

# Classifying Conspiratorial Narratives at Scale: False Alarms and Erroneous Connections

Ahmad Diab, Rr. Nefriana, and Yu-Ru Lin

School of Computing and Information, University of Pittsburgh  
 {ahd23, rrn14, yurulin}@pitt.edu

## Abstract

Online discussions frequently involve conspiracy theories, which can contribute to the proliferation of belief in them. However, not all discussions surrounding conspiracy theories promote them, as some are intended to debunk them. Existing research has relied on simple proxies or focused on a constrained set of signals to identify conspiracy theories, which limits our understanding of conspiratorial discussions across different topics and online communities. This work establishes a general scheme for classifying discussions related to conspiracy theories based on authors' perspectives on the conspiracy belief, which can be expressed explicitly through narrative elements, such as the agent, action, or objective, or implicitly through references to known theories, such as chemtrails or the New World Order. We leverage human-labeled ground truth to train a BERT-based model for classifying online CTs, which we then compared to the Generative Pre-trained Transformer machine (GPT) for detecting online conspiratorial content. Despite GPT's known strengths in its expressiveness and contextual understanding, our study revealed significant flaws in its logical reasoning, while also demonstrating comparable strengths from our classifiers. We present the first large-scale classification study using posts from the most active conspiracy-related Reddit forums and find that only one-third of the posts are classified as positive. This research sheds light on the potential applications of large language models in tasks demanding nuanced contextual comprehension.

## 1 Introduction

Conspiracy theories,<sup>1</sup> or CTs have long been the subject of curiosity, interest, and even skepticism (Van Prooijen and Douglas 2017). In recent years, the proliferation of CTs has been fueled by the rise of mis- and disinformation on the Internet. While some conspiracy narratives may seem harmless or even entertaining, others can have serious consequences. For instance, conspiracy theories related to COVID-19 (Allington et al. 2021), Pizzagate (Tangherlini et al. 2020; Bleakley 2023), and election fraud (Albertson and Guiler 2020; Bond and Neville-Shepard 2023) can

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>**Disclaimer:** This paper contains analyses and discussions of various conspiracy theories. The inclusion of these theories is solely for the purpose of academic or investigative analysis and should not be interpreted as an endorsement or validation of them.

jeopardize public health, democracy, and public trust. This can result in real-world repercussions such as outbreaks of diseases (Farhart et al. 2022; Romer and Jamieson 2020) and violent insurrections (Vegetti and Littvay 2022). Therefore, it is crucial to accurately identify conspiratorial content in order to understand the prevailing narratives and mitigate potential consequences.

Conspiracy theories are often convoluted and intricate, involving actors, events, and narratives that imply or explicitly suggest a plot behind the events. They typically lack verifiable details and rely instead on anecdotal evidence, hearsay, or speculation. These characteristics make it extremely difficult to detect conspiratorial narratives in text. In the past, many studies have employed various strategies, such as relying on simple proxies or a small, predetermined set of explicit signals, despite their limitations. For example, a straightforward approach that considers all content from CT-related forums to be conspiratorial (Phadke, Samory, and Mitra 2021a; Klein, Clutton, and Dunn 2019; Bessi et al. 2015) can produce a large number of false positives, whereas a keyword-driven method that uses predetermined keywords to extract conspiracy theories on a specific topic (Kim and Kim 2023; Hoseini et al. 2023) may overlook the false negatives and is frequently limited in scope and generalizability. Other methods, such as pattern matching (Kou et al. 2017; Introne et al. 2020), which match textual messages with pre-defined syntactical elements, can be labor-intensive and may miss narratives that lack exact matching. Recent works have also employed machine learning techniques to automate the classification of CTs, but they are frequently limited to specific topics (Shahsavari et al. 2020; Tangherlini et al. 2020) or lack clear and consistent criteria for identifying conspiratorial content (Platt, Brown, and Venske 2022). As a result, it is difficult to compare the study results of CT narratives that were extracted using various methods.

Furthermore, existing approaches that rely on simple proxies or lack theory-grounded criteria can lead to misinterpretation. For example, all discussions about CTs can be mistaken as endorsements of those theories. Our study challenges the appropriateness of previous methods and their limitations in analyzing conspiracy theory prevalence and emergence across a wide range of online forums.

