

Does Geo-Co-Location Matter? A Case Study of Public Health Conversations During COVID-19

Paiheng Xu, Louiqa Raschid, Vanessa Frias-Martinez

University of Maryland, College Park
{paiheng,lrashid,vfrias}@umd.edu

Abstract

Social media platforms like Twitter (now X) have been pivotal in information dissemination and public engagement. The objective of our research is to analyze the effect of localized engagement on social media conversations. This study examines the impact of geographic co-location, as a proxy for localized engagement. Our research is grounded in a COVID-19 dataset. A key goal during the pandemic for public health experts was to encourage prosocial behavior that could impact local outcomes such as masking and social distancing. Given the importance of local news and guidance during COVID-19, we analyze the effect of localized engagement, between public health experts (PHEs) and the public, on social media. We analyze a Twitter Conversation dataset from January 2020 to November 2021, comprising over 19 K tweets from nearly five hundred PHEs, and 800 K replies from 350 K participants. We use a Poisson regression model to show that geo-co-location is indeed associated with higher engagement. Lexical features associated with positive sentiment, emotion and personal experiences were more common in geo-co-located conversations. The relationship between geo-co-location and engagement also varies by the PHE's profession. To complement our statistical analysis, we also applied a large language model (LLM)-based method to explore how linguistic and thematic content varied by geo-co-location, offering a complementary lens on discourse differences. This research provides insights into how geographic co-location influences social media engagement and can inform strategies to improve public health messaging.

1 Introduction

Social media platforms such as Twitter (now X) play a crucial role in information dissemination and public engagement. Our research aims to analyze the effect of localized engagement on social media conversations. We use geographic co-location (defined in the paper) as a proxy for localized engagement. The COVID-19 pandemic presented a complex communication challenge: public health experts had to provide vital guidance, in a scenario where the science was constantly being updated, and in the midst of a social media deluge of mis- and dis-information. A key goal for public health experts (PHEs) was to encourage pro-social behavior that could impact In Real Life (IRL) outcomes

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

(Fischer 2020; Miles et al. 2022). Given the importance of local news and guidance during COVID-19, our research aims to analyze the effect of geo-co-located engagement on social media conversations. Prior research has shown that physical distance between social media users impacts the likelihood of online interactions, but not necessarily the content of those interactions. For example, the likelihood of interaction decreases with greater physical distance (Leskovec and Horvitz 2008). Bozarth et al. (2023) found that news sharing and user interactions occur more often within a state on Reddit, when compared to across states. While there has been some recent research on social media engagement during COVID-19 there is limited research on conversations involving PHEs and the public (Rao et al. 2025).

In this paper, we hypothesize that geographic co-location between public health experts (PHEs) and the public is associated with greater engagement on social media. Our novel research advances the state of the art by analyzing the impact of geographic co-location between PHEs and the public on the following: (1) Engagement as measured by the count of participants and replies per conversation around an original tweet. (2) The linguistic style, emotional tone, and COVID-19 topics expressed in conversations. (3) The profile (profession) of the public health expert. We note that our research is the first to study both the relationship between geographic co-location and engagement, and the content characteristics of conversations influenced by such geo-co-location. The study also contributes to the literature on the effectiveness of public health messaging during COVID-19.

Our research is based on an X (Twitter) dataset spanning January 2020 through November 2021. Starting with a seed set of thirty public health experts (PHEs), we identified a larger group of almost five hundred PHEs or their expert peers in adjacent domains, e.g., medical communication. We then curated a dataset of conversations, comprising over 19 K posts (tweets) from the PHEs, and approximately 800 K replies from almost 350 K participants in the conversations. We used Carmen (Dredze et al. 2013; Zhang, DeLucia, and Dredze 2022) to determine geo-location (US state). We then identified geo-co-located pairs - where the PHE and participant are in the same US state, and non-geo-co-located pairs, respectively. Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) was used to identify lexical features that capture levels of self-disclosure and psy-

chological processes. An LLM-based method (Zhong et al. 2022, 2023) was used to further explore how linguistic and thematic content varied by geo-co-location, offering a complementary lens on discourse differences. Our analysis reveals the following key findings:

1. We find that engagement is significantly higher when the PHEs and the public are geo-co-located, compared to when they are not. This result is based on a Poisson regression model predicting reply counts for each original tweet. The finding is robust when we apply a mixed-effects specification that accounts for author-level heterogeneity, and is also supported by a non-parametric statistical test comparing engagement rates across all tweets.
2. The sharing of personal experiences and emotional expressions, e.g., anxiety, in the original post, are associated with higher geo-co-located engagement. These findings are confirmed using a large language model (LLM) based approach to generate and validate hypotheses.
3. Replies from geo-co-located participants are more positive in sentiment, and express more emotions and personal issues, when the PHEs initially share personal experiences / feelings.
4. The relationship between geo-co-location and engagement varies by the PHE's profession.

To summarize, this research provides novel insights on the impact of geo-co-location, on engagement. It has the potential to inform how public health messaging strategies can harness geo-co-location to enhance engagement and to maintain a positive discourse around public health guidance.

2 Related Work

Geo-co-location in social networks Many studies have shown that users on social media platforms tend to connect or interact preferentially with users who are geographically closer over those who are farther away (Leskovec and Horvitz 2008; Laniado et al. 2018). Our study focuses on message-exchange interactions such as replying to an original post, on platforms such as Twitter, where interactions typically reflect a shared interest in specific topics. We specifically investigate geo-location of users; this differs from prior research on the correlation between user geo-location and content (Pavalanathan and Eisenstein 2015; Cheke et al. 2020). Twitter interactions in the form of replies and mentions have been successfully used as predictors of geographically closer connections, but there is limited research on the influence of geo-co-location on user interactions. Cuevas et al. (2014) count the number of retweets and compare how tweets are shared across countries. Bozarth et al. (2023) studied the geographical diffusion of news articles on Reddit. None of these studies investigated the specifics of social media interactions. The data collection for this research was facilitated when Twitter released API v2 in late 2020; the API supported the retrieval of a conversation, i.e., retweets and replies to an original post.

Geo-location data on Twitter Twitter provides multiple sources of geo-location information in the metadata of tweets and users, including coordinates, a Place object

within a tweet, and location information within a user profile. Further, the geo-location of a tweet can also be inferred from its content (Izbicki, Papalexakis, and Tsotras 2019). Given that our study focuses on the geo-location of users, we rely on tools that extract geo-location from user-provided metadata in tweets and user profiles (Dredze et al. 2013; Zhang, DeLucia, and Dredze 2022). We acknowledge that Twitter users and those who use Twitter's geo-location services may be a biased sample of the user population. Despite that limitation, Twitter's geo-location data has been shown to be valuable in many applications such as disaster response (Hong and Frias-Martinez 2020), urban planning (Frias-Martinez et al. 2012) and public health surveillance (Jordan et al. 2018; Xu, Dredze, and Broniatowski 2020; Xu, Broniatowski, and Dredze 2024).

Content understanding of social media interactions

Numerous text analysis tools have been employed to comprehend user interactions on social media. This includes topic modeling (Cheke et al. 2020; Li et al. 2024, 2025), linguistic analysis (Pavalanathan and Eisenstein 2015; Wood-Doughty et al. 2021; Choi et al. 2021), text classification models for constructs like sentiment (Lwin et al. 2020). More recently, large language model (LLM) based approaches have shown the potential for more nuanced and deeper analysis of social interactions (Ziems et al. 2024; Zhou et al. 2025). In this research, we use a wide range of text analysis tools, including topic modeling (Eisenstein, Ahmed, and Xing 2011), LIWC for linguistic analysis (Pennebaker et al. 2015), sentiment classification (Loureiro et al. 2022), and LLM-based methods for hypothesis generation, in the context of studying geo-co-located engagement.

3 Data

The Twitter conversation dataset used in this study is centered around COVID-19 discussions. The creation of the dataset commenced with a curated seed user list of 30 Public Health Experts (PHEs) (see Appendix A), both academics and health professionals, who were active in COVID-19 discussions. These 30 experts were handpicked by academic colleagues in public health. PHEs include individuals with advanced degrees in medicine, epidemiology, genomics, infectious diseases, public policy, and economics (Rao et al. 2025). A retweet network of these 30 seed experts was then constructed using a dataset of over 1 billion publicly available COVID-19 Tweets (Chen et al. 2020). All users in the retweet network were ranked using eigenvector centrality (Ghosh and Lerman 2010), and the Top 500 ranked users were identified. Eigenvector centrality measures a user's influence based not only on how many times they are retweeted, but also on the influence of the users who are retweeting them. We then manually removed organizations or bots. We obtained an expanded set of 489 PHEs or PHE adjacent influential users. An example of a PHE-adjacent user may be a health reporter, a policy expert, or an appointee in a health agency, as well as other influential users (high-centrality nodes) in the retweet network.

The next step was to collect all of the original posts (tweets) of PHEs, and the resulting conversations, com-

prised of replies¹. The dataset covered a period from the early days of the pandemic, starting in January 21, 2020, through November 4, 2021. We only consider the original posts (tweets) of the 489 PHEs, i.e., we exclude their own retweets, replies, and quoted tweets, giving a dataset of 144 K tweets. The final step was to collect all of the replies for the original tweets. Unfortunately, restrictions and rate limits imposed by Twitter on the academic API limited our ability to rehydrate all of the replies (conversations). We were able to obtain all reply tweets for 19.5 K original tweets, from 462 PHEs, that had at least one reply. The final dataset included approximately 786 K reply tweets from 345 K unique participants in conversations with the 462 PHEs. Each tweet received an average of 40.24 replies, with a median of 5 replies per original PHE tweet, and approximately 70% of the tweets did not have replies.

In this paper, we refer to an original PHE tweet as a *tweet* and a reply tweet as a *reply*. We refer to the PHEs as either *authors* or *PHE authors*. All users who responded with a reply tweet to an original tweet are referred to as *participants* in the conversation. Finally, we refer to the union of participants who reply to any of the original tweets of a PHE author as the *audience* of the PHE author.

4 Geo-Co-Location and Engagement

We first compare the activity statistics of PHE authors and participants who disclose geo-location and those who do not. We then define the geo-co-located (*gcl*) audience and participant count (reply count) for a tweet. We use a Poisson regression model that predicts the reply count to demonstrate the higher engagement of *gcl* participants. We then define the engagement rate and use the Kolmogorov-Smirnov (KS) test (Hodges Jr 1958) to determine the statistical difference between the *gcl* and *non-gcl* engagement rates.

4.1 Geo-Located Activity Statistics

We use the Carmen library (Dredze et al. 2013; Zhang, DeLucia, and Dredze 2022) to infer the geo-location of the PHE authors and the participants. Carmen uses the following features: location information and self-description from the user profile, coordinates, and the Place object from the metadata associated with tweets. The most frequently mentioned (inferred) state is selected as the home state for the user. We note that Carmen could infer the geo-location (state) of over 74% of the PHEs (Table 1) and over 1/3 of the participants (Table 2). We chose to conduct our analysis at the state level to ensure data reliability. The accuracy of the Carmen tool is substantially higher at this level (over 99%) compared to city-level analysis (around 53%).

