Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach

Anonymous ACL submission

Abstract

Susceptibility to misinformation describes the degree of belief in unverifiable claims, a latent aspect of individuals' mental processes that is not observable. Existing susceptibility studies heavily rely on self-reported beliefs, which can be subject to bias, expensive to collect, and challenging to scale for downstream applications. To address these limitations, in this work, we propose a computational approach to efficiently model users' latent susceptibility levels. As shown in previous work, susceptibility is influenced by various factors (e.g., demographic factors, political ideology), and directly influences people's reposting behavior on social media. To represent the underlying mental process, our susceptibility modeling incorporates these factors as inputs, guided by the supervision of people's sharing behavior. Using COVID-19 as a testbed, our experiments demonstrate a significant alignment between the susceptibility scores estimated by our computational modeling and human judgments, confirming the effectiveness of this latent modeling approach. Furthermore, we apply our model to annotate susceptibility scores on a large-scale dataset and analyze the relationships between susceptibility with various factors. Our analysis reveals that political leanings, etc., psychological factors exhibit varying degrees of association with susceptibility to COVID-19 misinformation, and shows that susceptibility is unevenly distributed across different professional and geographical backgrounds.[1]

1 Introduction

False claims spread on social media platforms, such as conspiracy theories, fake news, and unreliable health information. They mislead people's judgment, promote societal polarization, and exacerbate distrust in government (Pennycook and Rand, 2021; Nan et al., 2020). The harm is especially significant in various contentious events, including elections, religious persecution, and the global response to the COVID-19 pandemic (Ecker et al., 2022). Many works have investigated the *observable* behavior of misinformation propagation such as where the information propagates (Taylor et al., 2023), how people share it (Yang et al., 2020), and what people discuss about it (Gupta et al., 2022). However, it is still crucial but challenging to understand the *unobservable* mental and cognitive processes of how individuals believe misinformation (Ecker et al., 2022). Individual susceptibility (i.e., the likelihood of believing and being influenced by misinformation) plays a pivotal role in this context. If one is more susceptible to misinformation, they are not only more likely to share but also more prone to being misled by them (Scherer et al., 2020).

Previous works have investigated the psychological, demographic, and other factors that may contribute to the high susceptibility of an individual (Brashier and Schacter, 2020; Pennycook and Rand, 2017). However, these studies heavily rely on self-reported belief towards false claims collected from questionnaire-based participant surveys (Escolà-Gascón et al., 2021; Rosenzweig et al., 2021), which presents several limitations. For instance, different participants might interpret belief levels differently. Moreover, the data collection is labor-intensive, thereby limiting the scale of downstream research on the size, scope, and diversity of the target population (Nan et al., 2022).

People's mental processes, which are unobservable and influenced by various factors, directly affect several externalized behaviors, such as reposting on social media (Mitchell et al., 2019; Brady et al., 2020; Islam et al., 2020; Altay et al., 2022). Building on these prior works, we propose a computational method to efficiently model individuals' unobservable susceptibility levels only based on their observable social media posting and shar-

[1]We will release all the code used in our paper, along with our trained model and all collected data.

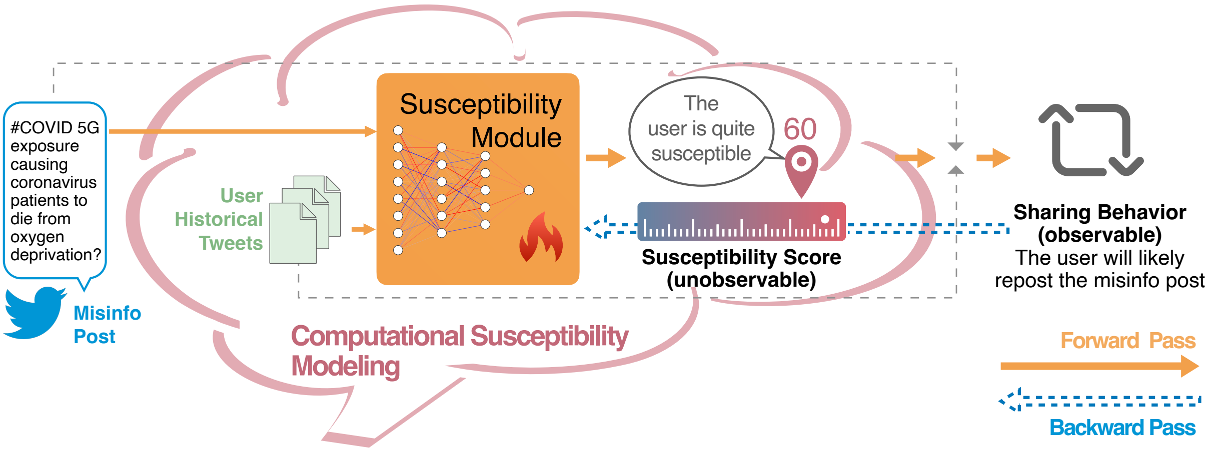

Figure 1: **Computational Modeling of Susceptibility to Misinformation**. We represent user susceptibility as a latent variable, which we capture using a shallow neural network. Our model is trained with the supervision of users' observable sharing behaviors, employing two loss functions: *binary classification entropy* and *triplet loss*.

ing behaviors. We represent users based on their historical posts and perform multi-task learning to simultaneously learn to classify whether a user would share a post, as well as to rank susceptibility scores among similar and dissimilar users when the same content is seen. This computational modeling method unlocks the scales of misinformation-related studies and provides a novel perspective to reveal users' belief patterns.

In this paper, we focus our experiments on COVID-19 misinformation, and our evaluations demonstrate that the estimations from our model are highly aligned with human judgment when assessed through a susceptibility comparison task. The correlation study between estimated and human-annotated susceptibility verifies the effectiveness of the indirect susceptibility modeling method. To further illustrate the significance of our work, we employ our model to annotate susceptibility levels on a large-scale dataset. Building upon this extensive susceptibility labeling, we then conduct a set analysis to examine how various factors relate to susceptibility. Our analysis reveals that psychological factors, professional fields, and political leanings are associated with susceptibility to varying degrees. Notably, this large-scale analysis enabled by our computational susceptibility modeling corroborates the findings of previous studies based on self-reported beliefs, e.g. confirming that stronger analytical thinking is an indicator of lower susceptibility. Moreover, the results of our analysis show the potential to extend findings in the existing literature. For example, we demonstrate that the distribution of COVID-19 misinformation susceptibility in the U.S. exhibits a certain degree of correlation with political leanings.