In this study, we propose a general, topic-independent classification scheme for conspiracy theories that consider

their multifaceted nature. Drawing from the extensive research on conspiracy theories, our approach takes into account various aspects of conspiratorial narratives, including the author’s perspective towards the conspiracy belief (e.g., promoting or debunking). This can be manifested either (1) explicitly through the use of narrative elements such as the presence of an agent, a particular action, or an objective,<sup>2</sup> or (2) implicitly through referencing and aligning with well-known conspiracy theories.<sup>3</sup> We seek to leverage the recent advancements in large language models (LLMs), such as various BERT models and GPT to develop techniques capable of identifying conspiracy narratives within vast online corpora. To this end, we have centered our research on the following key research questions:

**RQ1.** *Can we formulate a general, topic-independent, conspiracy theory classification scheme capable of identifying established theories and emerging ones?*

**RQ2.** *How feasible is it to use large language models (including the BERT family and GPT) to automatically classify online conspiracy theory narratives?*

**RQ3.** *How prevalent are conspiracy theory narratives in conspiracy theory-related forums?*

The main contributions of this work include:

- We establish a general classification scheme that enables the systematic identification of conspiratorial content, taking into account the complex and multifaceted nature of CT narratives.
- We develop the first general CT classifier based on the BERT-family of models and demonstrate the effectiveness of incorporating LLMs for detecting online conspiracy narratives. Our best model achieved 0.787 AUC.
- We identify the advantages and disadvantages of GPT in comparison to other machine classifiers. While GPT is well-known for its expressiveness and contextual understanding, our results show that our classifiers have comparative strength in terms of classification performance. Furthermore, our analysis of GPT’s results provides valuable insights into the challenges posed by generative AI in the context of CT detection.
- With the capacity of our best classifier, we present the first large-scale classification study investigating the prevalence of CTs in the most active CT-related Reddit forums. Our analysis reveals that only one-third of the posts are classified as CT narratives, with the remainder of the posts possibly not constituting CTs or not meant to promote CTs. The classification allows us to obtain a more precise picture of the scope and reach of conspiratorial content across online communities. Further investigation of the results revealed that posts promoting CT narratives tend to receive more engagement, which suggests that promoters of such theories might have a better chance to further leverage the platform algorithms and promote their content more widely.

<sup>2</sup>E.g., “The government (Agent) is in COVID. They made a larger virus (Action) for population control (Objective).”

<sup>3</sup>E.g., “Watch out for chem trails (known CT) in the UK. I suspect something is hidden from us.”

## 2 Related Work

The analysis of CTs often involves 1) identifying CT text and 2) comprehending its content. Existing works have employed classification techniques or text mining as a part of their analysis. We group current approaches to analysis into four categories: 1) proxy-based CT extraction, 2) keyword-driven approach, 3) analysis with heterogeneous contents, and 4) topic- and cluster-based analysis.

**Proxy-Based CT Extraction** In previous works, the elicitation of conspiracy messages from social media has often relied on a simple proxy: considering all content from conspiracy-theory-related forums, such as subreddits (Phadke, Samory, and Mitra 2021a; Samory and Mitra 2018; Klein, Clutton, and Dunn 2019; Engel et al. 2022), subverses (Papasavva et al. 2021), or Facebook pages (Bessi et al. 2015; Zollo et al. 2017), as conspiratorial. As an illustration, Klein, Clutton, and Dunn (2019) investigated the language Reddit users use in a conspiracy forum, r/conspiracy, and the associated social environments. While they implemented a filtration process, it primarily focused on excluding bot accounts and non-active users without involving further CT classification/identification techniques.

Several studies employing a direct approach have examined banned forums or forums that share names with some banned forums (Engel et al. 2022; Papasavva et al. 2021; Phadke, Samory, and Mitra 2021a). Phadke, Samory, and Mitra (2021a), for instance, referenced 17 banned QAnon-related subreddits from various press sources in their investigation of social imaginaries and self-disclosures of dissonance within online conspiracy discussion communities. Engel et al. (2022) analyzed a cohort of users active in 19 QAnon-focused subreddits that were shut down as part of a moderation effort, in their study characterizing the Reddit participation of individuals engaging with QAnon conspiracy theories. Papasavva et al. (2021) focused on examining QAnon in Voat subverses that shared analogous or identical names with banned subreddits, rather than delving into the banned subreddits themselves.

While this approach has been commonly employed, our empirical evidence, as detailed later, demonstrates that not all posts from these conspiracy-related subreddits can be categorized as conspiratorial (e.g., some users actively debunked conspiratorial narratives; others express skepticism towards a conspiracy theory (Klein, Clutton, and Polito 2018)). As a result, these false positives could compromise the validity of the subsequent analysis.

**Keyword-Driven Approach** Previous research has often relied on a set of predetermined keywords to identify CT instances (Kim and Kim 2023; Phadke, Samory, and Mitra 2021b; Hoseini et al. 2023; Pasquetto et al. 2022). For example, in their investigation of the QAnon conspiracy theory ecosystem on Facebook, Kim and Kim (2023) employed 14 core keywords that specifically encapsulated the essential characteristics of QAnon (e.g. “qanon,” “qarmy,” and “greatawakening”), alongside 40 extended keywords that encompassed broader aspects of the QAnon narrative and related conspiracy theories (e.g. “pedogate”). Similarly, Phadke, Samory, and Mitra (2021b) utilized the keyword

“conspiracy” and performed regular expressions to match the string “conspir” in subreddits’ names and descriptions to find conspiratorial subreddits. By employing Principal Component Analysis and Pointwise Mutual Information, they identified similar subreddits to the already identified conspiratorial subreddits and added them to the list. Then, they categorized all users within the identified conspiratorial subreddits as participants in conspiracy communities. Furthermore, Pasquetto et al. (2022) harnessed a range of keywords, hashtags, pseudonyms, and symbols frequently used by Italian QAnon influencers on Twitter to indicate their affiliation with the QAnon movement. This approach facilitated the authors in comprehensively studying QAnon influencers’ activities by analyzing their tweets and retweets, ultimately aiding them in understanding the intricate disinformation infrastructure of Italian QAnon supporters. However, this approach is often restricted to only one or a few CTs and may not generalize well to different CT topics.

