently process large social media health-related datasets. LLMs can extract themes from such data that human subject matter experts deem reasonable. However, we were unable to show that the LLMs we tested can replicate the depth of analysis from human subject matter experts by consistently extracting the same themes from the same data. There is vast potential, once better validated, for automated LLM-based real-time social listening for common and rare health conditions, informing public health understanding of the public's interests and concerns and determining the public's ideas to address them.

**Trial Registration:** Not applicable

**Keywords:** generative large language model; GPT; Claude; Twitter; X formerly known as Twitter; social media; inductive content analysis; COVID-19; vaccine hesitancy; infodemiology



## Introduction

Social media platforms can shed light on public health trends and patterns to inform targeted interventions and communication strategies [1]. The potential to leverage social media to better understand public sentiment about vaccines, which play a crucial role in preventing the spread of infectious diseases, saving lives, and ultimately promoting public health and well-being within society has been well researched [2], [3], [4], [5], [6], [7]. However, assessing unstructured user-generated content on social media can be time consuming, [8] limiting the ability to harness the full potential of this approach to understand and improve public health. In the past, this has led researchers to use foundational methods such as natural language processing (NLP), supervised machine learning, and other approaches to help interpret data [7], [9], [10]. For example topic modeling or classification of posts can be used as an initial step before subsequent manual analyses, but even those methods can be inaccurate and time consuming – perhaps even more so for analysis of a larger corpus of text [8], [11], [12], [13], [14], [15], [16]. Recently it has come to light that “few-shot” or “zero-shot” learners, such as generative large language models (LLMs), may have advantages for overcoming some of these limitations, including for extracting inference or reasoning from large corpora of text including health related content, but with potential inherent bias and other concerns [17], [18], [19], [20], [21], [22]. LLMs such as GPT4 (OpenAI, San Francisco), based on transformer architecture, are neural networks trained on very large corpora of natural text [23], [24].

While manual inductive thematic analyses [9] and other similar manual approaches used in the literature are valuable for assessing unstructured and unlabeled social media content and depict themes of public interest, they demand an extensive burden of human time and effort for detailed content analysis by well-trained human subject matter experts. This makes it unfeasible to conduct large-scale, nearer real-time studies of social media listening to routinely inform public understanding and policy decisions despite the time-sensitive nature and impact on public health of online discourses that constantly evolve during health emergencies [25], [26]. Although it has been suggested LLMs have potential for not only summarizing but also interpreting large amounts of text [18], [19], [20], it is not clear to what extent large language models can analyze text to glean subtleties of health-related meaning and convey the resulting themes in a clear and detailed fashion. In the past, early LLMs had restricted context volume windows, making it difficult to conduct such analyses of large documents or corpora [27]. However, several newer LLMs have become available with an increased context window to allow analysis of larger documents and corpora, initially including GPT4 and Claude 2 (Anthropic, San Francisco) [28], [29].

Application of LLMs to public health social listening approaches may have the potential to help expedite the processes of social media thematic analyses and make it more efficient than tasking human subject matter experts [30], [31], [32]. However, different LLMs can exhibit different biases or capabilities [33], [34], [35], including hallucinations (false information resulting from the token-prediction algorithm), that have not been appropriately evaluated [36]. Specifically, there have not been abundant studies validating the use of LLMs for thematic analysis of large corpuses of health-related social media content. It is foreseeable that public health and healthcare stakeholders are or will begin to more rapidly adopt LLMs to generate automated reports using large unstructured social media or similar health datasets [16], [17], [30]. Before assuming LLMs can achieve the equivalent of humans in the context of thematic identification or in-depth content coding, it is important to compare results from LLMs to those from human qualitative analysis on specific topics of public health importance, such as vaccine



rhetoric [37], [38]. The development of topic models is often misaligned with the needs of users who analyze social media data [39]. Evidence suggests that researchers frequently use topic models sub-optimally due to a lack of adequate methodological support for building and interpreting topics [39]. This gap in support leaves researchers struggling to fully leverage topic models in their analyses [39].

In this comparative study, we evaluate the feasibility of using LLMs for topic model selection or inductive thematic analysis of health-related social media posts on vaccine rhetoric discourse [34], [35]. We compare the output of three different LLMs to conduct the same analysis that members of our group had previously conducted in which they had used a combination of data mining, topic modeling, and manual content analysis in a prior published study examining vaccine rhetoric on Twitter. Here, we used the same corpus of social media content and guidelines for the LLM analysis as was conducted during human annotation in the prior study and then we conducted a comparative analysis [9]. Using the results from that prior study as background for our current LLM experiment, here we asked: Can LLMs conduct topic model selection and inductive analysis in a manner comparable to human performance, or at least reasonably as judged by subject matter experts? We also asked if all selected LLMs are equivalent in their ability, as well as how reliable is one LLM to conduct repeated analyses. We **hypothesized** that LLMs would select the same set of topics as had previously been chosen by humans following topic modeling output by an unsupervised NLP model [9], that LLMs would induce a similar set of themes as humans had [9], that there would be variability in the ability of different LLMs, and that an LLM should provide similar responses with low variability when prompts are repeated. The overall purpose and goal of this study were to: (a) task an LLM with the exact same set of data and tasks that humans were given (manual annotation of Twitter posts) and determine how similarly or different (or unreasonable) the LLMs results were compared to what humans result were; and (2) leverage the relatively new emerging larger context window LLMs for this purpose (i.e., LLMs that could finally allow us to provide all posts in a single prompt for the LLMs).

## Methods:

### Methods for selecting the top 5 most relevant topics that resulted from an unsupervised NLP model

#### *Brief review of methods from original published study: human selection of top 5 most relevant topics from an unsupervised NLP model [9]*

For comparison to this current study's approach using LLMs, we here first describe how the top 5 relevant topics were manually selected in our prior published analysis using the unsupervised topic model BTM [9]. In that prior study, we collected data from Twitter's (now "X") public streaming API from March 2020 to October 2020 (a critical time for the formation of both pro- and anti-vaccination opinions, as the topic of vaccine development was extensively debated and discussed during that period) and filtered it for COVID-19 specific key words ("coronavirus," "covid," and "pandemic," etc.). Of the resulting 3,999,726 Twitter posts, we then removing duplicate tweets (with the same Tweet ID) resulting in 118,971 messages. We next applied a second text filter to isolate anti-vaxx specific messages. We then used the biterm topic model (BTM) to organize our data into 20 different clusters based on hyperparameters set by the research team for the topic model as reported elsewhere, following which we manually screened the top 10 tweets that were most highly correlated to the 20 topic clusters [9]. Finally

using this set of top 10 tweets from 20 clusters, we identified the 5 BTM topics most relevant to our research question by manually identifying the 5 clusters that most closely included messages calling out or making claims about public figures that opposed vaccination or that called out groups of people such as scientists or political parties. We chose to focus our analysis on public figures as they are highly influential in our society, especially on social media. We aimed to assess how their online presence and discourse affect public attitudes and sentiments towards health recommendations and policies. By focusing on public figures, we sought to understand the role they play in shaping public opinion and the potential impact of their statements on the dissemination of anti-vaccine messages. The topics included: *[Tennis Pro]'s anti-vaccination stance*; *[Public Figure 1] and [Philanthropist]' relation to anti-vaccination beliefs*; *[Politician 1]' potential anti-vaccination stance*; *[Politician 2] and Amy Duncan (of note: Amy Duncan is a fictional character played by actress [Actress 1])*; and *Political party potential anti-vaccination views*. Although each topic comprised thousands of tweets, our analysis focused solely on comparing the top 10 most relevant tweets from each cluster, enabling us to efficiently identify the five clusters and corresponding themes most pertinent to our research question [9]. Of note, the set of top 10 most correlated tweets from 20 clusters is the same set of posts that we then used in the current study for LLM-based top 5 most relevant BTM topic selection, described in the next section below.

### *Methods for LLM-based top 5 most relevant BTM topic selection.*

For this current study, we sought to replicate the manual BTM topic selection process above of identifying the 5 clusters (that most closely included messages calling out or making claims about public figures as anti-vaxxers or that called out groups of people such as scientists or political parties). But here we used LLMs for this process in place of the previous more manual approach. To do this, we first prepared the same set of posts (the original set of the top 10 tweets that were most highly correlated to the 20 topic clusters from the original manual study) for use with LLMs by labeling each post with an original BTM topic group ID of 1-20. This was to allow the LLM to know which BTM topic group each post was part of. We then asked the LLMs to rank the BTM groups from 1-20 in the order of relevance as related to the guidelines for what was used by subject matter experts when they had manually selected the most relevant 5 topics in the prior publication. We then compared how well the top 5 (out of 20) topics ranked by LLMs compared to the 5 (out of 20) topics previously chosen manually. More details are as follows:

**LLMs and platforms:** We used GPT4-32K, Claude-instant-100K, and Claude-2-100K, accessing them via the poe.com (Mountain View) platform. To use the poe.com interface we manually pasted in prompt texts and copied out the results – we refer to each of these events as “test runs” in the manuscript.