2 Related Work

Measure of Susceptibility The common practice to measure susceptibility is to collect self-reported absolute or relative agreement or disagreement with (or perceived accuracy, credibility, reliability, or validity of) one or more claims verified to be false from a group of individuals (Roozenbeek et al., 2020; Escolà-Gascón et al., 2021; Rosenzweig et al., 2021; Nan et al., 2022). A small number of previous studies indirectly assess the susceptibility by its impact, however, they can only capture behaviors rather than people's beliefs (Loomba et al., 2021). Cheng et al. (2021) defines a heuristic susceptibility score as the radio of misinformation posts out of all user's posts, which unrealistically simplifies the definition of susceptibility. Instead of using expensive and limited self-reported beliefs, we propose a computational model to estimate susceptibility at scale.

Contributing Factors and Application of Susceptibility Relying on the manually collected susceptibility annotation, previous research investigates the psychological, demographic, and more factors that contribute to users' susceptibility (Bringula et al., 2021; van der Linden, 2022). These factors include emotion (Sharma et al., 2022) (e.g. anger and anxiety; Weeks, 2015), analytic thinking (hui Li et al., 2022), partisan bias (Roozenbeek et al., 2022a), source credibility (Traberg and van der Linden, 2022), and repetition (Foster et al., 2012). Many theories have been proposed about the reason behind suscetibility (Scherer et al., 2020), including limited knowledge acquiring and literacies capabilities (Brashier and Schacter, 2020), strong preexisting beliefs (Lewandowsky and Ecker, 2012), neglecting to sufficiently reflect about the truth (Pennycook and Rand, 2017) or overconfidence (Salovich et al., 2020).A better understanding of the phenomenon and mechanism of susceptibility can facilitate various downstream applications. These include analyzing the spread of bots (Himelein-Wachowiak et al., 2021), revealing community properties in information pathways (Taylor et al., 2023; Ma et al., 2023), combating misinformation by emphasizing publisher (Dias et al., 2020) and prebunking interventions based on inoculation (Roozenbeek et al., 2022b). However, the absence of a computational modeling framework significantly limits the scale of current susceptibility research.

Inferring Unobservables from Observables Latent constructs or variables refer to concepts that are not directly observable or measurable. Many studies have shown that unobservable variables can be inferred indirectly through models based on observable ones (Bollen, 2002; Borsboom et al., 2003). These unobservable variables can be estimated using various modeling techniques, including nonlinear mixed-effects models, hidden Markov models, or latent class models. In our work, we utilize a neural network-based architecture to model people's latent susceptibility level to misinformation, guided by the supervision provided by their observable sharing behaviors on social media.

3 Computational Susceptibility Modeling

Misinformation is characterized as information that is false, inaccurate, or misleading, which could be created deliberately or accidentally (Pennycook and Rand, 2021). The susceptibility to misinformation represents the belief in misinformation and related constructs, including discernment between true and false claims and the extent to which exposure to misinformation misleads subsequent decisions (Nan et al., 2022). Previous research on susceptibility and misinformation mainly relied on self-reported beliefs collected using surveys or questionnaires - they suffered from problems like being subject to bias, expensive to collect, and challenging to reproduce and scale up.

Existing studies indicating that believing a piece of misinformation can influence various outward behaviors, such as sharing actions. For example, previous studies of the inattention or "classical reasoning" account contend that people are committed to sharing accurate information, but the unique context of social media disrupts their capacity to critically assess the accuracy of news (Pennycook and Rand, 2021; van der Linden, 2022). These studies suggest that people are more likely to share things they genuinely believe (Altay et al., 2022). Inspired by this observation, we propose to model user's unobservable susceptibility only based on their historical posting and sharing behaviors, which are the most available and the easiest collectable data from social media (§3.1) as shown in Fig. 1. Therefore, our proposed framework can efficiently infer users' susceptibility levels to misinformation on a large scale, demonstrating the potential to expand the scope of previous misinformation-related research.

Furthermore, because social media users utilize posts to express their personal and inner thoughts, they reveal information about their characteristics through their posts. Therefore, our proposed susceptibility modeling can incorporate users' informative hidden factors, such as personality traits, analytical thinking, and emotion, to infer a user's susceptibility to misinformation. These additional pieces of information are otherwise very difficult to directly collect on social media.

3.1 Modeling Unobservable Susceptibility

Content-Sensitive Susceptibility In our work, we consider the susceptibility of user u when a particular piece of misinformation p is perceived (i.e. $s_{u,p}$). This allows us to account for the fact that an individual's susceptibility can vary across different content, influenced by factors such as topics and linguistic styles. By focusing on the susceptibility to specific pieces of misinformation, we aim to create a more nuanced, fine-grained, and accurate representation of how users interact with and react to different misinformation.

User and Misinfo Post Embeddings We induce user and post embeddings to reflect hidden factors of the user personality traits and content of the post. As a component of the computational model, we use SBERT (Reimers and Gurevych, 2019), which is developed upon RoBERTa-large (Liu et al., 2019), to compute the embedding vector to represent the information contained in the misinformation and user historical posts. We consider the misinformation post as a sentence and produce its representation with SBERT. For the user embedding, we calculate the average of sentence representations for the user's recent original posts. More specifically, for every user-post pair (u, p), we gather the historical posts written by user u within a 10-day window preceding the creation time of the misinformation post p, to learn a representation of user u at that specific time.[2]

Computational Model for Susceptibility Given the input of user historical posts for the user u and the content for misinformation post p, the susceptibility computational model is expected to produce the *susceptibility score* $s_{u,p}$ as shown in Eq. 1, reflecting the susceptibility of u when p is perceived.

$$s_{u,p} = suscep(E(u), E(p)) \tag{1}$$

We first obtain the embeddings $E(p)$ and $E(u)$ for post p and user u, where u is represented by the user's historical tweets and E is the frozen SBERT sentence embedding function. The susceptibility score is calculated by the function $suscep$, which is implemented as a multi-layer neural network, taking the concatenation of the user and post embeddings as inputs. During the training phase, we maintain the sentence embedder as a fixed component and exclusively train the weights for the $suscep$ function. Then the learned $suscep$ function can be applied to generate susceptibility scores for new pairs of users (u) and posts (p) during the inference process.

3.2 Training with Supervision from Observable Behavior

Susceptibility is a latent variable and cannot be directly observed. Consequently, it is impractical to directly apply supervision to $s_{u,p}$ since only the user u themselves know their own beliefs regarding content p. To address this challenge, we regard susceptibility as a crucial factor for sharing behavior and train the susceptibility computational model using the supervision signals obtained from the observable behavior of sharing misinformation.

To determine the probability of user u sharing post p, we compute the dot product of the embeddings of the user and post content, incorporating the susceptibility score for the same pair of u and p estimated by our model as a weighting factor, and pass the resulting value through a sigmoid function, as illustrated in (2).

[2]We chose the 10-day timeframe because it provides a substantial amount of data to represent a user and is also recent enough to capture their dynamics.

$$p_{\text{rp}} = \sigma\left(E(u) \cdot E(p) \cdot s_{u,p}\right) \tag{2}$$

It is important to highlight that we do not directly utilize the *susceptibility score* to estimate sharing probability because sharing behavior depends not solely on susceptibility levels but also on various potential confounding factors. For instance, it is possible that a user may possess a significantly high susceptibility score for a piece of misinformation but decides not to share it, potentially influenced by factors such as their personality, the impact of social influence, concerns about potential repercussions, and their emotional state at that specific moment, among other variables. To account for these potential confounding factors as comprehensively as possible, we incorporate a dot product of the user and post embeddings into our model.