We report on activity statistics comparing PHE authors / participants who share / do not share geo-location, in Tables 1 and 2, respectively. The mean and standard deviation (std) of the count of replies per participant, for tweets from PHE authors with geo-location, is 2.29 ± 6.55 , in comparison to 1.91 ± 7.82 for tweets from PHE authors

without geo-location. We used the two-sample Kolmogorov-Smirnov (KS) test (Hodges Jr 1958), a non-parametric statistical test, to assess the statistical significance of the difference between the samples. It rejected the null hypothesis ($p < 0.001$), pointing to a statistically significant increase in engagement from participants for PHE authors with geo-location. The mean and std of the count of replies per tweet, for PHE authors with geo-location is 23.08 ± 53.93 , in comparison to 25.98 ± 60.79 for PHE authors without geo-location. However, that difference is not statistically significant. We conclude that geo-location information for PHE authors does appear to impact engagement at the participant level, with higher engagement for PHEs that have geo-location information.

Next, we consider the statistics for the *participants* in Table 2. We observe that participants with geo-location information post significantly more replies, with a mean and standard deviation of 2.45 ± 8.33 , than participants without geo-location, with mean and std of 2.38 ± 7.43 . This difference is statistically significant (two-sample KS test rejected the null hypothesis with $p < 0.05$).

To summarize, these activity statistics reflect that sharing geo-location information is an important signal for engagement in social media conversations, for both PHE authors and participants. This finding is also consistent with prior results that social media users who share their location information are more active (Rzeszewski and Beluch 2017).

4.2 A Model of Engagement

Geo-co-location (*gcl*) For each (tweet, reply) pair in the dataset, we refer to the participant being geo-co-located (*gcl*) with the author, if they are both located in the same state. Conversely, the participant and the author are not geo-co-located (*non-gcl*) if they are located in different states. Participants without geo-location are excluded from the *gcl* and *non-gcl* groups for further analysis.

Audience For a tweet T from author A , we define the *gcl audience* W as the set of all *gcl participants* across all the tweets created by A in the dataset. We can similarly define the *non-gcl audience* for A . Since the popularity of an author can vary over time, we define the audience as a cumulative measure over the time period of the data collection. We note that while *gcl* participants are limited to those from the same state as the tweet author, *non-gcl* participants can be from all other states. Hence, the *gcl* audiences are inherently smaller than the *non-gcl* audiences.

We model engagement using a Poisson regression model that can predict the count of *gcl* and *non-gcl* participants (reply count) while controlling for the difference in the audience count. For each tweet, we treat the count of *gcl* and *non-gcl* replies as separate data points for the regression model. Each data point is also associated with a categorical variable indicating the origin as the *gcl* or *non-gcl* audience.²

²The model assumes that the decision to reply from the two audiences of each author are independent. Given that more than 40% of replies in the dataset are directly replying to the original tweets and 95% of the participants are *non-gcl* or *no-loc*, interactions between the *gcl* and *non-gcl* audience are expected to account for

¹<https://developer.twitter.com/en/docs/twitter-api/conversation-id>

	# Authors	# Tweets	# Replies	# Participants	Mean (std) followers	Mean (std) replies per tweet	Mean (std) replies per participant
With state	342	16,465	380k	166k	157k (654K)	23.08 (53.93)	2.29 (6.55)
No state	120	4,432	115k	60k	160k (438K)	25.98 (60.79)	1.91 (7.82)

Table 1: Statistics for PHEs with and without geo-location in the COVID-19 Conversation Dataset. We **bold** the cases where the comparison between the PHEs with and without state is significant using the two-sample KS test ($p < 0.001$).

	# Participants	Mean (Std) replies
With state	73,912	2.45 (8.33)
No state	131,760	2.38 (7.43)

Table 2: Statistics for participants with and without geo-location in the COVID-19 Conversation Dataset. We **bold** the cases where the comparison of participants with and without state is significant using the two-sample KS test ($p < 0.05$).

Thus, the dependent variable is the count of *gcl* or *non-gcl* participants (replies), denoted as Y_T , and the independent variable, X_T , is the audience type ($gcl = 0$ and $non-gcl = 1$). To control for the size of the audiences, we include the corresponding audience count W_T as an exposure variable. The model is formulated as follows:

$$\log(\mathbb{E}[Y_T]) = \beta_0 + \beta_1 \cdot X_T + \log(W_T). \quad (1)$$

We can interpret e^{β_1} as the multiplicative difference in the expected engagement between the two groups, accounting for exposure.

We also fit a mixed-effects Poisson regression (Seabold, Perktold et al. 2010) to account for author heterogeneity, which extends Eq. (1) by including a random intercept (u_j) for each author:

$$\log(\mathbb{E}[Y_{ij}]) = \beta_0 + \beta_1 X_{ij} + u_j + \log(W_{ij}), \quad (2)$$

where the variables correspond to those in the main model but are indexed by tweet i and author j .

The regression results in Table 3 (simpler model) show that the engagement of the *non-gcl* audience is $e^{-0.12} \approx 88.6\%$ of that of *gcl* audience, while controlling for the audience size ($N = 31,852$, $p < 0.001$).³ For example, if the engagement (replies per original tweet) for the *gcl* group is x , then the engagement for the *non-gcl* group would be approximately $.886 \times x$. This indicates that the *non-gcl* audience tends to engage at a lower rate compared to the *gcl* audience, even after accounting for differences in the audience size. The mixed-effects model further confirms our findings. See Appendix B for details.

4.3 Comparison of Engagement Rates

The Poisson regression model demonstrated the greater engagement of the *gcl* audience across all tweets in the dataset.

only a small proportion of the replies.

³We also fitted a Negative Binomial model with the same setup and obtained similar results. However, the estimated dispersion parameter α was very small, indicating minimal over-dispersion, suggesting that the Poisson model is more appropriate.

Variable	Coefficient	IRR	95% CI (IRR)
Intercept	-4.92	0.007	(0.007, 0.007)
$X_T(gcl = 0, non-gcl = 1)$	-0.12	0.886	(0.870, 0.901)

Table 3: Poisson regression model to predict the effect of engagement type ($gcl = 0$; $non-gcl = 1$) on the count of replies per original tweet. IRR is the incidence rate ratio. Statistically significant results ($p < 0.001$) are bolded.

We are further interested in comparing the difference between *gcl* and *non-gcl* engagement, per original tweet, so as to identify the characteristics of interesting subsets that exhibit the greatest difference in the level of engagement. We use a measure of engagement rate for this analysis.

Engagement Rate The *gcl engagement rate* for a tweet T from author A is defined as the fraction of A 's *gcl audience* that replied to the tweet T :

$$gcl \text{ engagement rate} = \frac{\text{count of } gcl \text{ participants for } T}{\text{count of } gcl \text{ audience for } A}$$

The *non-gcl engagement rate* for tweet T of author A is similarly defined using the *non-gcl* participants and *non-gcl* audience.

The *gcl* (or *non-gcl*) engagement rate for tweet T , representing the fraction of audience participation, enables a direct comparison between the two levels of engagement per tweet. A direct comparison would not have been possible using the un-normalized *gcl* and *non-gcl* reply counts from the previous Poisson model. On average, the *gcl* replies take up 5.7% of the replies for a tweet while *non-gcl* replies take up 29.4%. For this computation, we filter the dataset to only include authors that have both audience types, which ensures that for every tweet in the analysis, both *gcl* and *non-gcl* engagement rate can be calculated. The filtered subset includes about 97% of the tweets from authors with state information.

Table 4 displays the engagement rates (ER) for *gcl* and *non-gcl* audiences, as well as for audiences where the location of the participant could not be inferred (labeled as *no-loc*). The table shows that a higher average engagement rate is observed for *gcl* audiences, in comparison to *non-gcl* audiences (mean values: 1.89% vs 1.76%). A two-sample KS test confirmed the significance of this difference ($n = 15,926$, $p < 0.001$). In addition, the *no-loc* engagement rate is very similar to the *non-gcl* rate (mean values: 1.76% vs 1.78%). This validates that our focus on comparing *gcl* versus *non-gcl* subsets is a reasonable experimental filter.

Figure 1 provides a histogram of the *gcl* (red hatched bars) and *non-gcl* (blue bars) engagement rates. The significant rightward skew of the *gcl* histogram clearly illustrates its

	Mean (Std) participants per tweet	Mean (Std) ER
<i>gcl</i>	0.99 (3.32)	1.89% (8.05%)
<i>non-gcl</i>	5.16 (13.09)	1.76% (6.10%)
<i>no-loc</i>	11.18 (26.56)	1.78% (5.53%)

Table 4: Engagement rate (ER) for participants. *no-loc* means there is no location information for the participants. We **bold** the statistic comparison between *gcl* and *non-gcl* groups that is significant from a two-sample KS test ($p < 0.001$).

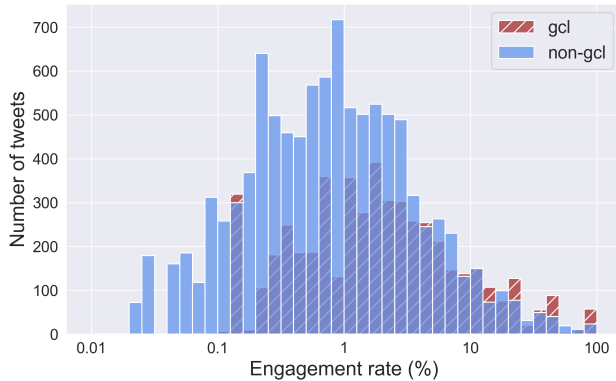


Figure 1: Distribution of *gcl* and *non-gcl* engagement rate. Tweets with 0% engagement rate are not shown in the plot.

dominance in the region with a higher engagement rate region; for example, we see the histogram extend to a 100% engagement rate.

In summary, the Poisson regression model of participant counts, the statistical significance of the *gcl* versus *non-gcl* engagement rate distributions using the KS test, and the insights from the histograms, confirm our hypothesis that geographical co-location between PHE authors and participants is associated with greater engagement on social media.

5 The Nuances of Geo-co-location on Engagement

Having established that there is higher engagement of the *gcl* audience, we next explore the nuances of this engagement as follows: First, we explore whether greater *gcl* (*non-gcl*) engagement is dependent on the type of COVID-19 issue being discussed. We then drill down into the content, to identify lexical features that may reveal additional insights into the emotions and concerns that are expressed in *gcl* versus *non-gcl* conversations. Finally, we determine if the PHE author profile, and in particular, their profession, has an impact on attracting greater levels of *gcl* engagement.

5.1 What Tweets Attract More *gcl* Engagement?

Impact of COVID-19 Topics We first examine whether the *gcl* engagement rates vary across Tweets with different topics, and how these variations relate to *gcl* versus *non-gcl* engagement. We adopt a method from Rao et al.