**Data:** The original corpus of posts contained 193 posts, labeled with one of 20 original BTM Topic numbers. This list of Topic number + post content was included in the prompt shown below.

**The content ranking prompt we used** for all 3 LLMs varied slightly between LLMs but was basically as shown in Figure 1a (this example was used for GPT4).

### *Methods for Statistical Assessment of the top 5 (of 20) topics ranked by LLMs*

We tested the null hypothesis that the LLM's top 5 BTM topic rankings in this current study

would be independent of the top 5 BTM topics chosen in the previous study by the human raters. We modeled LLM choice under this null hypothesis as random sampling without replacement (i.e., the number of topics chosen by the model that had been chosen by the human was assumed given by the hypergeometric distribution, under the null hypothesis). We chose this approach since, under the assumption that the LLM picked choices randomly and independent of the human choices, the number of agreements with the human choices is given by the hypergeometric distribution. If the LLM agreed with the human more often than the hypergeometric would lead us to expect, we conclude the LLM is more likely to pick the human choices than chance alone would indicate. For each LLM, we first determined the number  $N$  of the human-chosen top 5 BTM topics that the LLM ranked, as its top 5 topics (necessarily,  $N$  is in the range 0-5). We then computed the probability using the hypergeometric distribution. Using the number of matches as a test statistic, the probability that the LLM would have picked as many or more of the human choices as we observed therefore provides a probability value as a way to assess that the LLM choices were unrelated to the human choices. In this way, the more of the original 5 BTM topics chosen by humans that were also ranked in the top 5 by LLMs, the lower the probability that the result was by chance alone. A small probability value therefore indicates that LLMs made similar topic choices as humans.

## Methods for comparison of inductive thematic analysis by humans vs. by LLMs

### *Methods of Original study, human inductive thematic analysis to identify and define 5 themes [9]*

For comparison to this current study's approach using LLMs to automatically identify key themes based on the content of the posts, we here first briefly describe how the top 5 themes were manually identified, selected and defined in our initial published manual analysis [9]. In the previous paper, the human team used grounded theory, allowing for themes to emerge while coding rather than prespecifying the content of interest. After the first round of manual review, we inductively developed a codebook for the qualitative content analysis and categorization of Twitter posts. We then reapplied our codebook to the 768 Twitter messages in our sample, while iteratively continuing to develop existing codes and definitions as well as new codes. Ultimately from about 7 themes that we considered met our criteria (to identify the top clinical, social, epidemiological or otherwise significant themes), we selected 5 of them to narrow the focus and describe in the manuscript: **Theme A.** *Neutral-Absence of expression of a clear judgement even if the message is related to the topic*; **Theme B.** *Insults a person because they are an anti-vaxxer; says something derogatory to someone because they are or have been accused of being an anti-vaxxer*; **Theme C.** *Negative public health impact-States or implies that anti-vaxxers and anti-vaccine behaviors have a negative impact on public health*; **Theme D.** *Anti-vax accusation-Accuses or asserts a specific person or groups of people are anti-vaxxers*; **Theme E.** *Defending anti-vax stance- Defends or upholds an anti-vax position*. These findings were important as they suggested a reciprocal influence between public health recommendations and attitudes toward public figures, challenging the previously-described notion of a one-way, outsized influence of celebrities on vaccination attitudes. This nuanced understanding of vaccine sentiment and its interplay with public figures challenged conventional narratives about the influence of celebrities on vaccination attitudes. It also highlighted the complex relationship between public health recommendations, societal perceptions of authority figures, and individual beliefs, underscoring the need for tailored interventions and messaging strategies. This social listening study provided insights into the dynamics of vaccine discourse on Twitter and informing potential strategies for public health officials and policymakers to

craft more effective communication strategies to promote vaccine acceptance and uptake. But this manual human inductive thematic analysis process in our prior study took many days and hours of effort for all team members. Therefore, in the current study we sought to leverage LLMs to try to conduct the same analysis and assessed the outcome (described below). Of note, the exact same set of 768 Twitter messages in the published human assessment study is the same set of posts that we then used in the current study for LLM inductive analysis to identify and define 5 themes with titles, definitions and representative posts, described in the next section below.

### *Methods for LLM inductive analysis to identify and define 5 themes with titles, definitions and representative posts.*

In this current study, we have used the above-described original set of 768 Twitter messages from the published inductive thematic analysis manual study, to include them in our LLM content analysis prompt shown below to prompt the LLM to deduce themes (a very similar prompt was used for all 3 LLMs). Using the prompt submission below, results were obtained twice (i.e. test runs 1 and 2) each for GPT4-32K, Claude-instant-100K, and Claude-2-100K. Each test run was independent of any other run. Completing these tasks took approximately 45 minutes of one researcher's effort. **The content analysis prompt we used** for all 3 LLMs varied slightly between LLMs but was basically as shown in Figure 1b (this example was used for GPT4).

### *Methods to assess hallucination (generation of phantom posts) in responses given by LLMs*

Before assessing the themes identified by the LLMs, we first reviewed the "example posts" provided in the LLM responses to assess how many of the social media post examples provided in the LLM responses were actually part of the original 768 posts we had provided compared to how many post examples provided by the LLMs were "phantom posts" fabricated via hallucination by the LLMs and not in the original set of 768 posts we had provided the LLM. We assessed the accuracy of these example posts to determine if the LLMs generated phantom posts, ensuring that they identified themes accurately without altering the original post content. For each example post provided in each LLM response, we assessed its similarity to the original post from the prompt that had been presented to the LLM and classified it as an: identical example post (a verbatim copy of a post from the original LLM prompt), near identical example post (very similar to an original post in the LLM prompt, but not completely identical such as a missing period or added number) or a phantom example post (the LLM provided us an example of an original post that was not obviously similar to any original post in the LLM prompt). We then summarized the results for each of these three categories overall and tabulated the totals by LLM platform and test run.

### *Methods to assess themes in the responses given by LLMs*

To assess the themes identified by the LLMs, two subject matter experts, who were authors and manual annotators of the original manuscript, reviewed the themes identified by each LLM test run output response. For each response we identified (a) how many of the 5 themes provided matched the original manuscript themes, (b) which themes matched and, (c) how reasonable on a scale of 1-3 was each derived theme, despite whether it matched the original manuscript theme. We evaluated the match between the original themes and the LLM derived themes by assessing for relevance, accuracy, and fidelity to the original themes. We then reported the individual and total matches per LLM test run. For assessing reasonableness, each LLM derived theme was scored by our subject matter expert team members from 1 to 3 where "1" meant not



reasonable (the LLM theme matched poorly with little relevance to the original manuscript), “2” meant reasonable (the LLM theme matched moderately well, with some relevance to the original manuscript), and “3” meant very reasonable (the theme matched closely and accurately with high relevance to the original manuscript). We then averaged the scores of the two authors to assign a reasonableness score.

## **Methods for comparison of inductive thematic analysis by humans vs. pseudo-inductive thematic analysis outputs from the topic models LDA and BERTopic**

Finally, for comparison to the ability and utility of using LLMs for assessing themes, we also compared that to using 2 topic models to produce outputs from the 768 posts and assessing those outputs as we did for LLMs (did they match the original human themes and how reasonable are they), as well as an additional assessment to compare how clear the meaning of the outputs of the topic models were for humans compared to how clear in general the meaning of the outputs of the LLMs were for humans.

### *Methods for topic modeling and for producing word frequency topic grids for human assessment*

Using the same set of original 768 posts we had provided, we used 2 topic modeling frameworks, latent dirichlet allocation (LDA) and Bidirectional Encoder Representations from Transformers with class-based TF-IDF (BERTopic) to develop topics which we then visualized for humans to assess.

LDA is a topic modeling technique to extract topics from a given set of texts, it converts each text into a bag of words and categorize them into k different clusters based on their similarity. It then outputs the correlation score between each text and the cluster, with the correlation score between each word and cluster, the text or word that has higher correlation score are more likely to contain topics related to the rest of the corpus. In our assessment, first we extracted 5 topics from the dataset and for each we output the top 20 keywords that have the highest correlation score to represent the topic for each cluster, and then output a list of texts sorted based on the correlation score from each topic for human assess.

BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions[40]. Instead of converting each text into a bag of words, it uses pre-trained deep learning model and embedding each text into a text vector, and then categorizes them using clustering model. In our assessment, we used KMeans as our clustering model to avoid the output of outliers and set the minimum size of each cluster to 10 and then extracted 5 topics. For each topic we then output the top 20 keywords that have the highest correlation score from each cluster to represent the topic. However, since KMeans directly assigns each text into different topics, we are unable to get the text correlation score for BERTopic.

For visualization of the results from both LDA and BERTopic we created topic grids. To create the topic grid for each topic, the top 20 keywords were included in cells within the grid. The background colors ranged from dark to light and font size ranged from largest to smallest, reflecting the correlation between each keyword to the topic, from the most to least relevant, respectively.