Objectives To better train our computational model, we perform multi-task learning to utilize different supervision signals. First, we consider a binary classification task of estimating repost or not with a cross-entropy loss. Additionally, we perform the triplet ranking task (Chen et al., 2009; Hoffer and Ailon, 2014) to distinguish the subtle differences among the susceptibility scores of multiple users when the same false content is present.

During each forward pass, our model is provided with three user-post pairs: the anchor pair (u_a, p), the similar pair (u_s, p), and the dissimilar pair (u_{ds}, p). We regard the similar user u_s as the user who reposted p if and only if user u_a reposted p. The dissimilar user u_{ds} is defined by reversing this relationship. When multiple candidate users exist for either u_s or u_{ds}, we randomly select one. However, if there are no suitable candidate users available, we randomly sample one from the positive (for "reposted" cases) or negative examples (for "did not repost" cases) and pair this randomly chosen user with the misinformation post p.

In Eq. 3, we define our loss function. Here, y_i takes the value of 1 if and only if user u_i reposted misinformation post p. The parameter α corresponds to the margin employed in the triplet loss, serving as a hyperparameter that determines the minimum distance difference needed between the anchor and the similar or dissimilar sample for the loss to equal zero. Besides, λ is the control hyperparameter, which governs the weighting of the binary cross-entropy and triplet loss components.

$$\mathcal{L}_{\text{bce}}(u_i, p) = -\left(y_i \log(p_{\text{rt}}(u_i, p)) + (1 - y_i)\log(1 - p_{\text{rt}}(u_i, p))\right)$$
$$\mathcal{L}_{\text{triplet}}(u_a, u_s, u_{ds}, p) = \text{ReLU}(\|s_{u_a,p} - s_{u_s,p}\|_2^2 - \|s_{u_a,p} - s_{u_{ds},p}\|_2^2 + \alpha)$$
$$\mathcal{L}(u_a, u_s, u_{ds}, p) = \frac{\lambda}{3}\sum_{i\in\{a,s,ds\}} \mathcal{L}_{\text{bce}}(u_i, p) + (1-\lambda)\mathcal{L}_{\text{triplet}}(u_a, u_s, u_{ds}, p) \quad (3)$$

4 Dataset and Experiment Setup

We have chosen Twitter as our data source because it hosts a diverse collection of users and allows for free-text personal and emotional expression. Furthermore, Twitter provides crucial metadata, including timestamps and location data, which are useful for our subsequent analysis.

Misinformation Tweets We consider two misinformation tweet datasets: the `ANTi-Vax` dataset (Hayawi et al., 2021) was collected and annotated specifically for COVID-19 vaccine misinformation tweets. And `CoAID` (Covid-19 Healthcare Misinformation Dataset; Cui and Lee, 2020) encompasses a broader range of misinformation related to COVID-19 healthcare, including fake news on websites and social platforms. The former dataset contains 3,775 instances of misinformation tweets, while the latter contains 10,443. However, a substantial number of tweets within these two datasets do not have any repost history. Hence, we choose to retain only those misinformation tweets that have been retweeted by valid users. Finally, we have collected a total of 1,271 misinformation tweets for our study.

Positive Examples We define the positive examples for modeling as (u_{pos}, p) pairs, where user u_{pos} viewed and retweeted the misinformation post p. We obtained all retweeters for each misinformation tweet through the Twitter API.

Negative Examples Regarding negative examples, we define them as (u_{neg}, p) pairs where user u_{neg} viewed but did not retweet misinfo post p. However, obtaining these negative examples poses a considerable challenge, because the Twitter API does not provide information on the "being viewed" activities of a specific tweet. To address this issue, we construct potential negative users u_{neg} who are highly likely to have viewed a particular post p but did not repost it, following these heuristics: 1) u_{neg} should be a follower of the author of the misinformation post p, 2) u_{neg} should not retweet p, and 3) u_{neg} was active on Twitter within 10 days before and 2 days after the timestamp of p.

	Total	Positive	Negative
# Example	7658	3811	3847
# User	6908	3669	3255
# Misinfo tweet	1271	787	1028

Table 1: **Data Statistics** of our constructed training dataset. We show the statistics for the number of the user-tweet pairs (*# Example*), unique users (*# User*), and unique misinformation tweets (*# Misinfo tweet*) in the overall dataset and the positive and negative subsets.

In the end, we collected 3,811 positive examples and 3,847 negative examples, resulting in a dataset consisting of a total of 7,658 user-post pairs. We divide the dataset into three subsets with an 80% - 10% - 10% split for train, validation, and test purposes, respectively. The detailed statistics of the collected data are illustrated in Tab. 1. We provide the training details of our model in Appendix B.

5 Evaluation

We demonstrate the effectiveness of our susceptibility modeling by directly comparing our estimations with human judgment (§5.1) and indirectly evaluating for assessing sharing behavior (§5.2, §5.3).

5.1 Validation with Human Judgement

Due to the abstract nature of susceptibility and the absence of concrete ground truth, we encounter challenges in directly assessing our susceptibility modeling. As a result, we tend to human evaluations to validate the effectiveness of our modeled susceptibility. Given the inherent subjectivity in the concept of susceptibility, and to mitigate potential issues arising from variations in individual evaluation scales, we opt not to request humans to annotate a user's susceptibility directly. Instead, we structure the human evaluation as presenting human evaluators with pairs of users along with their historical tweets and requesting them to determine which user appears more susceptible to overall COVID-19 misinformation. We provide more details regarding the human judgment framework and the utilized interface in Appendix C.

Subsequently, we compared the predictions made by our model with the human-annotated pre-

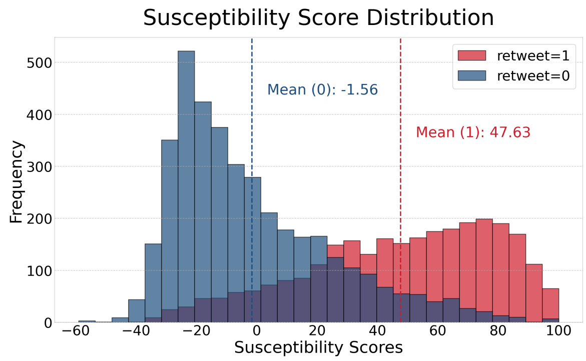

Figure 2: **Susceptibility Score Distribution** among positive and negative user-tweet pairs. The distribution of susceptibility levels, estimated by our computational modeling, among positive (red) and negative (blue) examples exhibits a significant difference.

dictions. To obtain predictions from our model, we compute each user's susceptibility to overall COVID-19 misinformation by averaging their susceptibility scores to each COVID-19 misinformation tweet in our dataset. As presented in Tab. 2, our model achieves an agreement of 72.90% with human predictions, indicating a solid alignment with the annotations provided by human evaluators. Additionally, we consider a baseline that directly calculates susceptibility scores as the cosine similarity between the user and misinformation tweet embeddings. Compared to this baseline, our susceptibility modeling brings a 9.35% improvement. Moreover, we conduct a comparison with ChatGPT by providing it with instructions based on the task description of the susceptibility level comparison setting in a zero-shot manner (more details are in Appendix E). We notice that our model even outperforms predictions made by ChatGPT, despite ChatGPT being a significantly larger model than ours. These results of the human judgment validate the effectiveness of our proposed susceptibility modeling, showcasing its capability to reliably estimate user susceptibility to COVID-19 misinformation.