(2023) to identify tweets that discuss different COVID-19 topics. Their method extracts issue-relevant keywords from Wikipedia articles by identifying the most distinctive keywords when comparing issue-specific articles with general articles on the pandemic and politics in the U.S. using SAGE (Eisenstein, Ahmed, and Xing 2011). These manually verified keywords are then used to label tweets through keyword matching. Rao et al. (2023) validate their method by evaluating its results on a test sample with manual annotations. We replicate this method and retain topics that have F1 scores higher than 0.7; the topics are *masking*, *lockdowns*, *education*, and *vaccines*. Table 12 shows sample tweets.

A descriptive comparison of engagement rates for tweets discussing the four topics is shown in Table 5. We observe that tweets on *masking* are associated with the largest gap between *gcl* and *non-gcl* engagement. This is consistent with Rao et al. (2023) that finds that *masking*-related tweets gained more replies. Similarly, *lockdown* and *education* also show higher engagement rates for *gcl* audiences. On the other hand, **vaccine-related tweets are associated with more engagement from non-gcl audiences**. These differences were all statistically significant through bivariate KS tests. The discussions around topics such as *masking*, *lockdowns*, and *education* often had a strong local angle, and this may have led to a greater level of *gcl* engagement. In contrast, *vaccine*-related discussions often take place within the context of a national policy (Karami et al. 2021).

To test the robustness of these associations, we employ a multivariate linear regression model to examine the relationship between topics and engagement while controlling for an author’s number of followers, a recognized driving factor in social media engagement (Cha et al. 2010; Bakshy et al. 2011). We define the model as:

$$\text{Engagement Rate} = \beta_0 + \sum_{T_i \in \text{Topics}} (\beta_i T_i) + \beta_c F + \epsilon,$$

where T_i is the binary indicator for each topic, F is a control for the author’s min-max-normalized follower, and β_0 and ϵ are the constant and error term respectively. We repeat the analysis for three types of engagement rate as the dependent variable, i.e., *gcl*, *non-gcl* and the difference between *gcl* and *non-gcl* (denoted as *Diff*). Table 13 (Appendix E) shows that, after controlling for authors’ followers, only the topic of *vaccines* remains a statistically significant factor.⁴ This suggests the association between *vaccine* discussions and *non-gcl* audience is a more robust finding, while the initial associations observed for the other topics are not significant when author popularity is controlled in the model.

Impact of Lexical Features Our next step is to drill down into the content, to identify lexical features that may reveal additional insights into the emotions and concerns expressed in *gcl* versus *non-gcl* conversations. LIWC has been extensively used to study psychological patterns and emotional states in social media texts, providing insights into user behavior and social dynamics (Schwartz et al. 2013; Jiang and

⁴The results are consistent with the topic coefficients shown in our full lexical model in Table 8.

	Masking	Lockdowns	Education	Vaccines
<i>gcl</i> ER	2.32%	1.49%	1.94%	1.52%
<i>non-gcl</i> ER	1.88%	1.39%	1.80%	1.73%
Diff	0.44%	0.10%	0.14%	-0.21%
# authors	162	107	168	202
# Tweets	929	326	812	2240

Table 5: Difference between *gcl* and *non-gcl* engagement rate (ER) for Tweets discussing different COVID-19 topics. Significant ($p < 0.001$) KS test results for *gcl* ER and *non-gcl* ER are bolded.

Wilson 2018; Choi et al. 2021). To explore whether psychological and emotional states - measured via LIWC categories - might affect the *gcl* engagement rates, when compared to *non-gcl* audiences, we propose the following approach. We count the occurrences of the LIWC categories in each original Tweet and normalize them by the number of tokens in the Tweet. We use the TweetTokenizer from NLTK (Bird, Klein, and Loper 2009) for tokenization. Next, we run coefficient regression analyses with the occurrences of LIWC categories in the original Tweets as independent variables, and engagement rates as dependent variables, to assess the effect of LIWC categories on engagement rates.

We determine a list of LIWC categories relevant to the COVID-19 context based on a preliminary lexical analysis and existing literature. Categories in LIWC (Pennebaker et al. 2015) encompass two dimensions: grammar and psychological processes. We select several subcategories in grammar and most of the subcategories in psychological processes. We include: first person singular pronouns (*i*), first person plural pronouns (*we*), and second person pronouns (*you*) because the usage of pronouns reveal levels of self-disclosure (De Choudhury and De 2014; Wang, Burke, and Kraut 2016) and they are also some of the most distinctive features in the tweets with high *gcl* engagement rate based on our preliminary analysis (see Appendix D). Additionally, we choose subcategories in psychological processes potentially relevant to COVID-19 discussions.⁵ These include: (1) all subcategories in the affective process, i.e., positive emotion (*posemo*), negative emotion (*negemo*) and its subcategories (anxiety (*anx*), anger and sadness (*sad*)); (2) selected subcategories in the social processes, i.e., *family* and *friends*; (3) selected subcategories in the biological processes, i.e., *body* and *health*; (4) all subcategories in the personal concerns, i.e., *work*, *leisure*, *home*, *money*, *religion* (*relig*), and *death*; and (5) swear words (*swear*) from informal language, resulting a total of 19 variables. We define the model as:

$$\text{Engagement Rate} = \beta_0 + \sum_{L_j \in \text{LIWC}} (\beta_j L_j) + \sum_{T_i \in \text{Topics}} (\beta_i T_i) + \beta_c F + \epsilon,$$

⁵We exclude subcategories that are hard to interpret in this context, such as gender references, cognitive processes, time orientations, and relativity.

where L_j and β_j are selected LIWC categories and their coefficients. The topic indicator, T_i , and follower count, F , are included as control variables.

Table 8 shows the coefficient regression analysis. Similar results were observed without the controls. We focus on the results using *Diff* as the dependent variable. Recall that the *non-gcl* audience counts are inherently larger than the *gcl* counts. The counts are proportional across authors, i.e., an author with a larger *non-gcl* audience is more likely to have a larger *gcl* audience. The results show that using *Diff* cancels out the effect of LIWC categories that are significant for both *gcl* and *non-gcl* engagement rates, for e.g., *posemo*, *family*. It allows us to focus on the LIWC categories that are significant to the *gcl* engagement rate.

We observe that the coefficients for *i* and *anx* are significant and positive, indicating that a 1% increase of word usage in these two categories would increase the difference between *gcl* and *non-gcl* engagement rates by 0.1% and 0.13% respectively (*gcl* ER is higher). This suggests that **PHEs sharing personal experiences and feelings (especially anxiety) are associated with relatively higher engagement from *gcl* audiences than that of the *non-gcl* audiences**. We also experimented with a mixed-effect model (Lindstrom and Bates 1988) grouped by authors and found similar results (Table 15 in Appendix E).

Qualitative Analysis Table 6 presents sample tweets and their *gcl* and *non-gcl* replies. The author of the first tweet is located in California while the tweet is about an event in Michigan, resulting in a higher *non-gcl* engagement rate. The next two examples demonstrate that sharing personal experiences (with first-person pronouns) is associated with relatively higher *gcl* and *non-gcl* engagement rates. However, we note that in both cases, the *gcl* engagement rate is higher than the *non-gcl* rate. This is consistent with the regression results in Table 8, given the occurrence of *i* (sharing personal experiences or feelings) in the tweets.

LLM-based Content Analysis We employed an approach that leveraged LLMs to identify distributional differences between two text corpora (Zhong et al. 2022, 2023). As one option to create two corpora, we used the *Diff* metric to select original tweets in the top 20% (higher *gcl* engagement) and the bottom 20%. For this comparison, we randomly sampled 1000 tweets from each corpus to serve as an exploration (training) set and another 1000 tweets per corpus as the validation set. we used GPT4o-2024-08-06 as a “proposer” LLM, which was prompted with sampled tweets to generate $k = 20$ hypotheses describing how the two corpora differ. Each hypothesis was then evaluated by a separate LLM, which estimated a validation score, V' . We adopted the T5 model with 3B parameters that are fine-tuned for judging whether a hypothesis is true given a text sample.⁶ A positive V' indicates that the hypothesis is more prevalent in the top 20% corpus (higher *gcl* engagement). The absolute value $|V'|$ reflects the strength of this difference. We report on these results in Table 7.

⁶https://huggingface.co/ruiqi-zhong/d5_t5_validator_3B

Tweet - Author state	<i>gcl</i> ER	<i>non-gcl</i> ER	<i>gcl</i> reply - participant state	<i>non-gcl</i> reply - participant state
2 emerging US surges: Michigan (B.1.1.7) and Northeast (B.1.1.7+B.1.526). Both variants are vaccine responsive. They need very aggressive vaccination efforts – CA	0.7%	1.8%	The little inflection that is not receiving adequate attention [URL] – CA	Michigan has a comprehensive VAX plan in place. [url] – MI
Anybody have ideas for pandemic-era holiday gifts? <u>I</u> m thinking books, obviously, and magazine subscriptions. Gift certificates to nearby restaurants with delivery or safe take-out? Fun masks – to grim? Outdoor gear? [url] – DC	42.9%	5.7%	Gift certificates to nearby theatres that are streaming like @[user] – DC	Netflix subscription, Zoom subscription.... Oh, do you mean for me or other people? – WA
Had another one of my <u>I</u> m in a crowded indoor place where no one is masked and then <u>I</u> suddenly remember Covid” dreams last night. How many of you have those too? – NY	41.7%	38.3%	All the time. But I wonder in the dream, “wait. Is ‘it’ over so it’s ok no one around me is masked??” – NY	I’ve had about 15 ... no joke! Horror films in my head running for exits through unmasked hoards. – CA

Table 6: Sampled tweets and replies with *gcl* and *non-gcl* engagement rate (ER). First-personal pronouns in the tweets that indicates self-disclosure are underlined.

To validate the reliability of the V' score, one author annotated a random sample of 100 tweet-hypothesis pairs. The agreement between the human judgments and the validator LLM was strong, with an accuracy of 96% and a Cohen’s Kappa of 0.896.

From Table 7, hypotheses with the highest positive values of V' indeed share personal experiences and contain emotional expressions. Thus, the complementary LLM-based hypotheses are aligned with the results of our prior regression analyses and serve to confirm and validate our results.

To ensure the robustness of the findings, we experimented with prompt variations that involve different levels of contextual information in the prompts. Table 7 uses a minimal-context prompt shown in Table 16, while Zhong et al. (2023) adopts a more detailed prompt (Table 17) that includes explicit information about the corpora. However, we observed that when context about *gcl* is included, the LLM tends to hallucinate references to “locality” (Table 23).⁷ When we encourage the model to focus on linguistic and psychological patterns (prompt shown in Table 18), the overall content and distributional patterns remained similar.

We also considered additional options to create the two corpora, including varying the metrics used to define the groups, e.g., *gcl*, *non-gcl*, etc., and adjusting the thresholds. The dominant hypotheses continued to emphasize the sharing of personal experience and emotional expression when the threshold is 10%. Finally, the comparison between *gcl*-based and *non-gcl*-based groupings echoes patterns observed in our regression analysis (Table 8). Details and additional results are in Appendix F.