### *Methods to develop and assess the pseudo-themes in the outputs provided by topic model outputs*

For LDA and BERTopic, for each topic a pdf file was produced with the topic grid and a list of word frequencies. These 10 topic model output pdf pages (1 per topic, 5 topics per topic model) were provided to one of the team members that had also assessed the output of the LLMs vs. the original human themes. Unlike the LLM outputs, the topic models provide no theme title, no description of a theme, and no reasons for why any example posts support that theme – making it difficult to compare the topic model results to the themes of the original manuscript. Therefore, to assess the utility of using these topic model outputs for thematic induction, our team member subject matter experts reviewed the 5 topic outputs from LDA and BERTopic and then manually developed a pseudo-theme for each topic. For each, they reviewed the topic grid and list of word frequencies and then they manually developed a title for each topic, which we will refer to in this manuscript as the topic's "pseudo-theme". Assessment of pseudo-theme outputs was then conducted similar to assessment of LLM outputs: For each topic model output (LDA and BERTopic), we identified (a) how many of the 5 topic pseudo-themes matched the original manuscript themes, (b) which original manual theme they matched and, (c) how reasonable on a scale of 1-3 was each derived topic pseudo-theme, despite whether it matched the original manuscript themes. We then reported the individual and total matches per topic model output.

As the topic model output grid was much less of a clear comprehensive narrative than the output of the LLMs, for scoring matches to original themes, for assessing reasonableness, each topic pseudo-theme was assigned a "reasonableness score" by our subject matter expert team members. Scores ranged from 1 to 3 where "1" meant not reasonable (the topic pseudo-theme matched poorly with little relevance to the original manuscript), "2" meant reasonable (the topic pseudo-theme matched moderately well, with some relevance to the original manuscript), and "3" meant very reasonable (the topic pseudo-theme matched closely and accurately with high relevance to the theme of the original manuscript). This scoring process was repeated, while masked to the first scoring round results. We then averaged the scores of the two rounds to assign a reasonableness score to each topic model's pseudo-theme.

Finally, as we noted potentially more difficulty for humans to interpret the topic model grid outputs compared to interpreting LLM outputs, we assigned a "clearness score" comparing our ability to understand the topic model output pseudo-theme to how easily and quickly we had been able to confidently understand the meaning and theme of the typical LLMs theme outputs. We assigned a clearness score from 1-3 as: "1" (this topic model result is much harder for me to easily and quickly confidently understand the meaning and theme of compared to the LLM outputs), "2" (this topic model result is about the same for me to easily and quickly confidently understand the meaning and theme of compared to the LLM outputs), "3" (this topic model result is much easier for me to easily and quickly confidently understand the meaning and theme of compared to LLM outputs). This clearness scoring process was repeated, while masked to the first scoring round results. We then averaged the scores of the two rounds to assign a clearness score for each topic model's pseudo-theme.

### **Ethical Considerations**

As this study used deidentified, publicly available social media data, the Institutional Review Board of University of California, San Francisco classified our proposal as exempt from review (IRB 13-12815). Public figure names have been de-identified by replaced them with

generalized names [in square brackets].

## Results

### Results for selecting the top 5 most relevant topics that resulted from an unsupervised NLP model

#### *Brief review of results from original published study: human selection of top 5 most relevant topics from an unsupervised NLP model*

The 5 topics chosen by humans as the 5 most relevant topics in the published original manual analysis [9] are described in the methods and shown here in Table 1 in the headers for columns 2-6.

#### *Results of LLM-based top 5 most relevant BTM topic selection*

We obtained results from each LLM using three test runs per LLM. Completing these tasks took approximately 1 hour of researcher effort. Each LLM was able to assign rank orders to the 20 topics. In Table 1 for each LLM analysis (each LLM row), columns 2-6 show the relevance-ranking (1 being most relevant, 20 being least relevant) assigned by the LLM for each of the 5 original topics chosen by humans in the original manuscript. Overall, the results suggest LLMs make many of the same 5 topic choices as humans did - GPT was the most successful, followed by Claude 1. A more detailed assessment follows.

#### *Statistical Assessment of the top 5 (of 20) topics ranked by LLMs*

Table 1 shows the names of the top 5 most relevant topics that were chosen manually in the original manuscript [9] and the rankings (out of 20) assigned by the LLM for each of those 5 original topics. When comparing how many of the LLM's top 5 ranked topics were the same 5 topics from the manuscript, different LLMs yielded different results. GPT-4's top 5 ranked topics in test runs included 3 or 4 of the 5 topics from the original manuscript (with a mean of 3.3 over three runs). Claude 1's top 5 ranked topics in test runs included 2 or 3 of the 5 from the manuscript (with a mean of 2.7), while Claude2's top 5 ranked topics in test runs included between 2 or 3 of the 5 from the manuscript (with a mean of 2.3). When broadening to include the LLM's top 10-ranked (rather than just the top 5-ranked) topics for comparison to the 5 topics from the manuscript, the three GPT's top 10 ranked topics in test runs included 4 of the 5 topics from the manuscript (with a mean of 4), the Claude 1's top 10 ranked topics in test runs included 3 or 4 of the 5 from the manuscript (with a mean of 3.3), and finally Claude 2's top 10 ranked topics in test runs included 2 or 3 of the 5 from the manuscript (with a mean of 3.0).

The overall result for each of the 9 LLM test runs conducted (each row) is in Table 1, Column 7. Values shown represent the probability that by chance alone we would have seen as many or more matches of the LLM-chosen top 5 with the 5 chosen by humans, compared to what we actually observed in that test run. A low probability is evidence that the LLMs choices agree with those chosen by humans more than expected by chance alone. Combining all 9 test run results, testing the hypothesis that all LLMs were independent of each other and the humans, we reject this null hypothesis ( $P < 0.001$ , overall comparison, Monte Carlo  $N = 16384$ ), and conclude that the LLMs are more likely to include the top human 5 groups in their own top rankings than would occur by chance.

**Table 1. Relevance-ranking (out of 20 BTM topic groups) by LLMs.** Rows: Each row contains the results of an LLM test run to assess the corpus of 193 posts, where each was labeled in the original manuscript BTM methods as being from one of the 20 BTM topics. For each row, the ranking (1 most relevant, 20 least relevant) assigned by the LLM to the 5 topics shown in the header of each column is shown. Column 1: LLM platform and test run number; Column 2-6: Top 5 most relevant topics (out of 20 BTM topic groups) assigned by human raters in the original manuscript; Column 7: Looking at the top 5 topics selected by the LLM test run in that row, this column shows the probability that by chance alone we would have seen as many or more matches of the LLM-chosen top 5 with the 5 chosen by humans compared to what we actually observed. The hypergeometric probability shows the probability that the LLM would agree with as many or more of the human choices by chance alone.

	BTM topic: [Tennis Pro] antivaxxer stance	BTM topic: [Public Figure 1]/[Philanthropist] relation to anti-vax beliefs	BTM topic: [Politician 1] potential anti- vaxxer stance	BTM topic: Political party potential anti-vax views	BTM topic: Amy Duncan (actress: [Actress 1]), [Politician 2]	Hyper-geometric probability
GPT4, 1st test run	1	2	3	6	19	0.0726
GPT4, 2nd test run	1	2	3	10	16	0.0726
GPT4, 3rd test run	1	2	3	5	19	0.0049
Claude 1, 1st test run	1	6	5	4	12	0.0726
Claude 1, 2nd test run	1	5	4	11	18	0.0726
Claude 1, 3rd test run	1	19	2	10	16	0.3661
Claude 2, 1st test run	1	8	12	4	17	0.3661
Claude 2, 2nd test run	1	5	4	11	18	0.0726
Claude 2, 3rd test run	1	19	2	10	16	0.3661

## Results for comparison of inductive thematic analysis by humans vs. LLMs

### *Brief results of original published study, human inductive thematic analysis to identify and define 5 themes [9]*

The 5 themes identified and defined by humans in the published original manual analysis [9] are described in the methods section above.

### *Overall results of LLM inductive analysis to identify and define 5 themes with titles, definitions and representative posts.*

In general, all LLMs were able to provide responses outputted in the requested format. An



example of output for GPT-4 32k is shown below in Table 2 and additional examples for Claude are shown in the Multimedia Appendix 1, Tables 6 and 7.



**Table 2: Example of GPT-4 32K inductive thematic analysis.** The prompt asked the LLM to read 768 posts and define themes, including with a title, description and post examples. GPT4 test run 1 output is shown.