	Our	Baseline	ChatGPT
Agreement	72.90	63.55	62.62

Table 2: **Comparison with Human Judgement.** `Baseline` refers to a direct comparison based on cosine similarity between user and misinformation embeddings, while `ChatGPT` denotes prompting the ChatGPT model (engine *gpt-3.5-turbo*) for determining the more susceptible user in a zero-shot manner.

5.2 Inferred Susceptibility Score Distribution

We provide a visualization showing the distribution of susceptibility scores produced by our model for both the positive and negative examples within the training data. As illustrated in Fig. 2, there is a significant disparity in the distribution between positive and negative examples. The difference in the means of the positive and negative groups is statistically significant, with a p-value of less than 0.001. This confirms our assumption that the susceptibility level to misinformation is a fundamental influencing factor for subsequent sharing behavior.

5.3 Resulting Sharing Behavior Prediction

Additionally, as described in §3, a high susceptibility level to misinformation is highly likely to lead to subsequent sharing behavior on social media. Here, we reinforce this assumption by showcasing that our learned susceptibility model exhibits a strong capability to predict subsequent sharing behavior. When tested on the held-out test set, our model achieves a test accuracy of 78.11% and an F1 score of 77.93. These results indirectly demonstrate the validity of our computational modeling for latent susceptibility within the human thought process.

6 Analysis

To further illustrate the significance of our work for the Computational Social Science community in susceptibility and misinformation research, we conducted a large-scale analysis on our collected large Twitter datasets and analyzed the correlation between user's susceptibility and their psychological factors (§6.1), professional backgrounds (§6.2), and geographical distribution (§6.2). Our findings demonstrate that the large-scale analysis enabled by our proposed efficient susceptibility modeling not only corroborates the results of previous questionnaire-based studies, but also shows the potential of further extending the scope of research on susceptibility and misinformation.

6.1 Correlation with Psychological Factors

Previous research on human susceptibility to health and COVID-19 misinformation primarily relied on questionnaire surveys (Scherer et al., 2020; Nan et al., 2022; van der Linden, 2022). These studies have identified several psychological factors that influence individuals' susceptibility to misinformation. For instance, analytical thinking (as opposed to intuitive thinking), trust in science, and positive emotions have been linked to a greater resistance to health misinformation. Conversely, susceptibility to health misinformation is frequently associated

Factors	Coeff.	Factors	Coeff.
Analytic Thinking	-0.31	Emotion - Positive	-0.08
Political Leaning	0.13	Emotion - Anxious	0.08
Ethnicity	0.09	**Emotion - Angry**	0.16
Religiosity	0.10	**Emotion - Sad**	0.14
Technology	-0.09	**Swear**	0.18
Illness	0.09	Wellness	-0.02

Table 3: **Correlation Coefficients** between our modeled susceptibility levels and various psychological factors. Our model reveals correlations that are consistent with findings from prior questionnaire-based health susceptibility studies. The factors with absolute scores greater than 0.1 are highlighted in red (+) and blue (-).

with factors such as conspiracy thinking, religiosity, conservative ideology, and negative emotions. In this part, we analyze the correlation coefficients between our modeled susceptibility scores and the aforementioned factors to determine whether our results align with previous research findings.

To achieve this, we compute factor scores for each user in our dataset based on their historical tweets using LIWC Analysis.[3] We mainly consider the following factors: *Analytic Thinking*, Emotions (*Positive* emotions, *Anxious*, *Angry* and *Sad*), *Swear*, *Political Leaning*, *Ethnicity*, *Technology*, *Religiosity*, *Illness* and *Wellness*. These factors have been extensively studied in previous works and can be inferred from a user's historical tweets. We calculate and plot the Pearson correlation coefficients between each factor and the susceptibility level estimated by our model in Tab. 3.

In our analysis, the correlations are consistent with findings from previous social science studies that relied on surveys to assess participants' health susceptibility. For instance, *Analytic Thinking* is a strong indicator of low susceptibility, with a correlation coefficient of -0.31. Conversely, certain features such as *Swear*, *Political Leaning*, and *Angry* exhibit a weak correlation with a high susceptibility level. These results not only corroborate the conclusions drawn from previous questionnaire-based studies (van der Linden, 2022; Nan et al., 2022) but also provide further validation for the effectiveness of our computational modeling for susceptibility.

6.2 Community Differences

We further leverage our computational model to investigate how susceptibility level differs and compares between different community groups on social networks. Specifically, two different types of communities are considered: professional and geographical communities.

To perform a reliable analysis among different communities, a large-scale user dataset is needed. To address this requirement, we sample 100,000 users across the world from the existing COVID-19 Tweet Dataset (Taylor et al., 2023) which contains all COVID-19-related tweets for a certain time. [4] Furthermore, for better interpretability, we normalize the resulting susceptibility scores within the range of -100 to 100 using *Min-Max* normalization. We define -100 to indicate that the individual holds the most resistance to misinformation, while 100 means the individual is easiest to believe in misinformation when encountered. To obtain an aggregated susceptibility score for a community, we calculate the mean of individual susceptibility scores for all users within that community.

Occupation and Professional Community We first explore how susceptibility varies among users with different occupations. There is a social consensus regarding the susceptibility of the practitioners within a specific occupation community. For example, susceptibility scores towards health misinformation are expected to be significantly lower among experts in health-related fields compared to the general population (van der Linden, 2022; Nan et al., 2022). We consider the following professional communities and compare their average susceptibility scores: Education (*Edu*), Society and Public (*S&P*), Health and Medicine (*H&M*), Finance and Business (*F&B*), Science and Technology (*S&T*), Arts and Media (*A&M*), as well as *N/A* for Twitter users who do not specify their occupation in their user descriptions.