While the LLM-based approach may provide nuanced insights, there are several limitations and concerns that do

⁷We consider referring to locality as hallucination because the co-location between users and the mentioned location in the tweets cannot be inferred from the provided tweets in the prompts.

not arise with our prior statistical approaches and models. First, LLMs can hallucinate content. An example is that the LLM referenced `local` aspects in a hypothesis, as introduced earlier. We observe that this issue is largely mitigated when contextual information about *gcl* is excluded from the prompt. Although such hallucination is relatively easy to identify, other hypotheses and their associated validation scores may be more difficult to assess for accuracy and reliability. Further, the granularity and phrasing of hypotheses can be inconsistent. For instance, while the model often detects emotional tone, it rarely distinguishes between positive and negative sentiment. Hypotheses can range from being too vague to being overly specific, and they are often redundant. These limitations underscore that **while LLM-based methods may provide insights that go beyond the simpler LIWC features, these methods cannot be used in lieu of careful validation and traditional modeling.**

5.2 How *gcl* Replies Differ from *non-gcl* Ones?

Our linguistic analysis has focused on the original Tweets. Now, we shift to the replies and attempt to understand the differences in language usage between *gcl* replies and *non-gcl* replies. We first do a sentiment analysis and find *gcl* replies are slightly more positive and less negative than *non-gcl* replies (details in Appendix E). We then run a series of regression analyses to discover additional insights into the differences between *gcl* and *non-gcl* replies by analyzing the effect of LIWC features on *gcl* and *non-gcl* engagement.

From Original Tweets to Replies We extend our analysis by investigating how specific language (LIWC categories) in the original tweets might elicit similar language (LIWC categories) in their corresponding *gcl* and *non-gcl* replies. To carry out this analysis, we run logistic regressions with the independent variables being all the LIWC categories in the original Tweet T , and the dependent variable being one

Hypothesis	V'
include expressions of personal emotions or sentiments, such as fear, pride, or frustration	0.081
focus on personal experiences and anecdotes related to the pandemic	0.070
discuss everyday life adjustments and challenges during the pandemic, like dating or travel	0.061
incorporate personal milestones or special occasions and how they've been affected by the pandemic	0.055
...	
mention trends or statistics related to COVID-19 cases or immunity	-0.061
include updates or opinions about scientific research and data	-0.120

Table 7: LLM-generated hypotheses with $|V'| \geq 0.05$. A positive V' score indicates that the hypothesis is more prevalent in the top 20% corpus (higher *gcl* engagement), compared to the bottom 20% corpus. Full results in Appendix F.

	<i>Diff</i>	<i>gcl</i>	<i>non-gcl</i>
Selected LIWC categories			
i	0.102***	0.420***	0.318***
posemo	0.038	0.087***	0.049***
negemo	-0.070	0.012	0.082**
anx	0.128*	-0.052	-0.179***
family	0.104	0.236***	0.132**
death	0.020	-0.041	-0.061**
Control variables			
# followers	-0.001	0.011	0.012*
masking_True	0.004	0.004	0.000
education_True	-0.001	-0.001	-0.000
lockdowns_True	-0.003	-0.003	-0.000
vaccines_True	-0.004**	-0.005***	-0.001
R-squared	0.002	0.024	0.031
R-squared Adj.	0.000	0.023	0.029

Table 8: LIWC regression results of LIWC categories' impact on different types of engagement rates. We only show LIWC categories that are significant in at least one of the regression models (columns). * $p < .1$, ** $p < .05$, *** $p < .01$. Full table in Appendix E.

binarized LIWC category present in the replies for that tweet T .⁸ Formally, the model is defined as:

$$L_i^{\text{gcl-replies}} = \beta_0 + \sum_{L_i^T \in \text{LIWC}} (\beta_i L_i^T) + \epsilon,$$

where $L_i^{\text{gcl-replies}}$ is a binary value of 1 if the i -th LIWC category is present in at least one of the *gcl* replies to T , and L_i^T is the i -th LIWC score of T . We run this model for each

⁸A reply-level model would violate the assumption of independent observations, as all replies to a single tweet share identical independent variables.

binarized LIWC category (dependent variable), and for *gcl* and *non-gcl* replies, separately. A comparison of the regression coefficients for the *gcl* and *non-gcl* groups (see Figure 2) reveals a strong diagonal effect for both groups, with slightly larger diagonal coefficients in *gcl* replies for most personal concerns (*work*, *home*, *relig*, and *death*) and swear words. This suggests that *gcl* replies may be more likely to discuss personal concerns compared to *non-gcl* replies, when these concerns have been raised in the original Tweets. However, there is a stronger agreement in *non-gcl* replies for LIWC categories associated with affective processes, including *posemo* and *negemo* and all its subcategories, *anx*, *anger*, and *sad*. This suggests that, while *gcl* and *non-gcl* replies discuss similar types of personal concerns when they are mentioned in the original Tweets, *non-gcl* replies are more likely to share similar emotions to the original Tweets. We hypothesize that this may be due to *non-gcl* participants having weaker social connections with the authors, leading them to share similar emotions and potentially avoiding conflict, while *gcl* participants express more freely a diverse set of emotions independently of the ones that are raised in the original Tweets from local authors.

More interestingly, when observing the heatmaps horizontally, the coefficients for *i* are significant and positive in almost all regressions for *non-gcl* replies, while only significant and positive in *posemo*, social processes (*family* and *friend*), and some personal concerns (*leisure*, *home*, *relig*) for *gcl* replies. This pattern indicates that **authors sharing personal experiences tend to elicit all kinds of psychological and emotional states from *non-gcl* replies but mostly positive *gcl* replies. This finding reveals that geo-co-located messaging could be used as a strategy for enhancing engagement and maintaining a positive discourse in online communities.**

Difference in Replies To examine differences in language usage between *gcl* and *non-gcl* replies, we run a logistic regression where the dependent variable indicates whether a reply is *gcl* or *non-gcl* and the independent variables are the LIWC categories described in Section 5.1. Table 9 shows that negative emotions (particularly *anger*) are more prevalent in *non-gcl* replies; while the use of pronouns such as *we* and *you* is more common in *gcl* replies, which might suggest more direct communication with the *gcl* PHE authors. Interestingly, *gcl* and *non-gcl* replies are associated with different types of personal concerns. For example, *gcl* replies involve topics related to *work*, *home*, and *money* more frequently, while *non-gcl* replies are more likely to discuss *relig* and *death* and biological processes (i.e., *body* and *health*).

We applied the LLM-based method mentioned in Section 5.1 to examine distributional differences between *gcl* and *non-gcl* replies. As shown in Table 26, however, all generated hypotheses yielded V' scores below 0.05. This indicates that the method did not uncover any strong or consistent distributional distinctions between the two reply types, unlike the LIWC-based analysis, which revealed clear linguistic and topical divergences.

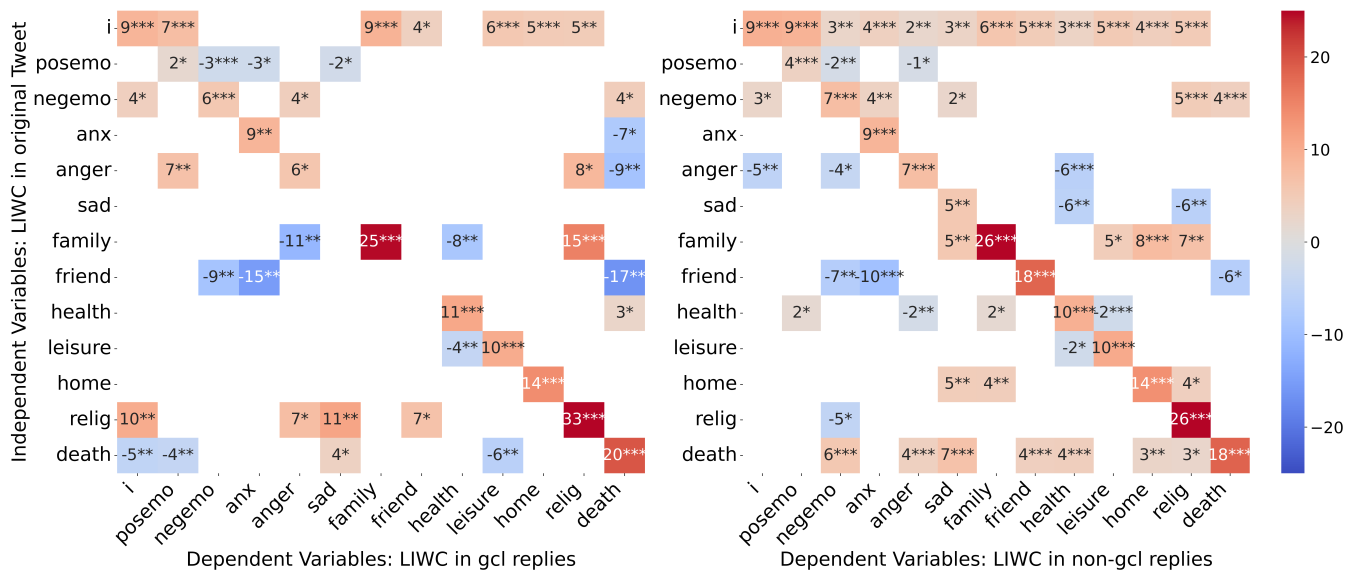


Figure 2: How *gcl* replies and *non-gcl* with specific LIWC categories correlate with LIWC features in the original Tweet. Each column is a logistic regression. Each cell is the coefficient. * $p < .1$, ** $p < .05$, *** $p < .01$. Only selected LIWC categories are shown for demonstration purposes. A complete version is in Figure 4.

Variable	OR	95% CI	Variable	OR	95% CI
i	0.63*	(0.39, 1.03)	body	0.14***	(0.05, 0.39)
we	4.25***	(2.15, 8.40)	health	0.27***	(0.14, 0.52)
you	7.73***	(5.12, 11.68)	work	2.08***	(1.38, 3.13)
posemo	1.04	(0.78, 1.40)	leisure	0.86	(0.37, 1.99)
negemo	0.39***	(0.20, 0.76)	home	44.48***	(12.12, 163.25)
anx	0.60	(0.16, 2.20)	money	3.85***	(1.78, 8.31)
anger	0.24***	(0.08, 0.69)	relig	0.29*	(0.11, 0.82)
sad	2.01	(0.73, 5.54)	death	0.17***	(0.06, 0.50)
family	0.25*	(0.06, 1.06)	swear	1.93	(0.74, 4.98)
friend	0.26	(0.05, 1.30)			

Table 9: Logistic regression results for *gcl* vs. *non-gcl* replies. Odds Ratios (OR) over 1 indicate a category is more frequent in *gcl* replies. * $p < .1$, ** $p < .05$, *** $p < .01$.

5.3 Which PHEs Have More *gcl* Engagement?

This analysis shifts the focus to PHE authors and explores PHE features that may attract more *gcl* engagement.