<b>Theme 1</b>	(a) Title: <u>"Public Perception and Discussion of Celebrities as Anti-Vaxxers"</u> (b) Definition: This theme captures discussions around public figures, notably celebrities such as athletes and actors, being perceived as anti-vaxxers. The public's reaction to these figures espousing anti-vaccination views can influence societal attitudes towards vaccination, potentially impacting public health initiatives and vaccination rates. Additionally, this theme showcases the role of social media in rapidly disseminating and amplifying these views. (c) Examples: "Post 6: [Tennis Pro] is an anti-vaxxer cool cool cool." This post highlights a casual reaction to the news about a celebrity being an anti-vaxxer, indicating how such information can be normalized within social networks. "Post 80: Wow. I had no idea that [Tennis Pro] is an anti-vaxxer." This post exemplifies the surprise and shock that can occur upon learning a well-known figure holds anti-vaccination beliefs. This can prompt further discussion and sharing of the information.
<b>Theme 2</b>	(a) Title: <u>"Expressions of Disappointment and Betrayal Towards Anti-Vaxxer Celebrities"</u> (b) Definition: This theme involves posts expressing disappointment, surprise, or a sense of betrayal upon learning that admired public figures are anti-vaxxers. This reaction can reflect the personal connection fans feel to celebrities and the influence those public figures can have on their fans' beliefs and behaviors. (c) Examples: "Post 77: Oh [Tennis Pro]! I was such a fan (üüü #antiantivacc Coronavirus: [Tennis Pro] reveals he's an anti-vaxxer and it may stop his return to tennis   World News   Sky News." This post exemplifies the disappointment fans can feel when they learn a celebrity they admired holds views they disagree with. "Post 97: [Tennis Pro] being an anti-vaxxer...i guess i have to stan [Tennis Pro 2] now." This post shows the potential for such revelations to shift fan allegiances and influence public opinion about celebrities.
<b>Theme 3</b>	(a) Title: <u>"Concerns About the Impact of Celebrity Anti-Vaxxers on Public Health"</u> (b) Definition: This theme involves posts expressing concern about the potential public health impacts of celebrities espousing anti-vaccination views. These posts reflect an understanding of the influence celebrities can have on public opinion and behavior, and the potential consequences for vaccination rates and public health. (c) Examples: "Post 18: In other news [Tennis Pro] is an anti-vaxxer on top of being an opponent of equal prize money for women tennis players. Puts me more firmly in Federer's corner." This post connects the anti-vaccination stance with other controversial views, suggesting a broader concern about the influence of such public figures. "Post 97: [Tennis Pro] being an anti-vaxxer...i guess i have to stan [Tennis Pro 2] now." This post highlights the potential for celebrity anti-vaxxer views to influence public opinion and behavior, including decisions about vaccination.
<b>Theme 4</b>	(a) Title: <u>"Expression of Negative Sentiments Towards Anti-Vaxxer Celebrities"</u> (b) Definition: This theme involves posts expressing strong negative sentiments towards celebrities who are anti-vaxxers. These posts often include harsh language and criticism, reflecting strong feelings of opposition towards the anti-vaccination stance. (c) Examples: "Post 17: He's an anti-vaxxer? Fuck [Tennis Pro]." This post reflects strong negative sentiment towards the celebrity, highlighting the emotional intensity that discussions around vaccination can provoke. "Post 101: Idiot. [Tennis Pro] reveals he's an anti-vaxxer and it may stop his return to tennis." This post uses derogatory language to express disapproval of the celebrity's anti-vaccination stance, reflecting strong opposition to their views.
<b>Theme 5</b>	(a) Title: <u>"Potential Consequences of Anti-Vaxxer Beliefs for Celebrities' Careers"</u> (b) Definition: This theme involves posts discussing the potential impact of anti-vaccination beliefs on celebrities' professional careers. This includes speculation about potential loss of opportunities or negative impacts on their public image. (c) Examples: "Post 19: #[Tennis Pro] is #antivax [Tennis Pro]'s opposition to vaccination may stop his return to tennis." This post reflects concerns that anti-vaxxer beliefs could have practical consequences for the celebrity's career, in this case potentially preventing a return to professional tennis. "Post 66: Coronavirus: [Tennis Pro] reveals he's an anti-vaxxer and it may stop his return to tennis. Twat." This post combines negative sentiment towards the celebrity with speculation about potential career impacts, highlighting the perceived seriousness of...

## Results assessing hallucination (generation of phantom posts) in responses given by LLMs

Overall, the LLMs rarely produced phantom examples of original posts when providing post

examples in their responses (2/60) and only by Claude 1. All other example posts provided by LLMs were identical (47/60) or near identical (11/60) to the original posts provided in the prompt. In general, GPT-4 performed the best (19/20 identical, 1/20 near-identical) compared to Claude1 and Claude2. Broken down by LLM platform and test run, the results are shown in Table 3.

**Table 3. Assessment of hallucination (generation of phantom posts) in responses given by LLMs Relevance-ranking (out of 20 BTM topic groups) by LLMs.** Each row shows the results of an LLM test run for which each of the 10 example posts provided by a given LLM response was compared to the original pool of posts presented to the LLM in the prompt and then classified as an: identical example post (a verbatim copy of a post from the original LLM prompt), near identical example post (very similar to an original post in the LLM prompt, but not completely identical such as a missing period or added number) or a phantom example post (the LLM provided us an example of an original post that was not obviously similar to any original post in the LLM prompt). The results for each of these three categories are tabulated by LLM platform and test run.

	Identical example post	Near identical example post	Phantom example post
GPT4, 1st test run	10/10	0/10	0/10
GPT4, 2nd test run	9/10	1/10	0/10
Claude 1, 1st test run	6/10	3/10	1/10
Claude 1, 2nd test run	7/10	2/10	1/10
Claude 2, 1st test run	8/10	2/10	0/10
Claude 2, 2nd test run	7/10	3/10	0/10

### *Results assessing themes in the responses given by LLMs*

We first assessed how many of the themes identified by LLMs were equivalent to the themes from the original manuscript [9]. Overall our team's two subject matter experts found that the inducted themes output by LLMs partially matched the five themes described in the manuscript [9]. Each human inductive analysis derived theme in the original manuscript was matched at least once successfully by an LLM test run, with the exception of the neutral category, which did not yield any corresponding matches. Table 4 shows the results for each LLM test run as compared to the original themes from the human thematic induction paper. In each cell the theme title provided by the LLM output is in quotes, and in each cell above the LLM theme title, we have indicated if the LLM theme matched one of the original paper's Themes A-E [9] or if there was no match with any of the original paper's themes. Claude 1 most closely matched the themes from the original human inductive thematic analysis but did not identify every theme and was closely followed by outputs of both Claude 2 test run 1 and GPT test run 1. The GPT output from test run 2 only identified one theme, and both Claude 1 test run 1 and Claude 2 test run 2 outputs did not successfully identify any of the original themes.

Next, we assessed how reasonable each theme derived from each LLM test run was (independently of whether it matched a theme from the original human study). Our team's two human subject experts determined that the majority of the LLMs themes were reasonable but

varied by LLM. In Table 5, each cell includes the reasonableness score of an LLM's theme, as an average of the 2 scores assigned by the 2 human assessors. As described in more detail in the methods, scores ranging from 1-3 (1=not reasonable), (2=reasonable), (3=very reasonable) were assigned to each theme in each test run. The average of all the scores assigned for a given test run are in the final column, and these ranged from 1.8 to 2.8. Reasonableness for each of the themes included in GPT test run 1 ranged from 2.5 to 3, while GPT test run 2 performed slightly worse. Themes included in Claude 1 test run 1 ranged from scores of 2-3, while the Claude 2 test run 1 performed relatively well, with the exception of a single theme that was determined to be a poor match. Both Claude 1 test run 2 and Claude 2 test run 2 performed relatively poorly, similar to how they had underperformed producing themes that matched the originals. Notably, both the matched themes and reasonableness were inconsistent between the two test runs for each given LLM.

### Results assessing pseudo-themes in the outputs provided by topic models

Regarding matches to the five human induced themes described in the original manuscript, overall our team's two subject matter experts found that the induced themes output by topic models never matched any of the original 5 themes (0/10, see LDA and BERTopic rows in Table 4). Regarding reasonableness, the topic model pseudo-themes had scores comparable to the lower-performing LLMs (see LDA and BERTopic rows in Table 5). Regarding being understandable, the mean clearness score (see LDA and BERTopic rows in Table 5) reflected that it was more difficult to easily and quickly confidently understand the meaning and theme of each topic model output compared to the output of LLMs (1.6 for LDA and 1.7 for BERTopic, where: 1 = much harder than LLMs, 2 = about the same as LLMs, 3 = much easier than LLMs).

**Table 4. Original Human Induced Themes and Matches With LLM Induced Themes or Topic Model Pseudo-themes.** Table 4 shows the results for each LLM test run (or topic model output) as compared to the original themes from the human thematic induction paper. Column 1 indicates the source, for each row, that leads to the themes provided in that row. Row 1 shows the 5 primary human induced themes from 768 posts in the published prior study. Subsequent rows indicate the LLM test run (or topic model output) in this current study used to derive themes from those same 768 posts, and the 5 LLM-induced theme titles from the LLM's output for that test run are shown in columns 2-6, in quotes. In each cell above the LLM theme title, we have indicated if the LLM theme (or topic model pseudo-theme) matched one of the original paper's Themes A-E and if so, we indicated the theme it matched. For LLMs or topic models where there was no match with any of the original paper's themes this is indicated with "(no match)."