The results are presented in Tab. 4. It is worth noting that professions closely associated with *S&T*, *H&M*, and *Edu* tend to exhibit lower susceptibility to COVID-19 misinformation. In contrast, occupations within the *A&M* area demonstrate comparatively higher susceptibility, possibly because of their greater exposure to misinformation and stronger emotional reactions. These findings align with our expectations and reinforce the pre-

[3] liwc.app. For each user, we compute the final factor score by calculating the average value across the user's historical tweets. However, for emotional factors like anxiety and anger, which may appear less frequently, we choose to use the maximum value instead to better capture these emotions.

[4] Besides, we make sure each sampled user has posted more than 100 historical tweets between January 2020 and April 2021. For each user, we utilize the Twitter API to gather their user descriptions and location information, after which we extract and categorize their occupations from their self-reported descriptions with ChatGPT in a zero-shot manner.

vious conclusions that expertise and knowledge in relevant fields serve as protective factors against misinformation (Nan et al., 2022)[5].

Occupation	Suscep.	# Users
N/A	4.6201	35145
Arts and Media	-0.1504	12635
Science and Technology	-2.2076	7170
Finance and Business	-5.4192	5844
Health and Medicine	-5.4762	6272
Society and Public	-6.7747	10973
Education	-7.8070	5261

Table 4: **Susceptibility Distribution by Professional Field**. We present the average susceptibility scores, estimated by our computational modeling, for 6 main professional fields. *S&T*, *H&M*, and *Edu* (highlighted in blue) tend to have lower susceptibility to COVID-19 misinformation, consistent with existing studies.

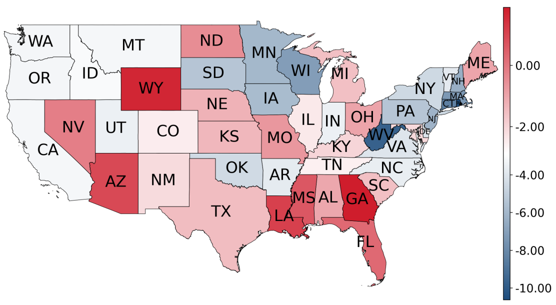

Figure 3: **Susceptibility Distribution by U.S. State**. We plot the susceptibility score, estimated by our computational modeling (with Bayesian smoothing), for each state in the U.S. The average susceptibility score in the overall U.S. (-2.87) is used as the threshold, with scores above it displayed in red, and those below it in blue. Due to insufficient data points, we are only displaying data for 48 contiguous states within the U.S.

Geographical Community We further investigate the geographical distribution of susceptibility to COVID-19 misinformation, specifically focusing on different U.S. states.[6] This analysis enables us to explore the influence of political ideology associated with different U.S. states (Gelman, 2008) on susceptibility to misinformation. Out of the 100,000 users sampled from around the world, 25,653 users are from U.S. states with more than 200 users for each state. As shown in Fig. 3, the distribution of susceptibility levels estimated by our computational modeling is imbalanced across U.S. states and demonstrates a certain degree of correlation with political leanings. In general, states known to have a more conservative population tend to have relatively higher susceptibility scores, while states that are considered more liberal have lower scores. The average susceptibility score for users in blue or red states is -3.66 and -2.82 respectively.[7] We observe that 60% of the ten states with the highest susceptibility scores are red states, and 90% of the ten states with the lowest susceptibility scores are blue states. This trend corresponds with the conclusion observed in various previous studies, where political ideology influences people's perspectives on scientific information (McCright et al., 2013; Baptista et al., 2021; Imhoff et al., 2022). However, it is important to acknowledge the limitations of our analysis, as it solely reflects the estimated susceptibility distribution of the sampled users within each state. In Appendix F, we present the average susceptibility scores calculated based on our sampled users for each U.S. state, along with the corresponding number of users.

7 Conclusion

In this work, we propose a computational approach to efficiently model people's latent susceptibility to misinformation. While previous research on susceptibility is heavily relied on self-reported beliefs collected from questionnaire-based surveys, our model trained in a multi-task manner can estimate user's susceptibility levels only based on their posting and sharing behaviors on social media. When compared with human judgment, our model shows highly aligned predictions on a susceptibility comparison evaluation task. To demonstrate the potential of our proposed computational modeling in extending the scope of previous misinformation-related studies, we leverage the susceptibility scores estimated by our model to analyze factors that influence susceptibility to COVID-19 misinformation. Our analysis considers a diverse population from various professional and geographical backgrounds, and the results obtained through our computational modeling not only align with but also support and extend the findings from previous survey-based social science studies.

[5] We notice that Twitter users who don't declare their occupation in their user description (*N/A*) exhibit a higher susceptibility to COVID misinformation. This may be because those who are willing to declare their profession are often public figures who care more about their reputation.

[6] Given the imbalance in the number of users from different U.S. states, we calculate average susceptibility scores for each state with Bayesian smoothing. We use the overall mean and overall standard deviation as priors, and the more users in the state, the less the overall mean will affect that state's score.

[7] Red and blue states are determined by the 2020 presidential election results, with red states leaning Republican and blue states leaning Democratic.

Limitations

Besides investigating the underlying mechanism of misinformation propagation at a large scale, the susceptibility scores estimated by our model have the potential to be used to visualize and interpret individual and community vulnerability in information propagation paths, identify users with high risks of believing in false claims and take preventative measures, and use as predictors for other human behaviors. However, while our research represents a significant step in computational modeling susceptibility to misinformation, several limitations should be acknowledged.

First, our model provides insights into susceptibility based on the available data and the features we have incorporated. However, it's important to recognize that various other factors, both individual and contextual, may influence susceptibility to misinformation. These factors, such as personal experiences and offline social interactions, have not been comprehensively incorporated into our modeling and should be considered in future research.

Moreover, our modeled susceptibility scores represent an estimation of an individual's likelihood to engage with misinformation. These scores may not always align perfectly with real-world susceptibility levels. Actual susceptibility is a complex interplay of cognitive, psychological, and social factors that cannot be entirely captured through computational modeling. Our modeling should be viewed as a valuable tool for identifying trends and patterns, rather than as a means for providing definitive individual susceptibility assessments.

Finally, our study's findings are based on a specific dataset and may not be fully generalizable to all populations, platforms, or types of misinformation. Especially when examining the geographical distribution of susceptibility, it's important to note that not all U.S. states have a sufficient amount of Twitter data available for analysis, due to the high cost of data collection. Furthermore, platform-specific differences and variations in the types of misinformation can potentially impact the effectiveness of our modeling and the interpretation of susceptibility scores.

Ethics Statement

Analyzing and modeling susceptibility to misinformation can potentially raise several ethical concerns, particularly when applied at an individual level. Due to its dual nature, our modeling can not only be used to identify users with a high risk of believing in misinformation and taking preventative measures to reduce harm, but it also holds the potential for misuse by malicious actors, leading to privacy violations, stigmatization, and targeted attacks. To minimize the risk, we refrained from using any personally identifiable information (PII) data in our work. Nevertheless, it remains important to carefully consider the ethical implications associated with the deployment of computational models like ours, enhance regulatory oversight, and ensure responsible and transparent utilization.