Profession The PHE authors provide detailed user profiles and generate substantial content. Similar to the use of LLMs to understand social context (Ziems et al. 2024), we use *gpt4* to predict the authors’ professions. Based on a preliminary manual review of a sample of author profiles, we selected four broad professional categories relevant to our analysis: media, academia, health, and politics. For each group, we included example professions in the prompt, e.g., journalists and reporters for the media profession. LLM generates a binary prediction for each group and the rationale for the prediction. More details are in Appendix H. The authors manually reviewed the classification results for 50 authors with above 90% accuracy for all four professions.

	Professions	Media	Academia	Health	Political
<i>gcl</i> ER		1.63%	2.09%	1.95%	1.72%
<i>non-gcl</i> ER		1.55%	1.89%	1.73%	1.65%
Diff		0.08%	0.20%	0.23%	0.07%
IRR		0.923	1.325	1.013	2.153
# authors		131	99	108	20
# Tweets		9107	5610	6734	1394
Test Acc.		88%	98%	94%	98%

Table 10: Differences between *gcl* and *non-gcl* engagement rates (ER) for tweets grouped by author profession. IRR are results from Poisson mixed-effects regression (Eq. 2); a value greater than 1 indicates a higher ER for the *gcl* group. Test Acc is *gpt4*’s accuracy on a sample of 50 authors. Statistically significant results ($p < 0.001$) are shown in bold.

The tweet-level analysis in Table 10 shows consistently higher *gcl* engagement rates across all professions. However, applying the mixed-effects Poisson regression (Eq. 2) separately to the tweets from each professional group, we observe that the relationship between geo-co-location and engagement varies by PHE’s profession. The strength of *gcl* engagement is the highest for political professionals (IRR = 2.153), followed by academic professionals (1.325). Media professionals are associated with significant but slightly lower *gcl* engagement (0.923), while health professionals show no significant difference (1.013).

6 Summary, Lessons Learned, Limitations and Future Work

Summary This study examines the role of geographic co-location in shaping social media engagement during the

COVID-19 pandemic. Using Poisson regression models and statistical tests, we find that users who are *gcl* with PHEs are associated with higher engagement rates. To explore the nature of this engagement, we analyzed conversational content, including topics, language use, and the characteristics of PHEs. We observe that emotional language and the sharing of personal experiences are associated with increased engagement from *gcl* participants, who more frequently express positive sentiment and personal issues in their replies. However, this relationship varies by PHE's professions. Our mixed-effects analysis reveals a positive association with *gcl* engagement for academic and political professionals, a modest negative association for media professionals, and non-significance for those in medical roles. We additionally applied an LLM-based method to generate hypotheses about content differences between tweet groups. Many of the resulting hypotheses, such as the emphasis on personal experience and emotional expression, were consistent with patterns observed in our regression analysis. However, limitations in specificity and reliability suggest that such methods should be interpreted with caution and warrant further refinement. Overall, these findings indicate that geographic co-location is an important factor to consider in the study of public health communication. Content that is locally resonant and personally expressive may be associated with higher engagement, offering possible directions for future research on effective messaging strategies in online public health discourse.

Lesson learned for public health messaging A key insight is that PHEs' effort to share personal experiences and use positive emotional language tends to receive higher *gcl* engagement and replies with more positive sentiment. This is an encouraging insight given that the overall COVID-19 discourse on social media is largely negative. While we have not attempted to explore causality, these findings suggest potential for PHEs to improve engagement by leading more localized discussions, linking guidance to community resources, and personalizing messages. Although PHEs cannot control their audience composition, they can take steps to strengthen connections with their *gcl* audience and foster more effective public health communication.

While the focus of this dataset was interactions between PHEs and the public during COVID-19, the analysis and models may be applied to other scenarios. Related work has studied engagement dynamics and polarization in the context of abortion and political discourse (Lerman et al. 2024; Nettasinghe et al. 2025). Research in (Rao et al. 2021) explored a national index of beliefs and attitudes around evidence based findings and science. Ongoing work aims to develop a national index of attitudes and emotions about community challenges including homelessness, substance abuse disorder, emotional health, etc. The hope is that such national indices, combined with findings on improved engagement, can lead to better approaches for PHEs and community decision makers to engage with the public.

Limitations and future work Twitter users who use location services may be more active (Rzeszewski and Beluch 2017), which may introduce a bias. Findings derived from

Twitter data may not generalize to other platforms or to populations that are less active or less visible on social media. This sampling bias may result in unequal targeting of messages, with engagement-optimized strategies disproportionately focusing on more vocal or digitally connected groups.

Additionally, users may provide inaccurate location information (Hecht et al. 2011) and the geo-location inference tool may be inaccurate. Our analysis uses state-level geo-location, a choice made to ensure data reliability, as the accuracy of currently available geo-inference tools degrades substantially at finer granularities (Zhang, DeLucia, and Dredze 2022; Masis and O'Connor 2024). The accuracy can be improved by using advanced geo-inference tools leveraging deep learning models (Zhang et al. 2023).

LIWC and the employed LLM-based method provided complementary insights, though both have limitations, i.e., LIWC in the scope of categories it captures, and the LLM method in consistency and specificity. As this is a rapidly evolving field (Zhong et al. 2024; Movva et al. 2025), these LLM-based methods should be interpreted with caution. While the natural language hypotheses are potentially more expressive than fixed LIWC categories, they can lack the precision of human-expert-curated ones, sometimes resulting in vague or imprecise statements. Furthermore, although the validator model showed strong agreement with human annotation on our test sample, its performance may not generalize to more complex or ambiguous text-hypothesis pairs. Finally, the method is purely descriptive. It is adept at identifying correlational patterns but cannot provide causal insight into why these differences in language emerge. Future work is needed to improve the specificity, robustness, and causal inference capabilities of these novel techniques.

There is also a risk that insights about effective message framing, such as the use of personal narratives or emotional language, could be misused to manipulate public sentiment rather than inform it. Future work should evaluate the impact of geo-co-location in other online contexts (e.g., Reddit), and expand content analysis to other domains beyond public health to assess the broader applicability and ethical considerations of this approach.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. CCF-2200256.

References

- Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.
- Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Bozarth, L.; Quercia, D.; Capra, L.; and Šćepanović, S. 2023. The role of the big geographic sort in online news circulation among US Reddit users. *Scientific Reports*, 13(1): 6711.

- Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, 10–17.
- Cheke, N.; Chandra, J.; Dandapat, S. K.; et al. 2020. Understanding the impact of geographical distance on online discussions. *IEEE Transactions on Computational Social Systems*, 7(4): 858–872.
- Chen, E.; Lerman, K.; Ferrara, E.; et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2): e19273.
- Choi, M.; Budak, C.; Romero, D. M.; and Jurgens, D. 2021. More than meets the tie: Examining the role of interpersonal relationships in social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 105–116.
- Cuevas, R.; Gonzalez, R.; Cuevas, A.; and Guerrero, C. 2014. Understanding the locality effect in Twitter: measurement and analysis. *Personal and Ubiquitous Computing*, 18: 397–411.
- De Choudhury, M.; and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 71–80.
- Dredze, M.; Paul, M. J.; Bergsma, S.; and Tran, H. 2013. Carmen: A twitter geolocation system with applications to public health. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence*.
- Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 1041–1048.
- Fischer, S. 2020. Local news availability does not increase pro-social pandemic response.
- Frias-Martinez, V.; Soto, V.; Hohwald, H.; and Frias-Martinez, E. 2012. Characterizing urban landscapes using geolocated tweets. In *2012 International conference on privacy, security, risk and trust and 2012 international conference on social computing*, 239–248. IEEE.
- Ghosh, R.; and Lerman, K. 2010. Predicting influential users in online social networks. *arXiv preprint arXiv:1005.4882*.
- Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 237–246.
- Hodges Jr, J. 1958. The significance probability of the Smirnov two-sample test. *Arkiv för matematik*, 3(5): 469–486.
- Hong, L.; and Frias-Martinez, V. 2020. Modeling and predicting evacuation flows during hurricane Irma. *EPJ Data Science*, 9(1): 29.
- Izbicki, M.; Papalexakis, V.; and Tsotras, V. 2019. Geolocating Tweets in any Language at any Location. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 89–98.
- Jiang, S.; and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–23.
- Jordan, S. E.; Hovet, S. E.; Fung, I. C.-H.; Liang, H.; Fu, K.-W.; and Tse, Z. T. H. 2018. Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*, 4(1): 6.
- Karami, A.; Zhu, M.; Goldschmidt, B.; Boyajieff, H. R.; and Najafabadi, M. M. 2021. COVID-19 vaccine and social media in the US: Exploring emotions and discussions on Twitter. *Vaccines*, 9(10): 1059.
- Laniado, D.; Volkovich, Y.; Scellato, S.; Mascolo, C.; and Kaltenbrunner, A. 2018. The impact of geographic distance on online social interactions. *Information Systems Frontiers*, 20(6): 1203–1218.
- Lerman, K.; Feldman, D.; He, Z.; and Rao, A. 2024. Affective Polarization and Dynamics of Information Spread in Online Networks. *npj Complexity*, 1(1): 8.
- Leskovec, J.; and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, 915–924.
- Li, Z.; Calvo-Bartolomé, L.; Hoyle, A. M.; Xu, P.; Stephens, D. K.; Fung, J. F.; Dima, A.; and Boyd-Graber, J. L. 2025. Large language models struggle to describe the haystack without human help: A social science-inspired evaluation of topic models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7583–7604.
- Li, Z.; Mao, A.; Stephens, D.; Goel, P.; Walpole, E.; Dima, A.; Fung, J.; and Boyd-Graber, J. 2024. Improving the TENOR of Labeling: Re-evaluating Topic Models for Content Analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 840–859.
- Lindstrom, M. J.; and Bates, D. M. 1988. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404): 1014–1022.
- Loureiro, D.; Barbieri, F.; Neves, L.; Espinosa Anke, L.; and Camacho-collados, J. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260. Dublin, Ireland: Association for Computational Linguistics.
- Lwin, M. O.; Lu, J.; Sheldenkar, A.; Schulz, P. J.; Shin, W.; Gupta, R.; and Yang, Y. 2020. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR public health and surveillance*, 6(2): e19447.
- Masis, T.; and O’Connor, B. 2024. Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)*, 86–98.
- Miles, A.; Andiappan, M.; Upenieks, L.; and Orfanidis, C. 2022. Using prosocial behavior to safeguard mental health