Source of the themes	Inducted themes, from original manuscript or from LLM test runs or topic models. For the LLM-rows: (matches to original themes A-E), "LLM theme title"					# Match to human's themes
Human inductive analysis derived themes - definition	Theme A. Neutral-Absence of expression of a clear judgement even if the message is related to the topic	Theme B. <i>Insults a person because they are an anti-vaxxer; says something derogatory to someone because they are or have been accused of being an anti-vaxxer.</i>	Theme C. Negative public health impact- States or implies that anti-vaxxers and anti-vaccine behaviors have a negative impact on public health.	Theme D. Anti-vax accusation- Accuses or asserts a specific person or groups of people are anti-vaxxers.	Theme E. Defending anti-vax stance- Defends or upholds an anti-vax position.	n/a
GPT4, 1st test run	(no match) "Public Perception and Discussion of	Match: Theme B. "Expressions of Disappointment and Betrayal Towards	Match: Theme C. "Concerns About the Impact of Celebrity Anti-	Match: Theme B. "Expression of Negative Sentiments	(no match) "Potential Consequences of Anti-Vaxxer	3 of 5

	<i>Celebrities as Anti- Vaxxers"</i>	<i>Anti-Vaxxer Celebrities"</i>	<i>Vaxxers on Public Health"</i>	<i>Towards Anti-Vaxxer Celebrities"</i>	<i>Beliefs for Celebrities' Careers"</i>	
GPT4, 2nd test run	(no match) "Public Perception of Celebrities and Vaccination Stances"	(no match) "Emotional Responses to Anti-vaccination Views"	(no match) "Public Criticism and Condemnation of Anti-Vaccination Views"	(no match) "Potential Consequences of Anti-Vaccination Views"	Match: Theme B. "Public Shaming and Ridicule of Anti-Vaccination Views"	1 of 5
Claude1, 1st test run	Match: Theme E. "Vaccine skepticism during the COVID-19 pandemic"	Match: Theme D. "Accusations of being "anti-vaxxers" in the political discourse"	Match: Theme B. "Negative reactions to celebrity anti-vaccine stances"	(no match) "Comparisons between anti-lockdown and anti-vaccine movements"	Match: Theme C. "Spread of anti-vaccine messaging during the pandemic"	4 of 5
Claude1, 2nd test run	(no match) "Anti-vaxxer sentiment"	(no match) "Debate over COVID-19 vaccines"	(no match) "COVID-19 vaccine promotion and misinformation"	(no match) "Lockdown and public health protest activity"	(no match) "Popular culture, celebrities, and public discussions"	0 of 5
Claude2, 1st test run	Match: Theme E. "Skepticism towards COVID-19 vaccines"	Match: Theme C. "Blaming deaths and outbreaks on anti-vaccine views"	(no match) "[Tennis Pro]'s COVID-19 diagnosis"	Match: Theme B. "Insults and criticisms of anti-vaccine people"	(no match) "Anti-vaccine views linked to other conspiracies"	3 of 5
Claude2, 2nd test run	(no match) "Anti-vaxxer sentiment"	(no match) "Politicization of vaccines"	(no match) "Vaccine misinformation"	(no match) "Vaccine hesitancy"	(no match) "Calls for vaccination"	0 of 5
LDA	(no match) "Protesting lockdowns"	(no match) "[Tennis Pro] with COVID-19"	(no match) "Distrusting politicians and COVID vaccine"	(no match) "Numerous themes"	(no match) "Anti-vaccination COVID conspiracies"	0 of 5
BER Topic	(no match) "Anti-vaccination and anti-lockdown"	(no match)" Distrust in the COVID-19 vaccine, and politicians"	(no match) "[Tennis Pro] anti-vaxxer stance"	(no match) " Anti-vaccination conspiracies"	(no match) "Anti-lockdown opinions and protests"	0 of 5

**Table 5. Reasonableness of LLM Induced Themes or Topic Model Pseudo-themes.** Test runs and LLM-induced themes (or topic model pseudo-themes) are the same as that shown in Table 4. Each row represents a test run with column 1 indicating the LLM or Topic model source, and columns 2-6 displaying the LLM-induced theme titles (in quotes). A reasonableness score, shown in parentheses in columns 1-6, is the average of scores assigned by the 2 human assessors of the LLM theme (or of the pseudo-theme assessments of the topic model outputs), based on a scale of 0-3 (0= not understandable, 1=not reasonable, 2=reasonable, 3=very reasonable). For topic models rows, a clearness score is also provided (scored as, when compared to the LLM outputs, the ability to easily and quickly confidently understand the meaning and theme of the topic model output was: 1 = much harder than LLMs, 2 = about the same as LLMs, 3 = much easier than LLMs). Column 7 shows the average of the column 2-6 scores for that LLM (or topic model) row, where "R" (all rows) is the mean reasonableness score for that row, and "C" (topic model rows) is the mean clearness score for the topic model outputs compared to that of LLM outputs.

Source	LLM Induced Theme Titles and Subject Matter Expert Assigned Reasonableness Scores (and Clearness Scores for LDA and BERTopic)					Mean score
GPT4, 1st test run	"Public Perception and Discussion of Celebrities as Anti-Vaxxers" (reasonable: 2.5)	"Expressions of Disappointment and Betrayal Towards Anti-Vaxxer Celebrities" (reasonable: 3)	"Concerns About the Impact of Celebrity Anti-Vaxxers on Public Health" (reasonable: 3)	"Expression of Negative Sentiments Towards Anti-Vaxxer Celebrities" (reasonable: 3)	"Potential Consequences of Anti-Vaxxer Beliefs for Celebrities' Careers" (reasonable: 2.5)	R: 2.8
GPT4, 2nd test run	"Public Perception of Celebrities and Vaccination Stances" (reasonable: 2.5)	"Emotional Responses to Anti-vaccination Views" (reasonable: 2.5)	"Public Criticism and Condemnation of Anti-Vaccination Views" (reasonable: 3)	"Potential Consequences of Anti-Vaccination Views" (reasonable: 2)	"Public Shaming and Ridicule of Anti-Vaccination Views" (reasonable: 2.5)	R: 2.5
Claude1, 1st test run	"Vaccine skepticism during the COVID-19 pandemic" (reasonable: 2.5)	"Accusations of being 'anti-vaxxers' in the political discourse" (reasonable: 3)	"Negative reactions to celebrity anti-vaccine stances" (reasonable: 3)	"Comparisons between anti-lockdown and anti-vaccine movements" (reasonable: 2)	"Spread of anti-vaccine messaging during the pandemic" (reasonable: 2.5)	R: 2.6
Claude1, 2nd test run	"Anti-vaxxer sentiment" (reasonable: 2)	"Debate over COVID-19 vaccines" (reasonable: 2.5)	"COVID-19 vaccine promotion and misinformation" (reasonable: 1.5)	"Lockdown and public health protest activity" (reasonable: 1.5)	"Popular culture, celebrities, and public discussions" (reasonable: 2)	R: 1.9
Claude2, 1st test run	"Skepticism towards COVID-19 vaccines" (reasonable: 3)	"Blaming deaths and outbreaks on anti-vaccine views" (reasonable: 3)	"[Tennis Pro]'s COVID-19 diagnosis" (reasonable: 2.5)	"Insults and criticisms of anti-vaccine people" (reasonable: 3)	"Anti-vaccine views linked to other conspiracies" (reasonable: 1.5)	R: 2.6
Claude2, 2nd test run	"Anti-vaxxer sentiment" (reasonable: 2)	"Politicization of vaccines" (reasonable: 1.5)	"Vaccine misinformation" (reasonable: 1.5)	"Vaccine hesitancy" (reasonable: 2)	"Calls for vaccination" (reasonable: 2)	R: 1.8
LDA	"Protesting lockdowns" (reasonable: 2.5; clearness: 1.5)	"[Tennis Pro] with COVID-19" (reasonable: 2.5; clearness: 1.5)	"Distrusting politicians and COVID vaccine" (reasonable: 2; clearness: 1.5)	"Numerous themes" (reasonable: 2; clearness: 2)	"Anti-vaccination COVID conspiracies" (reasonable: 2.5; clearness: 1.5)	R: 2.3 C: 1.6
BERTopic	"Anti-vaccination and anti-lockdown" (reasonable: 1.5; clearness: 1.5)	"Distrust in the COVID-19 vaccine, and politicians" (reasonable: 2; clearness: 1.5)	"[Tennis Pro] anti-vaxxer stance" (reasonable: 2.5; clearness: 3)	"Anti-vaccination conspiracies" (reasonable: 2; clearness: 1)	"Anti-lockdown opinions and protests" (reasonable: 1.5; clearness: 1.5)	R: 1.9 C: 1.7

	clearness: 1.5)	clearness: 2)			clearness: 1)	
--	-----------------	---------------	--	--	---------------	--





## Discussion

### Principal Findings:

Our principal results compared to our original research questions and hypotheses are described here overall and then in further detail below. In this study we asked if LLMs can conduct topic model selection from an analysis of a large corpus of health-related social media posts, equivalent to how humans did. We hypothesized that LLMs would select the same set of five most relevant BTM topics (out of 20) as had previously been chosen by humans. Overall, we have found that all LLMs studied could assess the large corpus of social media posts, provide outputs, and that some of these outputs identified the top 5 most relevant topic models compared to humans quite well. For example, the relevancy of BTM topic: “[*Tennis Pro*] *antivaxxer stance*” was ranked number 1 by all LLM test runs and the relevancy of BTM topic: “[*Politician 1*] *potential anti-vaxxer stance*” was ranked in the top 5 by 8 of 9 LLM test runs. One particular original top 5 theme though, BTM topic: “*Amy Duncan (actress: [Actress 1]), [Politician 2]*” was consistently deemed not relevant by LLMs (most likely due to being about a fictional character, discussed in Limitations section below).