We acknowledge the need for ongoing ethical scrutiny and are committed to the responsible release of our trained model, and this includes requiring users to sign a Data Use Agreement that explicitly prohibits any malicious or harmful use of our model. Within this agreement, researchers and practitioners will also be required to acknowledge the limitations (§7), that our modeling may not fully or accurately represent an individual's real susceptibility level.

References

Nasser Alsadhan and David Skillicorn. 2017. Estimating personality from social media posts. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 350–356. IEEE.

Sacha Altay, Anne-Sophie Hacquin, and Hugo Mercier. 2022. Why do so few people share fake news? it hurts their reputation. *New Media & Society*, 24(6):1303–1324.

João Pedro Baptista, Elisete Correia, Anabela Gradim, and Valeriano Piñeiro-Naval. 2021. The influence of political ideology on fake news belief: The portuguese case. *Publications*, 9(2).

Kenneth A. Bollen. 2002. Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1):605–634. PMID: 11752498.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. 2003. The theoretical status of latent variables. *Psychological Review*, 110(2):203–219.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

William J. Brady, Molly J. Crockett, Jay J. Van Bavel, and Jay J. Van Bavel. 2020. The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15:1010 – 978.

Nadia M. Brashier and Daniel L. Schacter. 2020. Aging in an era of fake news. *Current Directions in Psychological Science*, 29:316 – 323.

Rex Perez Bringula, Annaliza E. Catacutan, Manuel B. Garcia, John Paul S. Gonzales, and Arlene Mae C. Valderama. 2021. “who is gullible to political disinformation?” : predicting susceptibility of university students to fake news. *Journal of Information Technology & Politics*, 19:165 – 179.

Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhiming Ma, and Hang Li. 2009. Ranking measures and loss functions in learning to rank. In *Neural Information Processing Systems*.

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD ’21, page 148–157, New York, NY, USA. Association for Computing Machinery.

Cindy K Chung and James W Pennebaker. 2018. What do we know when we liwc a person? text analysis as an assessment tool for traits, personal concerns and life stories. *The Sage handbook of personality and individual differences*, pages 341–360.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Nicholas C. Dias, Gordon Pennycook, and David G. Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*.

Ullrich K. H. Ecker, Stephan Lewandowsky, Jonathan Cook, Philipp Schmid, Lisa K. Fazio, Nadia M. Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1:13 – 29.

Álex Escolà-Gascón, Neil Dagnall, and Josep Gallifa. 2021. Critical thinking predicts reductions in spanish physicians’ stress levels and promotes fake news detection. *Thinking Skills and Creativity*, 42:100934 – 100934.

Jeffrey L. Foster, Thomas Huthwaite, Julia A. Yesberg, Maryanne Garry, and Elizabeth F. Loftus. 2012. Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta psychologica*, 139 2:320–6.

Andrew Gelman. 2008. Red state, blue state, rich state, poor state: Why americans vote the way they do.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.

Samrat Gupta, Gaurav Jain, and Amit Anand Tiwari. 2022. Polarised social media discourse during covid-19 pandemic: evidence from youtube. *Behaviour & Information Technology*, 42:227 – 248.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2021. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23 – 30.

McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H. A. Schwartz, David H. Epstein, Lorenzo Leggio, and Brenda L Curtis. 2021. Bots and misinformation spread on social media: Implications for covid-19. *J Med Internet Res*, 23.

Elad Hoffer and Nir Ailon. 2014. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*.

Ming hui Li, Zhiqin Chen, and Li-Lin Rao. 2022. Emotion, analytic thinking and susceptibility to misinformation during the covid-19 outbreak. *Computers in Human Behavior*, 133:107295 – 107295.

Roland Imhoff, Felix Zimmer, Olivier Klein, João H. C. António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, Aleksandra Cichocka, Sylvain Delouvée, Karen M. Douglas, Asbjørn Dyrendal, Tom Etienne, Biljana Gjoneska, Sylvie Graf, Estrella Gualda, Gilad Hirschberger, Anna Kende, Yordan Kutiyski, Peter Krekó, Andre Krouwel, Silvia Mari, Jasna Milošević Đorđević, Maria Serena Panasiti, Myrto Pantazi, Ljupcho Petkovski, Giuseppina Porciello, André Rabelo, Raluca Nicoleta Radu, Florin A. Sava, Michael Schepisi, Robbie M. Sutton, Viren Swami, Hulda Thórisdóttir, Vladimir Turjačanin, Pascal Wagner-Egger, Iris Žeželj, and Jan-Willem van Prooijen. 2022. Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour*, 6(3):392–403.

A.K.M. Najmul Islam, Samuli Laato, Shamim Hayder Talukder, and Erkki Sutinen. 2020. Misinformation sharing and social media fatigue during covid-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, 159:120201 – 120201.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2022. The inductive bias of in-context learning: Rethinking pretraining example design. In *International Conference on Learning Representations*.

Stephan Lewandowsky and Ullrich K. H. Ecker. 2012. Psychological science in the public interest , in press misinformation and its correction : Continued influence and successful debiasing.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yanchen Liu, Timo Schick, and Hinrich Schtze. 2023. Semantic-oriented unlabeled priming for large-scale language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 32–38, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi Jane Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5:337 – 348.

Mingyu Derek Ma, Alexander K. Taylor, Nuan Wen, Yanchen Liu, Po-Nien Kung, Wenna Qin, Shicheng Wen, Azure Zhou, Diyi Yang, Xuezhe Ma, Nanyun Peng, and Wei Wang. 2023. Middag: Where does our news go? investigating information diffusion via community-level information pathways. *ArXiv*, abs/2310.02529.

Aaron M. McCright, Katherine E. Dentzman, Meghan Charters, and Thomas Dietz. 2013. The influence of political ideology on trust in science. *Environmental Research Letters*, 8.

Amy Mitchell, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. Many americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center*, 5:2019.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2020. Health misinformation.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social science & medicine*, 314:115398.

Gordon Pennycook and David G. Rand. 2017. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*.

Gordon Pennycook and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25:388–402.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jon Roozenbeek, Rakoen Maertens, Stefan M. Herzog, Michael Geers, Ralf H.J.M. Kurvers, Mubashir Sultan, and Sander van der Linden. 2022a. Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*.

Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John R. Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society Open Science*, 7.

Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022b. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8.

Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on "inoculation" theory can cut sensor to misinformation cross farming.

Leah R. Rosenzweig, Bence Bagó, Adam J. Berinsky, and David G. Rand. 2021. Happiness and surprise are associated with worse truth discernment of covid-19 headlines among social media users in nigeria. *Harvard Kennedy School Misinformation Review*.

Nikita A. Salovich, Amalia M. Donovan, Scott R. Hinze, and David N. Rapp. 2020. Can confidence help account for and redress the effects of reading inaccurate information? *Memory & Cognition*, 49:293–310.