- and foster emotional well-being during the COVID-19 pandemic: A registered report of a randomized trial. *PLoS one*, 17(7): e0272152.
- Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.
- Movva, R.; Peng, K.; Garg, N.; Kleinberg, J.; and Pierson, E. 2025. Sparse Autoencoders for Hypothesis Generation. In *Forty-second International Conference on Machine Learning*.
- Nettasinghe, B.; Rao, A.; Jiang, B.; Percus, A.; and Lerman, K. 2025. In-Group Love, Out-Group Hate: A Framework to Measure Affective Polarization via Contentious Online Discussions. In *Proceedings of the ACM on Web Conference 2025*, 560–575.
- Pavalanathan, U.; and Eisenstein, J. 2015. Confounds and Consequences in Geotagged Twitter Data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2138–2148.
- Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015.
- Rao, A.; Guo, S.; Wang, S. Y. N.; Morstatter, F.; and Lerman, K. 2023. Pandemic Culture Wars: Partisan Differences in the Moral Language of COVID-19 Discussions. In *2023 IEEE International Conference on Big Data (BigData)*, 413–422. IEEE.
- Rao, A.; Morstatter, F.; Hu, M.; Chen, E.; Burghardt, K.; Ferrara, E.; and Lerman, K. 2021. Political Partisanship and Antiscience Attitudes in Online Discussions about COVID-19: Twitter Content Analysis. *Journal of medical Internet research*, 23(6): e26692.
- Rao, A.; Sabri, N.; Guo, S.; Raschid, L.; and Lerman, K. 2025. Public Health Messaging on Twitter During the COVID-19 Pandemic: Observational Study. *Journal of Medical Internet Research*, 27: e63910.
- Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 502–518.
- Rzeszewski, M.; and Beluch, L. 2017. Spatial characteristics of twitter users—Toward the understanding of geosocial media production. *ISPRS International Journal of Geo-Information*, 6(8): 236.
- Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS one*, 8(9): e73791.
- Seabold, S.; Perktold, J.; et al. 2010. Statsmodels: econometric and statistical modeling with python. *SciPy*, 7(1): 92–96.
- Wang, Y.-C.; Burke, M.; and Kraut, R. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 74–85.
- Wood-Doughty, Z.; Xu, P.; Liu, X.; and Dredze, M. 2021. Using Noisy Self-Reports to Predict Twitter User Demographics. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 123–137.
- Xu, P.; Broniatowski, D. A.; and Dredze, M. 2024. Twitter social mobility data reveal demographic variations in social distancing practices during the COVID-19 pandemic. *Scientific reports*, 14(1): 1165.
- Xu, P.; Dredze, M.; and Broniatowski, D. A. 2020. The twitter social mobility index: Measuring social distancing practices with geolocated tweets. *Journal of medical Internet research*, 22(12): e21499.
- Zhang, J.; DeLucia, A.; and Dredze, M. 2022. Changes in tweet geolocation over time: A study with carmen 2.0. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 1–14.
- Zhang, J.; DeLucia, A.; Zhang, C.; and Dredze, M. 2023. Geo-seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4778–4794.
- Zhong, R.; Snell, C.; Klein, D.; and Steinhardt, J. 2022. Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, 27099–27116. PMLR.
- Zhong, R.; Wang, H.; Klein, D.; and Steinhardt, J. 2024. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37: 79350–79380.
- Zhong, R.; Zhang, P.; Li, S.; Ahn, J.; Klein, D.; and Steinhardt, J. 2023. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36: 40204–40237.
- Zhou, Y.; Xu, P.; Wang, X.; Lu, X.; Gao, G.; and Ai, W. 2025. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2302–2316.
- Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 1–55.

Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, in Section 6.**
 - (e) Did you describe the limitations of your work? **Yes, in Section 6.**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes, in Section 6.**
 - (g) Did you discuss any potential misuse of your work? **Yes, in Section 6.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes, we have some repeating hypotheses testing with a similar setup. We clearly state the assumption for the first test and briefer statements for the rest for readability.**
 - (b) Have you provided justifications for all theoretical results? **Yes**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes**
 - (e) Did you address potential biases or limitations in your theoretical framework? **Yes, in Section 6**
 - (f) Have you related your theoretical results to the existing literature in social science? **Yes**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, in Section 6**
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **Yes**
 - (b) Did you mention the license of the assets? **NA**
 - (c) Did you include any new assets in the supplemental material or as a URL? **No**
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **No**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **No**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **NA**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
 - (d) Did you discuss how data is stored, shared, and de-identified? **NA**

A Seed PHEs

The usernames for the 30 seed PHEs on Twitter are: EricTopol, PeterHotez, ashishkjha, trvr, EpiEllie, JuliaRaifman, devisridhar, meganranney, luckytran, asosin, DrLeanaWen, dremilyportermd, DrJaimeFriedman, davidwdowdy, BhramarBioStat, geochurch, DrEricDing, michaelmina_lab, Bob.Wachter, JenniferNuzzo, mtosterholm, MonicaGandhi9, cmyeaton, nataliexdean, angie_rasmussen, ProfEmilyOster, mlipsitch, drlucymcbride, ScottGottliebMD, CDCDirector, and Surgeon_Genera

B Supplement Results for Section 4.2

As a robustness check for the Poisson regression presented in Section 4.2, we fit two advanced models to account for author-level heterogeneity. Both use the same structure as the original model, with each original tweet corresponding to two data points, one for the *gcl* audience and another for *non-gcl*. Both models confirm that the positive association between *gcl* and engagement is statistically significant and robust. The results are summarized in Table 11.

First, we fit a **mixed-effects** Poisson regression. This model includes a random intercept for each PHE author, treating author-level effects as random variables. We define the model as:

$$\log(\mathbb{E}[Y_{ij}]) = \beta_0 + \beta_1 X_{ij} + u_j + \log(W_{ij})$$

where Y_{ij} is the reply count for tweet i by author j for a specific audience type, X_{ij} is a binary indicator for that audience type (*gcl* vs. *non-gcl*), u_j is the random intercept for author j , and W_{ij} is the corresponding audience size included as an exposure variable.

As a supplementary check, we also fit a Poisson regression with author **fixed effects**. This model accounts for author heterogeneity by including a separate fixed intercept for each author. The model is specified as:

$$\log(\mathbb{E}[Y_{ij}]) = \alpha_j + \beta_1 X_{ij} + \log(W_{ij})$$

where α_j is the unique fixed-effect intercept for author j . This model requires sufficient observations per author to reliably estimate each fixed effect. To ensure model convergence, this analysis was restricted to authors with at least three observations.

Model	Coefficient	IRR	95% CI (IRR)
Mixed-Effects	-0.122	0.885***	(0.870, 0.900)
Fixed-Effects	-0.105	0.901***	(0.883, 0.919)

Table 11: Robustness Checks for Poisson Regression. X_T ($gcl = 0$, $non-gcl = 1$). IRR is the incidence rate ratio. $p < 0.001$ is marked with ***.

C COVID-19 Issues

The lockdown issue comprises content pertaining to early state and federal mitigation efforts, such as quarantines, stay-at-home orders, business closures, reopening, and calls

for social distancing. The masking issue is defined by discussions on the use of face coverings, mask mandates, mask shortages, and anti-mask sentiment. Education-related content involves tweets about school closures, reopening of educational institutions, homeschooling, and online learning during the pandemic. The vaccine issue pertains to discussions about COVID-19 vaccines, vaccine mandates, anti-vaccine sentiment, and vaccine hesitancy in the U.S.

Issue	Sample Tweets
Lockdowns	This is a GREAT idea. We're all in this together. Take care of each other. #Stay-Home #TakeItSeriously #FlattenTheCurve #COVID19
Masking	We're in the middle of a pandemic and y'all are still coughing and sneezing without covering your mouths? Come on now.
Education	More glimmers of hope as we "safely" move forward and open up Texas AM University while containing #COVID19 .
Vaccines	You are joking right? Zero sympathy for antivaxxers who quit their jobs rather than get vaccinated . They put us all at risk and make the pandemic prolonged for the world.

Table 12: Sample Tweets for each COVID-19 issue. Issue-relevant keywords are bolded. Table from Rao et al. (2023).

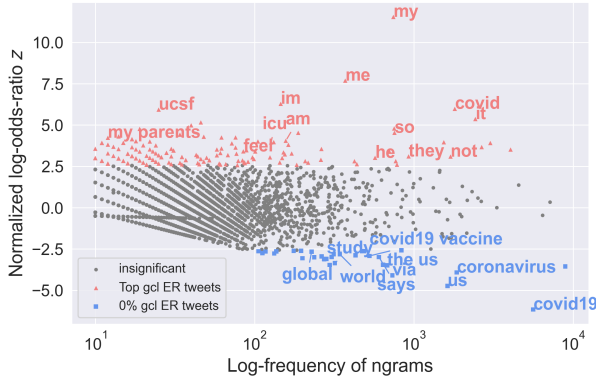
D Lexical Analysis for Geo-co-located Tweets

To gain insights into what lexical features are most distinctive for *gcl* tweets, we compare Tweets that have 0% *gcl* engagement rate ($N = 10,854$, bottom 68% of the tweets) with the top 10% percentile of tweets (3.4% *gcl* engagement rate with $N = 1,576$). We also compare 0% *gcl* ER tweets with the top 20% percentile of tweets (1.0% *gcl* engagement rate with $N = 3,190$).

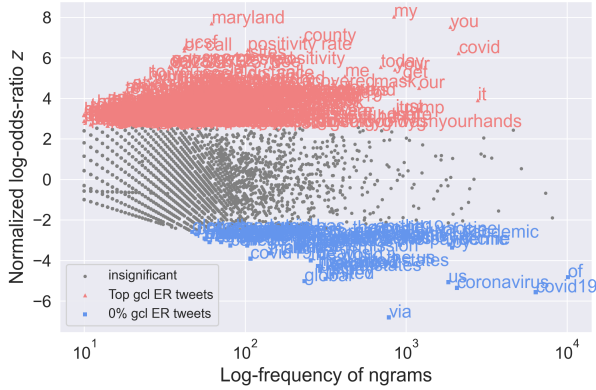
We follow Monroe, Colaresi, and Quinn (2008) and use normalized log-odds-ratio z to find out the n -grams ($n \in \{1, 2, 3\}$) more associated with Tweets with higher and lower *gcl* engagement rates. We find that the top group shares more personal experiences/feelings ("my", "me"), while the bottom group talks about the national or global situation of COVID-19 with keywords such as "covid19", "us", "united states", "study", "the world" (Figure 3).

E Supplemental Results for Section 5.1

Sentiment To investigate sentiment expressed in the *gcl* and *non-gcl* replies, we use a RoBERTa-base model (Loureiro et al. 2022) trained on 124M tweets from January 2018 to December 2021, and fine-tuned for sentiment analysis with the TweetEval benchmark (Rosenthal, Farra, and Nakov 2017). We classify each reply as positive, neutral, or negative. For each original Tweet T , we compute the positive rate for its *gcl* replies and *non-gcl* replies, respectively. We find that *gcl* replies are slightly more positive and less negative than *non-gcl* replies. The mean percentage of *gcl*



(a) Top 10% percentile of tweets



(b) Top 20% percentile of tweets

Figure 3: Most distinctive n-grams ($n \in \{1, 2, 3\}$): tweets with 0% *gcl* engagement rate (bottom 68% of the tweets) vs. (a) the top 10% percentile of tweets (3.4% *gcl* engagement rate) and (b) the top 20% percentile of tweets (1.0% *gcl* engagement rate). Only n-grams with 99.5% confidence level are colored ($z > 2.57$)

replies labeled as positive is 14% and the standard deviation is 29%, while the mean percentage for *non-gcl* replies labeled as positive is $13\% \pm 22\%$. The difference between the two distributions was statistically significant (KS test with $p < 0.001$ and $n = 3, 521$). The mean percentage of *gcl* and *non-gcl* replies labeled as negative were also significantly different ($p < 0.001$) with mean and standard deviation values of $43\% \pm 39\%$ and $45\% \pm 29\%$, respectively.