We also asked if LLMs can conduct inductive thematic analysis of a large corpus of health-related social media posts, equivalent to how humans did using the same corpus of 768 posts. We hypothesized that LLMs would induce a similar set of themes as humans had. Overall, we found that LLMs in our study identified several of the original themes identified by humans, with generally very low hallucination rates (almost no phantom posts were created in LLM responses). For example, Claude 2 identified a theme entitled “*Insults and criticisms of anti-vaccine people*” with 0/10 phantom examples of original posts when providing post examples in its response and we determined this LLM theme was a match for the original human inducted theme B: “*Insults a person because they are an anti-vaxxer; says something derogatory to someone because they are or have been accused of being an anti-vaxxer.*” Our findings add to a growing body of literature where LLMs are observed to provide similar (or at least reasonable) results as that provided by human assessors of a corpus of social media text. For example, a recent study of topic model detection from news stories by humans compared to topic model detection by LLM found only minor variations in their respective topic evaluation scores and also found GPT-4 outperformed other LLMs in their study [38], similar to its performance in our analysis. We did observe however that human coding appeared to have uncovered more depth and nuance, including that many posts were not amenable to a clear pro- or anti-vax stance, for example a post such as “*Maybe I'm an anti-vaxxer because no no no no no waaaaaaaaayyyyyyy*”. Future studies might investigate a hybrid approach as suggested by Haupt et al [16], where a small subset of messages is coded by humans to potentially assist the LLM in detecting prominent themes and narratives within large corpora with improved depth and nuance.

In addition, we had hypothesized that even if not identical to the themes determined by humans, the LLMs assessment of the original 768 posts would at least produce reasonable themes, as judged by subject matter experts. Overall, we found that despite not consistently matching the original themes, many of the unmatched themes generated by the LLMs were still



quite reasonable and relevant. For example GPT 4 test run 2 only resulted in one theme that matched the original 5 human determined themes, yet all of the themes it provided were rated by our subject matter experts with reasonableness scores ranging from 2 (reasonable) to 3 (very reasonable), with an overall average score of 2.5.

We also asked if all LLMs are equivalent in their ability and had hypothesized that there would be variation in the ability of different LLMs. Our results demonstrated some variation between LLMs in ranking of the 20 BTM topics and in the themes generated by different LLMs consistent with our hypothesis and with the well-known observation that different LLMs can yield substantially different performance even with the same size-class [28], [29], [41], with some LLMs identifying more of the original themes than others.

Finally, we had hypothesized that any given LLM would provide similar responses with low variability when test run prompts are repeated. But when using the same prompt with the same LLM, we found significant variation between test runs.

Overall, all of our results suggest that the utility of using the LLMs in our study for thematic analyses may be an efficient starting point, but do not currently match the ranking and especially the themes produced by a group of human subject matter experts that undertake in-depth qualitative content coding.

In our usage of topic models as a comparator to using LLMs, we found that use of topic models (rather than LLMs) to attempt to extract a pseudo-theme resulted in less effective matching of the original human themes (0/10). The topic models output's pseudo-themes had scores comparable to the lower-performing LLMs in terms of being reasonable based on the content of the corpus of posts. In addition, the topic model pseudo-themes were more unclear than the outputs from LLMs, and contained much less detail (no theme title, no description of the theme, no description of why posts examples represented the theme), requiring additional subject matter expertise to interpret theme titles and pseudo-themes from the topic model outputs, compared to LLM outputs.

## Relationship to other work

This current study serves as a direct follow-up to our initial unsupervised topic modeling and manual content annotation social listening study of Twitter data, aiming to explore the potential expansion and optimization of this field through the use of LLMs. Previous research [42] has examined the role of social media in medicine and healthcare. This current study contributes evidence of the utility of LLMs in conducting such research, and adds to the literature seeking to validate the use of LLMs, which is an evolving field [17], [30], [38].

## Limitations and discussion of less successful results

Despite fairly reasonable LLM results compared to humans in this study, some results were not consistent with the human-derived topics or themes, particularly for outputs generated by LDA and BERTopic. Exploring those differences can help us to improve performance in future studies or understand the limitations of our approach. For example, although many of the LLM rankings of the 5 most relevant BTM topics compared well to humans, the topic of "Amy Duncan (actress: [Actress 1])" consistently was ranked in the bottom quartile by LLMs (see column 6 Table 1). To understand the cause of this, we noted that in the prior study, to ensure

relevance to public discourse, we had manually selected clusters containing both verified and unverified Twitter accounts of public figures and groups, and therefore this particular BTM topic that we had manually selected had content regarding fictional characters. It is likely the LLMs recognized the content of this BTM topic as being about fictional character Amy Duncan and therefore it may have ranked this topic low as it is not about an actual public figure. This example demonstrates how humans may approach a task assessing a large corpus of social media posts differently than LLMs without very specific guidance to LLMs, and is an example of where iterative validations could help to improve the precision of an LLM prompt.

We also found that, despite decent performance by some of the LLMs, none of the LLMs actually generated themes that completely matched all 5 of the themes from the original analysis (see Table 4, right column) despite the use of very specific prompts that attempted to replicate the methods employed for manual annotation in the original paper by Honcharov et al [9]. We note that for a task such as thematic induction, there is always some level of subjectivity, even in our prior manual study. When we went back and reviewed the original prior study, we note that the initial manual analysis of the dataset had unveiled several supplementary key themes that were not incorporated into the report due to less overall agreement or difference in specific focus of themes. Hence, future studies should investigate whether the LLMs would have identified similar supplementary themes if directed to do so and how this may differ based on different specificity of prompts for topic modeling-related tasks requested of the LLMs. These results might not therefore be entirely surprising as they suggest that just as with humans, LLMs can exhibit subjectivity in interpreting large corpus of content, resulting in variation in results. This concept remains open for exploration in future analyses.

We observed significant variation between repeat runs of an identical prompt with the same content and same LLM, as expected over a web-based interface where it is not possible to set the temperature. In principle, a choice of temperature 0 should make the inference largely (though not perfectly) repeatable [43], but such setting was not possible using a web-based chat interface. Further work is needed to determine whether multiple runs at a larger temperature setting achieves greater flexibility than a single run at zero temperature. This variation also suggests the need for additional validation approaches that should be assessed through human supervision. Perhaps, LLMs could initially assess the outputs from multiple repeated test runs to assign a score of consistency between outputs, indicating areas with significant unexpected test retest variation for follow-up human supervision.

Other limitations to our study include the fact that we only used X (Twitter) content, we focused just on vaccine-related content, and we did not use all available LLMs. All of these limitations can be addressed in future comparative analysis studies to help draw more broad conclusions about the acceptable use of our approach for other content sources, health topics and different LLM platforms. In addition, we note that our analyses could not be fully masked, as original authors from the prior study conducted the assessment of LLM themes in this study.

## Conclusions

Our analysis demonstrates that LLMs can effectively and efficiently process some large social media health-related datasets and extract themes comparable to human researchers. While LLMs may not yet match human accuracy, this evolving field holds promise for greatly lowering the time and cost of analyses.

## Future Studies and Potential Future Significance

Future human – LLM comparative studies on larger datasets and diverse social media corpora are needed to support our current findings and before concluding that LLMs are a valid social listening tool to distill useful, relevant, unbiased and un-hallucinated themes. Researchers in other health science domains should further examine LLMs to assess large corpora of their social media posts and different prompts with varying specificity to topic modeling tasks to accurately choose relevant topics and to describe main themes for other health topics of interest as these LLMs may be more fine-tuned for vaccine or misinformation-related thematic detection. This can be done using results from additional prior manual inductive thematic analysis studies and comparing the original manually derived results to that of LLMs as conducted in this study. Ideally, such studies would be conducted for any particular health science field before assuming results from one field are sufficient for another.

Future studies also may help to further assess the utility of variability even between results of repeat test runs for a given LLM (see limitations above). While variability can be mitigated in LLMs by setting the LLM temperature parameter to 0, the variability may prove useful in deriving an ensemble thematic analysis, for potential increased performance (as is well-known for ensemble models in other fields [43], [44], [45]). Although the human effort and time needed to complete the tasks in our study was a fraction of the human hours of time that the original manual study took (several hours in the current study instead of 40+ hours in the original study), future studies would be needed specifically to measure, compare and substantiate claims of time savings, efficiency and costs savings of using LLMs for health-related social listening.

Once better validated, LLMs could find numerous social listening applications, including for disease forecasting (prediction) and nowcasting (providing data for situational awareness on what the public does, knows, or feels about health issues), data classification of established online health discourse topics and possible detection of new themes/trends, and efficiently grouping intersecting online health behavior queues and information seeking behavior for different health topics [46]. These applications could inform public health understanding of public interests and concerns, and to learn the public's ideas to address them.