Laura D. Scherer, Jonathon McPhetres, Gordon Pennycook, Allison Kempe, Larry A. Allen, Christopher E. Knoepke, Channing E. Tate, and Daniel D. Matlock. 2020. Who is susceptible to online health misinformation? a test of four psychosocial hypotheses. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*.

Prerika R Sharma, Kimberley A. Wade, and Laura Jobson. 2022. A systematic review of the relationship between emotion and susceptibility to misinformation. *Memory*, 31:1 – 21.

Louise Sundararajan, Rachel Sing-Kiat Ting, Shu-Kai Hsieh, and Seong-Hyeon Kim. 2022. Religion, cognition, and emotion: What can automated text analysis tell us about culture? *The Humanistic Psychologist*, 50(2):213.

Alexander K. Taylor, Nuan Wen, Po-Nien Kung, Jiaao Chen, Violet Peng, and W. Wang. 2023. Where does your news come from? predicting information pathways in social media. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Cecilie Steenbuch Traberg and Sander van der Linden. 2022. Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*.

Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28:460 – 467.

Wei Wang, Ivan Hernandez, Daniel A Newman, Jibo He, and Jiang Bian. 2016. Twitter analysis: Studying us weekly trends in work stress and emotion. *Applied Psychology*, 65(2):355–378.

Brian E. Weeks. 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, 65:699–719.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2020. The covid-19 infodemic: Twitter versus facebook. *Big Data & Society*, 8.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

A Potential Questions

We address here some potential questions readers might have about our work:

What is the goal of the method design? We aim to design a framework to estimate users' susceptibility indirectly with a comprehensive representation of their observable reposting behavior data rather than training their ground-truth susceptibility. The sentence embedding model (described in Appendix B) is selected to create rich representations of users and posts. Its effectiveness has been shown in existing works (Levine et al., 2022; Liu et al., 2022, 2023; Xu et al., 2023; Yang et al., 2024; Deng et al., 2024).

Proposed framework lacks novelty? *Multi-fold novelty:* Our work contributes to the literature in multiple dimensions. 1) Proposing a brand new task without existing data, baselines and evaluation setup. Modeling user's susceptibility efficiently and empirically while no ground-truth susceptibility to train or evaluate is provided; 2) Being the first to make large-scale susceptibility analysis possible, while previous works rely on expensive self-reported human-collected questionnaires; 3) Being the first large-scale analysis of the relationship between susceptibility and social/psych. factors, professional backgrounds and geographical distribution.

Conventional model is a secondary component under a bigger framework: Even though we use RoBERTa model trained in previous works to obtain user and post embeddings, we are the first to design an indirect estimation framework for susceptibility from users' history. The off-the-shelf sentence embedding model is a secondary component and it can be replaced with other models, such as LLMs.

Not just method novelty: The susceptibility modeling framework/method is only one part of our contribution, and more importantly, our other important core contribution is the large-scale analysis enabled by our proposed susceptibility modeling method and the interesting findings shown by this large-scale analysis. These findings not only corroborate the findings of previous questionnaire-based studies (which are not possible to scaled-up) but also showing the potential of extending the scope of misinfo research.

Reposting behavior not sufficient to provide a full understanding of believing? We acknowledge that modeling the user's susceptibility to misinformation only with the supervision of their sharing behavior on social media is a little bit limited. However, other information, like "user's intent behind reposting", is almost never indicated on any social media, and intractable to large-scaly collect and impossible to scale up, which goes against our original motivation. And actually, only users themselves know their sharing intents, whether they are expressing approval or perhaps irony, which we believe are rare cases. Therefore, we propose to infer a user's susceptibility to misinformation based solely on their historical tweets, because user's historical posts and reposting behavior are much easier to collect. This not only enables effective modeling of user's susceptibility and more importantly, it enables large-scale analysis to help people better understand the behind mechanism, patterns, influencing factors and distribution of human's susceptibility to misinformation, which has been shown in our work. We have also acknowledged the data unavailability in the Limitations section of this paper.

Why not use user personality features? We do not explicitly incorporate additional user characteristics into our modeling, because the additional information is relatively difficult to get on social media. However, users could display their personalities, etc., user characteristics information through their posts, thus justifying our design choice that our modeling solely based on users' historical posts could also take these user characteristics into account. To further confirm our point, there are lots of previous works proposing to predict/extract a user's personality from their posted or liked social media posts (Golbeck et al., 2011; Alsadhan and Skillicorn, 2017).

How reliable are LIWC scores used in analysis in Section 6.1? LIWC is a widely used, well-established, and convenient tool for analyzing text data in the field of computational linguistics and psychology. There are substantial works based on LIWC analysis and the reliability of LIWC has also been demonstrated in numerous studies across various domains (Wang et al., 2016; Chung and Pennebaker, 2018; Sundararajan et al., 2022; Boyd et al., 2022).

Why not add more comparing baselines? We work on a brand new task setting: estimate users' susceptibility **indirectly** without any ground-truth susceptibility labels provided. The only input to the estimation model is users' historical posts. The unique setting prevents us from finding any prior works that follow this challenging setting, which makes it impossible for us to conduct a direct comparison with existing susceptibility works. Hence, we come up with two methods: cosine similarity between embeddings and ChatGPT. We also observe that cosine similarity is not a weak baseline, and it yields even better performance than ChatGPT.

Why not include more ablation studies? Our work mainly focuses on developing a novel framework for susceptibility modeling and demonstrates its potential to enable large-scale analysis and facilitate susceptibility and misinformation-related research. Thus, we prioritize our emphasis on designing a reasonable modeling, rather than necessarily aiming for the optimal modeling. This is because, as previously stated, the unobservable nature and lack of ground truth for susceptibility prevent us from directly optimizing the modeling for susceptibility itself; instead, we can only do so for the indirect sharing predictions task. Consequently, ablation studies are of very limited significance in this context. We believe that including too many ablation studies could even deviate the audience's focus away from our research goals.

How to/what is the performance of adapting the proposed framework to other domains besides COVID-19? The advantage of our method is the capability to estimate susceptibility without the need for ground-truth user susceptibility labels. Using users' historical posts, target posts, and users' retweet behavior labels is sufficient to train the model. We will release our code, and people can try to robustness check it and extend it to more domains.

B Training Details

We use the *sentence-transformers/all-roberta-large-v1* model from `sentence-transformers` as our sentence embedder. Through grid search on learning rates ranging from 1e-5 to 5e-4 and λ values from 0 to 1, we train our model using a learning rate of 3e-5, set the hyperparameter λ to 0.9, and the margin α to 1 for 100 epochs on the training set, as detailed in §3. Following the training process, we select the checkpoint with the lowest validation loss and proceed to evaluate its performance on the test set.

C Human Judgement

Here, we provide details about the human judgment framework utilized in our work.