Regression Results for the Impact of COVID-19 Topics

We define the linear regression model for directly evaluating the impact of COVID-19 topics on engagement rates as follows:

$$\text{Engagement Rate} = \beta_0 + \sum_{T_i \in \text{Topics}} (\beta_i T_i) + \beta_c F + \epsilon,$$

where the dependent variable is one of three engagement rates (*gcl*, *non-gcl*, or *Diff*). The independent variables are binary indicators for each topic T_i (masking, lockdowns, education, and vaccines), and F is a control for the author's

	<i>Diff</i>	<i>gcl</i>	<i>non-gcl</i>
const	0.002** (0.001)	0.020*** (0.001)	0.018*** (0.001)
masking_True	0.003 (0.003)	0.004 (0.003)	0.001 (0.002)
education_True	-0.001 (0.003)	-0.000 (0.003)	0.000 (0.002)
lockdowns_True	-0.003 (0.005)	-0.003 (0.005)	0.000 (0.003)
vaccines_True	-0.004** (0.002)	-0.004** (0.002)	-0.001 (0.001)
# followers	-0.004 (0.008)	-0.031*** (0.007)	-0.027*** (0.006)
R-squared	0.000	0.002	0.002
R-squared Adj.	0.000	0.001	0.001

Table 13: Regression results for evaluating the impact of COVID-19 Topics. Standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

min-max normalized follower count to account for their general influence. The regression results are shown in Table 13.

Complete Results for the LIWC regression The complete results of the LIWC regression analysis (a complete version of Table 8) in Section 5.1 is shown in Table 14.

Results for a Mixed-effect Model The results for the mixed-effect model are shown in Table 15. We also attempted to fit a model with random slopes. However, the model failed to converge despite multiple attempts using different optimizers.

F LLM-based Content Analysis

We adapted the D5 framework proposed by Zhong et al. (2022, 2023), using the provided implementation.⁹ In our setup, we used GPT4o-2024-08-06 as the proposer to generate hypotheses, and a provided 3B parameter version of the T5 model, as the validator to assess each hypothesis.

For each comparison, we randomly sampled 1000 tweets from each corpus to serve as the exploration (training) set used by the proposer, and another 1000 tweets per corpus as the validation set used by the validator to assign a validation score, denoted as V' . During the generation phase, the proposer randomly selects 25 tweets from each corpus in each round and continues this process until $k = 20$ hypotheses are produced.

We applied this method across multiple experimental setups, varying: (1) The percentile thresholds used to define the corpora: 20% and 10%; (2) The engagement metrics used for partitioning: *Diff*, *gcl*, and *non-gcl*; (3) The level of contextual information included in the prompts.

By default, we used a minimal-context prompt shown in Table 16, which only describes the task and not the source or nature of the tweets. The original D5 setup uses a more detailed prompt (Table 17) that includes explicit information about the corpora. However, we observed that when context

⁹<https://github.com/ruiqi-zhong/D5>

	<i>Diff</i>	<i>gcl</i>	<i>non-gcl</i>
const	0.001 (0.001)	0.017*** (0.001)	0.016*** (0.001)
i	0.103*** (0.038)	0.448*** (0.037)	0.345*** (0.028)
we	-0.014 (0.030)	-0.036 (0.029)	-0.022 (0.022)
you	-0.022 (0.039)	0.017 (0.038)	0.039 (0.029)
posemo	0.038 (0.024)	0.092*** (0.023)	0.054*** (0.018)
negemo	-0.070 (0.046)	0.008 (0.045)	0.079** (0.034)
anx	0.128* (0.069)	-0.053 (0.067)	-0.181*** (0.051)
anger	0.106 (0.073)	0.037 (0.071)	-0.069 (0.054)
sad	0.057 (0.074)	-0.018 (0.072)	-0.075 (0.054)
family	0.105 (0.086)	0.248*** (0.084)	0.143** (0.063)
friend	-0.022 (0.093)	0.039 (0.090)	0.061 (0.069)
body	-0.066 (0.055)	-0.066 (0.053)	0.000 (0.040)
health	-0.017 (0.026)	-0.014 (0.025)	0.002 (0.019)
work	0.007 (0.018)	-0.011 (0.018)	-0.018 (0.014)
leisure	-0.059 (0.039)	-0.028 (0.038)	0.031 (0.029)
home	0.019 (0.060)	0.003 (0.058)	-0.016 (0.044)
money	-0.035 (0.042)	-0.050 (0.040)	-0.015 (0.031)
relig	-0.008 (0.081)	-0.079 (0.079)	-0.071 (0.060)
death	0.018 (0.041)	-0.055 (0.040)	-0.073** (0.030)
swear	-0.042 (0.157)	-0.070 (0.152)	-0.028 (0.115)
norm_followers_count	-0.003 (0.008)	-0.028*** (0.007)	-0.024*** (0.006)
masking_True	0.004 (0.003)	0.005 (0.003)	0.001 (0.002)
education_True	-0.001 (0.003)	-0.001 (0.003)	0.000 (0.002)
lockdowns_True	-0.003 (0.005)	-0.003 (0.005)	-0.000 (0.003)
vaccines_True	-0.004** (0.002)	-0.005*** (0.002)	-0.001 (0.001)
R-squared	0.002	0.014	0.015
R-squared Adj.	0.000	0.013	0.014

Table 14: Full LIWC regression results of LIWC categories' impact on different types of engagement rates. Standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

	<i>Diff</i>	<i>gcl</i>	<i>non-gcl</i>
Intercept	0.001 (0.001)	0.068*** (0.006)	0.068*** (0.006)
i	0.103*** (0.038)	0.313*** (0.036)	0.201*** (0.026)
we	-0.014 (0.030)	-0.017 (0.028)	-0.007 (0.020)
you	-0.022 (0.039)	0.010 (0.037)	0.028 (0.027)
posemo	0.038 (0.024)	0.037* (0.022)	0.004 (0.016)
negemo	-0.070 (0.046)	-0.011 (0.042)	0.053* (0.030)
anx	0.128* (0.069)	-0.038 (0.063)	-0.152*** (0.046)
anger	0.106 (0.073)	0.049 (0.067)	-0.047 (0.048)
sad	0.057 (0.074)	-0.017 (0.067)	-0.072 (0.049)
family	0.105 (0.086)	0.202** (0.079)	0.093 (0.057)
friend	-0.022 (0.093)	0.022 (0.085)	0.025 (0.062)
body	-0.066 (0.055)	-0.042 (0.050)	0.027 (0.036)
health	-0.017 (0.026)	-0.024 (0.024)	0.001 (0.018)
work	0.007 (0.018)	-0.005 (0.017)	-0.010 (0.012)
leisure	-0.059 (0.039)	-0.043 (0.036)	0.021 (0.026)
home	0.019 (0.060)	-0.014 (0.055)	-0.031 (0.039)
money	-0.035 (0.042)	-0.045 (0.039)	-0.014 (0.028)
relig	-0.008 (0.081)	-0.053 (0.074)	-0.057 (0.054)
death	0.018 (0.041)	0.010 (0.038)	-0.016 (0.028)
swear	-0.042 (0.157)	-0.106 (0.144)	-0.082 (0.104)
norm_followers_count	-0.003 (0.008)	-0.145** (0.070)	-0.146* (0.078)
masking_True	0.004 (0.003)	0.005** (0.003)	0.001 (0.002)
education_True	-0.001 (0.003)	-0.002 (0.003)	-0.001 (0.002)
lockdowns_True	-0.003 (0.005)	0.001 (0.004)	0.004 (0.003)
vaccines_True	-0.004** (0.002)	-0.005** (0.002)	-0.000 (0.001)
Group Var	0.000 (nan)	1.169*** (0.140)	2.803*** (0.302)

Table 15: Mixed-effect regression results of LIWC categories' impact on different types of engagement rates. The group variable is the authors. Standard errors in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$

about author or participant location is included—especially without being visible in the tweet text—the LLM tends to hallucinate references to “locality”, as seen in the generated hypotheses from Table 23.

We also tested a prompt that encourages the model to focus on linguistic and psychological patterns, broadly aligned with LIWC categories (prompt shown in Table 18 ; results in Table 25). While this prompt adjustment resulted in slightly different hypotheses, the overall content and distributional patterns remained similar to the default prompt (Table 20).

Similarly, we found only minor differences between results generated using 20% vs. 10% percentile thresholds. (Table 20 and Table 24) Across both setups, dominant hypotheses continued to emphasize the sharing of personal experience and emotional expression. Finally, the comparison between *gcl*-based and *non-gcl*-based groupings (Tables 21 and 22) yields consistent themes—namely, the salience of emotional expression and first-person narratives—echoing the patterns observed in our regression analysis (Table 8). These features are significant predictors across both *gcl* and *non-gcl* setups, and remain robust even when analyzing the *Diff* metric, which controls for shared effects across both dimensions.

Overall, the LLM-based content analysis complements our statistical findings by surfacing interpretable hypotheses that support the observed role of personal expression and affective language in shaping engagement.

Prompt used for hypothesis generation without *gcl* context

Group A: {sampled tweets from Corpus A}
Group B: {sampled tweets from Corpus B}

The dataset includes two sets of tweets with different engagement metrics. The two groups are generated based on different tweet engagement metrics. Group A contains tweets in Group A, while Group B contains tweets in Group B.

I am a public health communication researcher. My goal is to figure out what kind of tweets are more frequent in Group A compared to Group B.

Please write a list of hypotheses about the datapoints from Group A (listed using bullet points "-"). Each hypothesis should be formatted as a sentence fragment.
Based on the two sentence groups (A and B) above, more sentences in Group A ...

Table 16: Prompt used for hypothesis generation

G Supplement Results for Section 5.2

Table 9 shows the logistic regression results for comparing *gcl* vs. *non-gcl* replies. The heatmaps with all LIWC categories from Section 5.2 are shown in Figure 4.

Prompt used for hypothesis generation with *gcl* context

Group A: {sampled tweets from Corpus A}
Group B: {sampled tweets from Corpus B}

The dataset includes tweets with various geo-co-located and non-geo-co-located engagement rates. The two groups are generated based on the difference between geo-co-located and non-geo-co-located engagement. Group A contains tweets with higher geo-co-located engagement rate but lower non-geo-co-located engagement rate, while Group B contains tweets with higher non-geo-co-located engagement rate but lower geo-co-located engagement rate.

I am a public health communication researcher. My goal is to figure out what kind of tweets are more frequent in tweets with higher geo-co-located engagement rate but lower non-geo-co-located engagement rate.

Please write a list of hypotheses about the datapoints from Group A (listed using bullet points "-"). Each hypothesis should be formatted as a sentence fragment.
Based on the two sentence groups (A and B) above, more sentences in Group A ...