Such information could be used to revise and incorporate key current topics into outdated standard reported outcome forms, such as patient reported quality of life assessments or surgical treatment patient reported outcome forms, while also informing public health education and promotion campaigns with themes generated from extant online conversations closer to real-time when users experience and report them.

## Acknowledgements

This study was supported by a National Eye Institute of the National Institutes of Health grant (1R01EY024608-01A1; Principal Investigator [PI]: T.M.L.), a National Eye Institute Core Grant for Vision Research (EY002162; PI: Erik M Ullian), and a Research to Prevent Blindness Unrestricted Grant (PI: Jacque L Duncan).

This paper analyzes the results of generative artificial intelligence (AI)–where we studied the use of the generative AI tools GPT-4 (OpenAI) and Claude for LLM-based topic model selection and thematic assessment of social media posts. Generative AI was not used in ideation, or manuscript writing, or preparation.

## Data Availability

Deidentified data that contains associated twitter IDs associated with the original study, including the original subsets of 193 and 768 posts used in this current study, are available in a GitHub repository [47].

## Conflicts of Interest

T.K.M. and J.L. are employees of the startup company S-3 Research LLC. S-3 Research is a startup funded and currently supported by the National Institutes of Health—National Institute on Drug Abuse through a Small Business Innovation and Research contract for social media research and technology commercialization. T.K.M. is the Editor-in-Chief of JMIR Infodemiology. U.S. holds current research funding from the National Cancer Institute of the National Institutes of Health, California Healthcare Foundation, the Patient-Centered Outcomes Research Institute, and the Agency for Healthcare Research and Quality. She holds contract funding from InquisitHealth and RecoverX. She serves as a scientific/expert advisor for HealthTech 4 Medicaid (volunteer) and is a member of the American Medical Association's Equity and Innovation Advisory Group (honoraria). US is also on the Board of Directors of the Collaborative for Accountability and Improvement (volunteer). She is an advisor for Waymark (shares) and for Ceteri Capital I GP, LLC (shares). She has been a clinical advisor for Omada Health (honoraria), and an advisory board member for Doximity (honoraria, stock). These sponsors had no role designing or conducting this research.

## References:

- [1] G. Eysenbach, "Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet," *J Med Internet Res*, vol. 11, no. 1, p. e11, Mar. 2009, doi: 10.2196/jmir.1157.
- [2] H. Marani, M. Y. Song, M. Jamieson, M. Roerig, and S. Allin, "Public Officials' Engagement on Social Media During the Rollout of the COVID-19 Vaccine: Content Analysis of Tweets," *JMIR Infodemiology*, vol. 3, no. 1, p. e41582, Jul. 2023, doi: 10.2196/41582.
- [3] T. Mackey, C. Baur, and G. Eysenbach, "Advancing Infodemiology in a Digital Intensive Era," *JMIR Infodemiology*, vol. 2, no. 1, p. e37115, Feb. 2022, doi: 10.2196/37115.
- [4] H. C. Lyson *et al.*, "Social Media as a Tool to Promote Health Awareness: Results from an Online Cervical Cancer Prevention Study," *J Cancer Educ*, Jun. 2018, doi: 10.1007/s13187-018-1379-8.
- [5] G. M. Le *et al.*, "Perceptions of cervical cancer prevention on Twitter uncovered by different sampling strategies," *PLoS ONE*, vol. 14, no. 2, p. e0211931, Feb. 2019, doi: 10.1371/journal.pone.0211931.
- [6] Q. Xu, T. J. McMann, J. Li, C. Wenzel, and T. K. Mackey, "Characterization of COVID-19 vaccine clinical trial discussions on the social question-and-answer site Quora," *Trials*, vol. 24, no. 1, p. 790, Dec. 2023, doi: 10.1186/s13063-023-07837-5.
- [7] T. McMann *et al.*, "Detection and Characterization of Web-Based Pediatric COVID-19 Vaccine Discussions and Racial and Ethnic Minority Topics: Retrospective Analysis of Twitter Data," *JMIR Pediatr Parent*, vol. 6, p. e48004, Nov. 2023, doi: 10.2196/48004.
- [8] E. Karafillakis *et al.*, "Methods for Social Media Monitoring Related to Vaccination: Systematic Scoping Review," *JMIR Public Health Surveill*, vol. 7, no. 2, p. e17149, Feb. 2021, doi:

10.2196/17149.

- [9] V. Honcharov *et al.*, “Public Figure Vaccination Rhetoric and Vaccine Hesitancy: Retrospective Twitter Analysis,” *JMIR Infodemiology*, vol. 3, p. e40575, 2023, doi: 10.2196/40575.
- [10] A. J. Calac, M. R. Haupt, Z. Li, and T. Mackey, “Spread of COVID-19 Vaccine Misinformation in the Ninth Inning: Retrospective Observational Infodemic Study,” *JMIR Infodemiology*, vol. 2, no. 1, p. e33587, Mar. 2022, doi: 10.2196/33587.
- [11] M. S. Deiner *et al.*, “Clinical Age-Specific Seasonal Conjunctivitis Patterns and Their Online Detection in Twitter, Blog, Forum, and Comment Social Media Posts,” *Investigative ophthalmology & visual science*, vol. 59, no. 2, pp. 910–920, Feb. 2018, doi: 10.1167/iops.17-22818.
- [12] P. E. Kummervold *et al.*, “Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse,” *JMIR Medical Informatics*, vol. 9, no. 10, p. e29584, Oct. 2021, doi: 10.2196/29584.
- [13] M. S. Deiner *et al.*, “Facebook and Twitter vaccine sentiment in response to measles outbreaks,” *Health informatics journal*, p. 1460458217740723, Nov. 2017, doi: 10.1177/1460458217740723.
- [14] A. R. Ashraf, T. K. Mackey, and A. Fittler, “Search Engines and Generative Artificial Intelligence Integration: Public Health Risks and Recommendations to Safeguard Consumers Online,” *JMIR Public Health Surveill*, vol. 10, p. e53086, Mar. 2024, doi: 10.2196/53086.
- [15] U. Sarkar and D. W. Bates, “Using Artificial Intelligence to Improve Primary Care for Patients and Clinicians,” *JAMA Intern Med*, vol. 184, no. 4, pp. 343–344, Apr. 2024, doi: 10.1001/jamainternmed.2023.7965.
- [16] M. R. Haupt, M. Chiu, J. Chang, Z. Li, R. Cuomo, and T. K. Mackey, “Detecting nuance in conspiracy discourse: Advancing methods in infodemiology and communication science with machine learning and qualitative content coding,” *PLoS One*, vol. 18, no. 12, p. e0295414, Dec. 2023, doi: 10.1371/journal.pone.0295414.
- [17] M. S. Deiner *et al.*, “Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study,” *Journal of Medical Internet Research*, vol. 26, no. 1, p. e49139, Mar. 2024, doi: 10.2196/49139.
- [18] T. B. Brown *et al.*, “Language Models are Few-Shot Learners.” arXiv, Jul. 22, 2020. doi: 10.48550/arXiv.2005.14165.
- [19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners.” arXiv, Jan. 29, 2023. doi: 10.48550/arXiv.2205.11916.
- [20] J. A. Lossio-Ventura *et al.*, “A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data,” *JMIR Mental Health*, vol. 11, no. 1, p. e50150, Jan. 2024, doi: 10.2196/50150.
- [21] A. Rao *et al.*, “Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study,” *J Med Internet Res*, vol. 25, p. e48659, Aug. 2023, doi: 10.2196/48659.
- [22] C. W. Safranek *et al.*, “Automated HEART score determination via ChatGPT: Honing a framework for iterative prompt development,” *J Am Coll Emerg Physicians Open*, vol. 5, no. 2, p. e13133, Apr. 2024, doi: 10.1002/emp2.13133.
- [23] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Aug. 01, 2023. doi: 10.48550/arXiv.1706.03762.
- [24] OpenAI *et al.*, “GPT-4 Technical Report.” arXiv, Mar. 04, 2024. doi: 10.48550/arXiv.2303.08774.
- [25] M. Watkins, J. S. Mallion, D. Frings, J. Wills, S. Sykes, and A. Whittaker, “Public health messages during a global emergency through an online community: a discourse and sentiment analysis,” *Front Digit Health*, vol. 5, p. 1130784, Jun. 2023, doi: 10.3389/fdgth.2023.1130784.