During human judgment, annotators are tasked with selecting the more susceptible user based on five historical tweets for each user. We offer the user interface used for human judgment in Figure 4. In the task description, susceptibility is described as being more likely to believe, be influenced by, and propagate COVID-19 misinformation. To account for annotator uncertainty, we provide four options: *Definitely User A*, *Probably User A*, *Definitely User B*, and *Probably User B*. Furthermore, we also request annotators to identify the "most susceptible tweet" for the selected user, to enhance the reliability of annotations. This tweet should best exemplify the user's susceptibility to COVID-19 misinformation or be the basis for the annotator's decision.

Also, it is important to note that even when both users seem to have low susceptibility to COVID-19 misinformation, we still ask the annotator to make a choice. This is because our goal is to rank users based on their relative susceptibility, offering a comparative assessment rather than an absolute determination.

In total, we randomly sampled 110 user pairs and collected three annotations for each user pair. We recruited human annotators from Amazon Mechanical Turk (AMT) in the U.S. and compensated each annotator with $0.5 (hourly wage higher than the federal minimum wage). To determine the gold label for each user pair, we applied a weighted majority voting approach, assigning a value of 0.5 to *Probably User X* and a value of 1 to *Definitely User X*. We excluded user pairs with tied annotations, resulting in a final dataset of 107 user pairs. The kappa score for interrater agreement among the annotators is 0.74.

D Examples of User Posts and Susceptibility Scores

The user `KatCapps`'s susceptibility score is estimated as `38.62` when the user sees the tweet:

> The coronavirus infection rate is still too high. There will be a second wave | David Hunter [Link]

History tweets posted by the user are:

- RT @gregggonsalves: Study estimates 24 states still have uncontrolled coronavirus spread

- RT @JoeSudbay: OSHA chas not issued enforceable guidelines for protecting employees from covid-19, as it did during the H1N1 outbreak in

- RT @mlipsitch: New Oped with @rickmalley | Treating Mild Coronavirus Cases Could Help Save Everyone - The New York Times

- RT @stevesilberman: Texas church that rushed to reopen cancels masses after priest dies and others contract #coronavirus.

- RT @carlzimmer: Several cases of coronavirus reported after a swim party in Arkansas, governor says

- RT @GlennKesslerWP: They Survived the Worst Battles of World War II. And Died of the Virus.

Another user `AmitSin91018424`'s susceptibility score is estimated as `-12.27` when the user sees the tweet:

> Dominic Cummings has broken Covid-19 policy trust, say top scientists [Link]

History tweets posted by the user are:

- RT @guardian: The pandemic has laid bare the failings of Britain's centralised state | John Harris

- RT @guardian: Coronavirus world map: which countries have the most cases and deaths?

E ChatGPT Prompt Template

In Fig. 5, we present the template used to prompt ChatGPT for the susceptibility comparison task (§5.1).

F Average Susceptibility Scores and User Counts by U.S. State

We provide the aggregated susceptibility scores estimated by our computational modeling for each U.S. state (§6.2), along with the number of sampled users in Tab. 5.

Comparing Susceptibility to COVID–Related Misinformation

In this task, you will be presented with 2 Twitter users, each with 5 historical tweets presented in chronological order. Your objective is to determine which of the two users is more susceptible to COVID–related misinformation, which we define as being more likely to believe, be influenced by, and propagate such misinformation, e.g. through retweeting.

User A's Historical Tweets:

- ○ ${A_1}
- ○ ${A_2}
- ○ ${A_3}
- ○ ${A_4}
- ○ ${A_5}

User B's Historical Tweets:

- ○ ${B_1}
- ○ ${B_2}
- ○ ${B_3}
- ○ ${B_4}
- ○ ${B_5}

You have four options to choose from:

- ○ Definitely User A
- ○ Probably User A
- ○ Probably User B
- ○ Definitely User B

Choose option Definitely User A or Definitely User B, if you are highly confident that this user is more susceptible. And please choose option Probably User A or Probably User B, if you believe this user is more susceptible, but you are not entirely sure. It's necessary to make a choice even if both users appear to have low susceptibility to COVID misinformation. In such cases, you must select the user who, in your judgment, is relatively more susceptible.

Additionally, you are also tasked with selecting the most "susceptible" tweet for the user you have identified as more susceptible. This tweet should best reflect the user's susceptibility to COVID misinformation or be the tweet upon which you based your decision.

Figure 4: **Human Judgement Interface** utilized in our work. Participants are instructed to select the more susceptible user from a user pair based on five historical tweets for each user.

```
In this task, you will be presented with 2 Twitter users, each with 5 historical tweets presented in
chronological order. Your task is to determine which of the two users is more susceptible to COVID-
related misinformation, which we define as being more likely to believe, be influenced by, and
propagate such misinformation, e.g. through retweeting.

User A's Historical Tweets:
{userA_text}

User B's Historical Tweets:
{userB_text}

It is necessary to make a choice even if both users appear to have low susceptibility to COVID
misinformation.
In such cases, you must select the user who, in your judgment, is relatively more susceptible.

Please answer with one of the following options without any other text: A | B.
```

Figure 5: **ChatGPT Prompt Template** for the susceptibility comparison task.

State	Suscep.	# Users	State	Suscep.	# Users
Georgia	0.3935	669	Idaho	-3.2296	265
Florida	-0.2404	1592	Washington	-3.2577	526
Arizona	-0.5566	499	Montana	-3.2590	543
Louisiana	-1.3878	202	Oregon	-3.2612	260
Ohio	-1.6120	679	Utah	-3.3324	206
Texas	-1.7478	1627	Vermont	-3.3548	556
Missouri	-1.9076	308	Indiana	-3.3901	270
Nevada	-1.9857	294	Delaware	-3.4139	359
Michigan	-2.0996	575	Arkansas	-3.4179	418
Alabama	-2.3902	377	North Carolina	-3.5324	635
Maryland	-2.4763	527	South Dakota	-3.6020	351
South Carolina	-2.5456	298	Virginia	-3.7276	528
Mississippi	-2.5886	257	Oklahoma	-3.7577	291
Maine	-2.6193	208	New Hampshire	-4.1011	399
Illinois	-2.6294	816	Iowa	-4.1603	249
Nebraska	-2.6339	324	New York	-4.4226	2835
Kansas	-2.6541	328	West Virginia	-4.8056	285
Kentucky	-2.7774	469	Minnesota	-4.8423	372
Colorado	-2.8109	363	Pennsylvania	-4.8700	873
Tennessee	-2.8554	397	Rhode Island	-5.0661	488
New Mexico	-2.9178	518	Wisconsin	-5.2446	279
Wyoming	-2.9401	319	New Jersey	-5.2594	598
North Dakota	-2.9789	331	Connecticut	-5.6912	242
California	-3.2206	2849	Massachusetts	-6.3191	761

Table 5: **Susceptibility Scores Estimated by Our Computational Model and Number of Sampled Users per U.S. State**. Due to insufficient data points, we only consider 48 contiguous states within the U.S.