Table 17: Prompt used for hypothesis generation **with** *gcl* context

Prompt encouraging hypotheses aligned with LIWC categories

Group A: {sampled tweets from Corpus A}
Group B: {sampled tweets from Corpus B}

The dataset includes two sets of tweets with different engagement metrics. The two groups are generated based on different tweet engagement metrics. The Group A snippets tweets in Group A, while the Group B snippets tweets in Group B.

I am a public health communication researcher. My goal is to figure out what features related to **linguistic patterns, topics of discussion, and psychological processes** are more frequent in Group A compared to Group B.

Please write a list of hypotheses about the datapoints from Group A (listed using bullet points "-"). Each hypothesis should be formatted as a sentence fragment.
Based on the two sentence groups (A and B) above, more sentences in Group A ...

Table 18: Prompt encouraging hypotheses aligned with LIWC categories

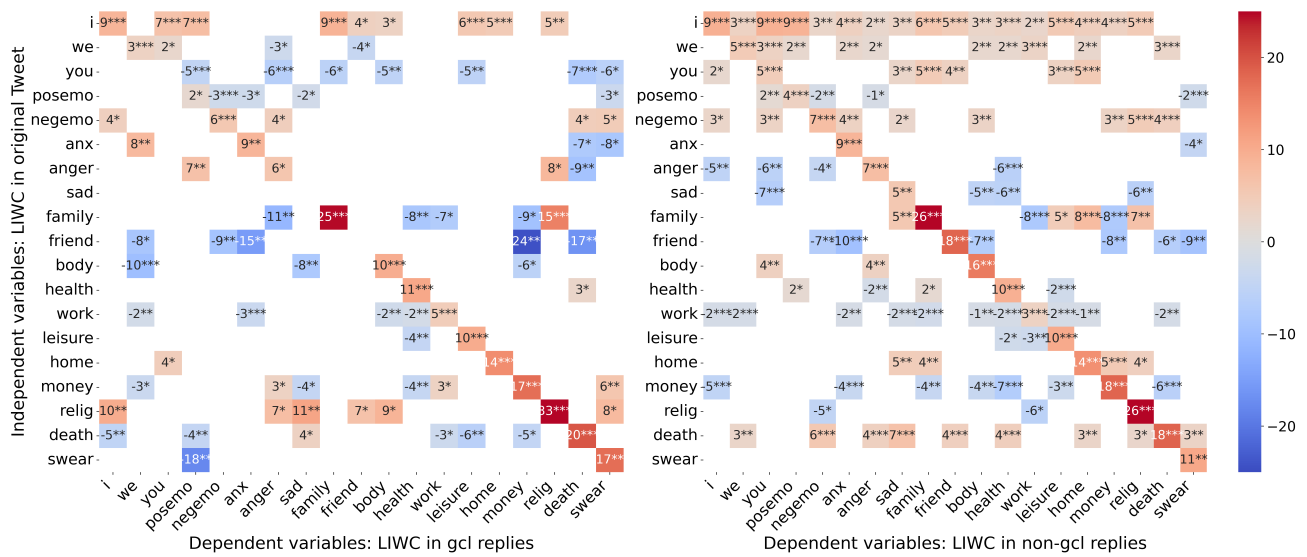


Figure 4: How *gcl* replies and *non-gcl* with specific LIWC categories correlate with LIWC features in the original Tweet. Each column is a logistic regression. Each cell is the coefficient. $p < 0.1$ is marked with *, $p < 0.05$ with **, and $p < 0.01$ with ***.

Prompt for profession classification task

For the following Twitter user profile description and username, infer whether the user's profession is related to each of the professions as follows: media professions (e.g., journalist, reporter, writer, editor, and author), academic (e.g., professor, researcher, and scientist), health professions (e.g., physician, doctor, epidemiologist), and politicians. For each Twitter user, output the results for each of the four professions and provide a brief explanation in the JSON format of:

```
{ "screen_name": USER_NAME,
  "Media_Professions": true /
false, "Academic": true / false,
  "Health_Professions": true / false,
  "Politicians": true / false,
  "Explanation": "[Explanation]" }
```

Only output true for the best-matched professions if possible. All outputs should be in the JSON format as described above.

```
username: {screen_name}
description: {description}
```

Table 19: Prompt used for classifying professions of Twitter users

H Prompt for Profession Classification

The prompt we used for profession classification in Section 5.3 is shown in Table 19. We instructed the LLM to infer whether each user's profile suggested affiliation with one or more of four target professions: media, academic, health, or political. The model was asked to output a binary judgment for each profession, along with a brief explanation, using a structured JSON format.

Note that {screen_name} and {description} are replaced with actual usernames and descriptions. A sample output from ChatGPT with an anonymized user name is:

```
{ "screen_name": "ANONYMIZED",
  "Media_Professions": false, "Academic":
true, "Health_Professions": true,
  "Politicians": false, "Explanation":
"The user is a Clinical Psychology PhD
Student and does research in Global
Mental Health, which indicates that
he is involved in academic and health
professions." }
```

Hypothesis	V'
include expressions of personal emotions or sentiments, such as fear, pride, or frustration	0.081
focus on personal experiences and anecdotes related to the pandemic	0.070
- discuss personal anecdotes and experiences related to COVID-19	0.063
discuss everyday life adjustments and challenges during the pandemic, like dating or travel	0.061
incorporate personal milestones or special occasions and how they've been affected by the pandemic	0.055
reference specific locations or cities, such as Scottsdale, AZ or Ohio	0.043
mention local government actions and leadership, especially in specific states or cities	0.039
describe impacts of COVID-19 measures on daily life and personal routines	0.037
contain direct calls to action or emphasize the importance of community actions (e.g., wearing masks or supporting essential workers)	0.024
address COVID-19 misinformation and actions taken by social media platforms	0.020
mention specific public health measures like mask mandates	0.015
focus on COVID-19 vaccine distribution and individual vaccination experiences	0.008
talk about the reopening or closure of specific facilities, such as nursing homes	0.005
highlight stories of specific individuals or public figures contracting COVID-19	0.000
report specific COVID-19 statistics or data related to local settings like neighborhoods or cities	-0.004
discuss innovations or changes in healthcare or public health practices due to COVID-19	-0.008
highlight the impact of COVID-19 on individual or specific communities (e.g., hospitals being full, grocery store experiences)	-0.025
mention specific public figures or local leaders in the context of COVID-19 response	-0.032
mention trends or statistics related to COVID-19 cases or immunity	-0.061
include updates or opinions about scientific research and data	-0.120

Table 20: LLM-generated distributional difference without *gcl* context: tweets bottom 20% percentile of the *Diff* values vs. the top 20% percentile of tweets. Full results for Table 7.

Hypothesis	V'
include humorous or light-hearted content related to coping with the pandemic	0.147
incorporate humor or light-hearted personal commentary about adjusting to pandemic life	0.146
- express personal experiences or anecdotes related to the COVID-19 pandemic	0.135
include casual language, emojis, or express personal emotions	0.097
reflect on the emotional or psychological impacts of COVID-19	0.071
.....	
mention specific locations or regions affected by the pandemic	-0.054

Table 21: LLM-generated distributional difference without *gcl* context: tweets with 0% *gcl* engagement rate vs. the top 20% percentile of tweets.

Hypothesis	V'
emphasize personal or anecdotal perspectives on pandemic-related issues	0.164
- focus on personal opinions and reflections related to COVID-19 experiences	0.128
express personal frustrations or emotions related to COVID-19, including feelings of worry or anger	0.123
engage in casual or conversational tone, often posing questions to the audience	0.092
include direct responses or criticisms toward specific political figures, particularly Donald Trump	0.070
.....	
draw comparisons with other countries or states to emphasize pandemic management or outcomes	-0.068
reflect the state-level impact of COVID-19, such as hospital capacities in Wisconsin or actions by California	-0.093
share insights on localized public health measures, like vaccination appointments in specific areas	-0.103
include hyperlinks or references to full articles or studies, possibly indicating a focus on data and evidence	-0.114

Table 22: LLM-generated distributional difference without *gcl* context: tweets with 0% *non-gcl* engagement rate vs. the top 20% percentile of tweets.

Hypothesis	V'
feature personal stories and specific events	0.085
cover anecdotes about personal interactions or encounters related to COVID-19	0.073
discuss the implications of COVID-19 on local healthcare systems or hospitals	0.043
report on local vaccination efforts or challenges	0.036
.....	
focus on domestic political issues and figures, such as the Trump administration and state-level responses	-0.050
focus on regional COVID-19 statistics and updates	-0.052

Table 23: LLM-generated distributional differences **with** *gcl* context: tweets with bottom 20% percentile *Diff* values vs. top 20% percentile of tweets. Hallucinations about "locality" are bolded

Hypothesis	V'
share light-hearted or personal sentiments about adapting to the pandemic	0.072
use humor or cultural references in relation to the pandemic	0.063
.....	
address misinformation or controversies related to COVID-19 testing, vaccines, or treatments	-0.052
report on statistics, forecasts, or models related to COVID-19 cases and mortality rates	-0.065
mention high-profile public figures in the context of pandemic-related policies	-0.094
focus on COVID-19 vaccination efforts or challenges	-0.112
discuss the ongoing effects and responses to COVID-19 in specific geographic regions or countries	-0.140

Table 24: LLM-generated distributional differences without *gcl* context: tweets with bottom 10% percentile *Diff* values vs. top 10% percentile of tweets.

Hypothesis	V'
use a conversational or colloquial tone	0.082
capture personal or anecdotal experiences	0.075
.....	

Table 25: LLM-generated distributional difference without *gcl* context, focusing on linguistic patterns, topics, and psychological processes: tweets with bottom 20% percentile *Diff* values vs. top 20% percentile of tweets.

Hypothesis	V'
convey frustration with the impact of legislation or policies on workers or specific groups	0.040
include a call to action or advocacy, such as petitions or protests	0.029
critique or express dissatisfaction with the federal government's response to the pandemic	0.027
express support or admiration for local leaders or officials	0.014
reference specific political events, policies, or government figures, often critically	0.013
discuss American politics or criticize political figures	0.011
mention COVID-19 in the context of state or federal government response or criticism	0.010
engage in direct confrontation or argument with other users	0.010
express gratitude or appreciation for someone's efforts or leadership	0.009
express frustration or emotional reactions, often in a casual or informal tone	0.008
reference data or statistics related to COVID-19 cases and mortality rates	0.008
debate the effectiveness of public health guidelines or actions	0.007
inquire about or critique the implementation of healthcare measures or policies	0.005
exhibit colloquial language or informal expressions	0.005
discuss the importance or implications of COVID-19 precautions, such as mask-wearing	0.001
mention political figures in the context of their response to the pandemic	-0.001
reflect concern for the long-term effects of COVID-19 or other health-related issues	-0.001
use sarcasm or mocking language	-0.005
contain personal insults or hostile language	-0.007

Table 26: LLM-generated distributional difference without *gcl* context: *gcl* replies vs. *non-gcl* replies.