- [26] A. D. Oxman *et al.*, “Health communication in and out of public health emergencies: to persuade or to inform?,” *Health Research Policy and Systems*, vol. 20, no. 1, p. 28, Mar. 2022, doi: 10.1186/s12961-022-00828-z.
- [27] L. Huang, S. Cao, N. Parulian, H. Ji, and L. Wang, “Efficient Attentions for Long Document Summarization.” arXiv, Apr. 11, 2021. doi: 10.48550/arXiv.2104.02112.
- [28] Anthropic, “Model Card and Evaluations for Claude Models.” Accessed: Apr. 05, 2024. [Online]. Available: <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf>
- [29] OpenAI, “GPT-4 and GPT-4 Turbo.” Accessed: Apr. 05, 2024. [Online]. Available: <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo?ref=maginaire.com>
- [30] S. De Paoli, “Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach,” *Social Science Computer Review*, p. 08944393231220483, Dec. 2023, doi: 10.1177/08944393231220483.
- [31] Y. Wang and Y. Chen, “Characterizing discourses about COVID-19 vaccines on Twitter: a topic modeling and sentiment analysis approach,” *J Commun Healthc*, vol. 16, no. 1, pp. 103–112, Mar. 2023, doi: 10.1080/17538068.2022.2054196.
- [32] A. Elyashar, I. Plohotnikov, I.-C. Cohen, R. Puzis, and O. Cohen, “The State of Mind of Health Care Professionals in Light of the COVID-19 Pandemic: Text Analysis Study of Twitter Discourses,” *Journal of Medical Internet Research*, vol. 23, no. 10, p. e30217, Oct. 2021, doi: 10.2196/30217.
- [33] J. Ashwin, A. Chhabra, and V. Rao, “Using Large Language Models for Qualitative Analysis can Introduce Serious Bias,” *Policy Research Working Paper Series*, Art. no. 10597, Nov. 2023, Accessed: Mar. 27, 2024. [Online]. Available: <https://ideas.repec.org/p/wbk/wbrwps/10597.html>
- [34] W. X. Zhao *et al.*, “A Survey of Large Language Models.” arXiv, Nov. 24, 2023. Accessed: Apr. 05, 2024. [Online]. Available: <http://arxiv.org/abs/2303.18223>
- [35] S. Minaee *et al.*, “Large Language Models: A Survey.” arXiv, Feb. 20, 2024. Accessed: Apr. 05, 2024. [Online]. Available: <http://arxiv.org/abs/2402.06196>
- [36] L. Huang *et al.*, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.” arXiv, Nov. 09, 2023. Accessed: Apr. 05, 2024. [Online]. Available: <http://arxiv.org/abs/2311.05232>
- [37] Z. Elyoseph and I. Levkovich, “Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study,” *JMIR Ment Health*, vol. 11, p. e53043, Mar. 2024, doi: 10.2196/53043.
- [38] A. Kosar, G. D. Pauw, and W. Daelemans, “Comparative Evaluation of Topic Detection: Humans vs. LLMs,” *Computational Linguistics in the Netherlands Journal*, vol. 13, pp. 91–120, Mar. 2024.
- [39] C. D. P. Laureate, W. Buntine, and H. Linger, “A systematic review of the use of topic models for short text social media analysis,” *Artif Intell Rev*, vol. 56, no. 12, pp. 14223–14255, Dec. 2023, doi: 10.1007/s10462-023-10471-x.
- [40] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” arXiv, Mar. 11, 2022. doi: 10.48550/arXiv.2203.05794.
- [41] L. Zheng, Y. Sheng, W.-L. Chiang, H. Zhang, J. E. Gonzalez, and I. Stoica, “Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings | LMSYS Org.” Accessed: Apr. 05, 2024. [Online]. Available: <https://lmsys.org/blog/2023-05-03-arena>
- [42] F. J. Grajales 3rd, S. Sheps, K. Ho, H. Novak-Lauscher, and G. Eysenbach, “Social media: a review and tutorial of applications in medicine and health care,” *J Med Internet Res*, vol. 16, no. 2, p. e13, Feb. 2014.
- [43] GPT-4 API result not stable with temp 0 - API,” OpenAI Developer Forum. Accessed: Apr. 09,

2024. [Online]. Available: <https://community.openai.com/t/gpt-4-api-result-not-stable-with-temp-0/302139>
- [44] J. Li, Q. Zhang, Y. Yu, Q. Fu, and D. Ye, “More Agents Is All You Need.” arXiv, Feb. 03, 2024. doi: 10.48550/arXiv.2402.05120.
- [45] J. Elder, “The Apparent Paradox of Complexity in Ensemble Modeling\*,” in *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*, R. Nisbet, G. Miner, and K. Yale, Eds., Boston: Academic Press, 2018, pp. 705–718. doi: 10.1016/B978-0-12-416632-5.00016-5.
- [46] G. Eysenbach, “Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health,” *American Journal of Preventive Medicine*, vol. 40, no. 5, pp. S154–S158, May 2011, doi: 10.1016/j.amepre.2011.02.006.
- [47] Mathison, “Mathison/coronavirus\_vaccine\_antivax\_research.” Aug. 15, 2023. Accessed: May 31, 2024. [Online]. Available: [https://github.com/Mathison/coronavirus\\_vaccine\\_antivax\\_research](https://github.com/Mathison/coronavirus_vaccine_antivax_research)

## Supplementary Files



## Figures

LLM Prompts used. (a) The content ranking prompt we used for all 3 LLMs varied slightly between LLMs but was basically as shown (this example was used for GPT4). (b) The content analysis prompt we used for all 3 LLMs varied slightly between LLMs but was basically as shown in Figure 1b (this example was used for GPT4).

A	B
<p><i>"This is an academic research project where you are assisting in monitoring content of tweets; some may contain misinformation or bad language. Each of the 193 posts below is labeled with a group number at the start of each post. There are 20 groups of posts. Please read ALL of the posts in each group. Your goal is to rank all 20 groups of posts in descending order, starting with groups whose content most closely represents posts calling out or making claims about public figures as anti-vaxxers or that called out groups of people such as scientists or political parties and in general that best represent groups of posts about anti-vaccination, and then going down as the content of post groups diverges from that goal. Here is more description to help you rank them from best (TOP RANKING) to worst (LOWER RANKING):</i></p> <p><i>1st: TOP RANKING GROUPS: List these first. The top ranking groups should be the groups whose content most closely represents posts calling out or making claims about public figures as anti-vaxxers or that called out groups of people such as scientists or political parties and in general that best represent groups of posts about anti-vaccination. Put the best one first and then in descending order.</i></p> <p><i>then</i></p> <p><i>2nd: MIDDLE RANKING: List these after the TOP RANKING groups. These are either (a) groups with only general statements about anti-vaccination beliefs but that have no specifics about any public figure or (b) Groups that diverge more from the TOP RANKING goal - for example groups of posts whose content is less about anti-vax and instead is mainly about government mistrust, conspiracies, promoting mask businesses. Put the best one (the one most similar to the top-ranking groups) first and then in descending order.</i></p> <p><i>then</i></p> <p><i>3rd: LOWER RANKING: List these last. These are groups that diverge the most from the groups mentioned above. Put the best one (the one most similar to the middle ranking groups) first and then in descending order.</i></p> <p><i>OUTPUT: For groups 1-20, list each group number in a descending rank order where the top of the list is the most relevant highest ranking group and the bottom of the list is the absolutely least relevant lowest ranking group. Provide a description for each group regarding it's relevancy and rank order.</i></p> <p><i>Do not create new posts following the ones given.</i></p> <p><i>Posts begin here:</i></p> <p><i>GROUP=0, POST CONTENT: [1<sup>st</sup> post content was here]</i></p> <p><i>GROUP=0, POST CONTENT: [2<sup>nd</sup> post content was here]</i></p> <p><i>GROUP=1, POST CONTENT: [3<sup>rd</sup> post content was here]</i></p> <p><i>...</i></p> <p><i>GROUP=1, POST CONTENT: [193<sup>rd</sup> post content was here]"</i></p>	<p><i>"This is an academic research project where you are assisting in monitoring content of tweets; some may contain misinformation or bad language.</i></p> <p><i>Read all of the content and then define 5 most clinically, socially, epidemiologically or otherwise significant themes as follows:</i></p> <p><i>Theme1</i></p> <p><i>a) Title: Descriptive title of theme (avoid specific personal names in the title):</i></p> <p><i>"[Title]"</i></p> <p><i>b) Definition: using 3-5 sentences provide a definition and describe the theme and any clinical, epidemiological, social and other significances of the theme; personal names are acceptable in this section.</i></p> <p><i>c) provide 2 verbatim examples of posts, using examples only from the set of posts provided to you (do NOT make any examples up or you will get in big trouble), each in quotes, followed by an explanation of how it is representative of the theme.</i></p> <p><i>Theme2</i></p> <p><i>a) etc. repeat above for Themes 2-5.</i></p> <p><i>Do not create new posts following the ones given.</i></p> <p><i>The content begins here:</i></p> <p><i>Post 1: [1<sup>st</sup> post content was here]</i></p> <p><i>Post 2: [2<sup>nd</sup> post content was here]</i></p> <p><i>Post 3: [3<sup>rd</sup> post content was here]</i></p> <p><i>...</i></p> <p><i>Post 768: [768th post content was here]"</i></p>

## **Multimedia Appendixes**

Table 6. Example of Claude 1 inductive thematic analysis, and Table 7. Example of Claude 2 inductive thematic analysis.  
URL: <http://asset.jmir.pub/assets/d34bc31f3cf729f8754ec869ce0207d4.docx>