**Analysis With Heterogeneous Contents** Several studies have sought to automate the classification of CT messages using machine-learning methods (Shahsavari et al. 2020; Tangherlini et al. 2020; Phillips, Ng, and Carley 2022; Pogorelov et al. 2021; Platt, Brown, and Venske 2022). However, these works often focus on a limited number of conspiracy topics. For example, when devising a graphical approach to identify conspiracy theories in social media and news automatically, Shahsavari et al. (2020) narrowed their focus to COVID-19 conspiracy theories, while Tangherlini et al. (2020) limited theirs to Bridgegate and Pizzagate. In a study by Phillips, Ng, and Carley (2022), the examination of conspiracy theories was confined to four specific topics (climate change, COVID-19 origin, COVID-19 vaccine, Epstein-Maxwell trial) within their experiment utilizing neural network classifiers for conspiracy, stance, and topic detections. Apart from these studies, there are others that, although not restricting their focus to a specific conspiracy topic, lacked clear criteria for defining what constitutes a conspiratorial message, making it hard to judge the quality of the results (Pogorelov et al. 2021; Platt, Brown, and Venske 2022). Unlike previous research, we propose a theory-grounded CT classifier applicable for classifying CT texts in various topics.

**Topic- and Cluster-Based Analysis** Previous studies have compared the contents between conspiracy and non-conspiracy datasets. In (Miani, Hills, and Bangerter 2022), the authors applied network analysis and text-mining techniques to the LOCO dataset’s (Miani, Hills, and Bangerter 2021) Latent Dirichlet Allocation (LDA) topics. They found that, compared to non-conspiracy texts, conspiracy texts exhibited higher levels of interconnectivity between topics, greater topical diversity, and higher similarity to one another. Similarly, Nerghees, Kerkhof, and Hellsten (2018), used topic modeling and semantic network analysis to assess user responses (comments and replies to the comments) to informational (non-CT) and CT videos related to the Zika virus on YouTube. Their research revealed that responses from viewers of informational videos primarily focused on the virus’s repercussions, whereas those of conspiracy theory videos

also emphasized the parties accountable for the outbreak. Samory and Mitra (2018) utilized a syntactic rule (a agent-action-target triplet) to extract conspiratorial statements in *r/conspiracy*. By semantically clustering these triplets into some clusters, they observed “narrative-motifs” such as governmental agency–controls–communications.

### 3 Identifying Conspiracy Narratives in Reddit Posts

#### 3.1 Theoretical Definition of Conspiracy Theories

Existing research offers a variety of perspectives on the precise nature of conspiracy theories, which generally fall into two categories: a) focusing on the constituent elements of conspiracy theories, and b) assessments of the veracity of these theories.

**Elements of conspiracy theories.** Several scholarly authors (Introne et al. 2020; Zonis and Joseph 1994; Mompelat et al. 2022; Wood and Douglas 2015) have examined the components of conspiracy theories. For example, Introne et al. (2020) highlighted six terms contained within a conspiracy theory: 1) events, 2) actors, 3) goal, 4) actions, 5) consequences, and 6) target. In contrast, Zonis and Joseph (1994) argued that the tenet consists of four rather than six basic components: 1) a number of actors joining together, 2) in a secret agreement, 3) to achieve a hidden goal, and 4) which is perceived to be unlawful or malevolent. Aside from these elements, this study added another factor that makes a conspiracy theory dangerous: how the individuals involved in the conspiracy were deviating from their usual behavior. This addition is consistent with the explanation of Mompelat et al. (2022) on conspiracy belief, stating that causal narratives of an event were not “random or natural occurrences” but rather a covert plan carried out by a secret cabal of people or organizations. Meanwhile, in addition to the element of secrecy in how the conspirators carried out their agenda, Wood and Douglas (2015) also included “systematic deception” in their definition. Specifically, they defined a conspiracy theory as “an allegation regarding the existence of a secret plot between powerful people or organizations to achieve some goal (usually sinister) through systematic deception of the public.”

**The veracity of the conspiracy theories.** Previous research has also looked into the veracity of conspiracy theories (Swami and Furnham 2014; Sunstein and Vermeule 2009). A conspiracy theory, according to Swami and Furnham (2014), is “a set of false beliefs in which an omnipresent and omnipotent group of actors are believed to work together in pursuit of malevolent goals.” According to their definition, theories that turned out to be true, such as the Project MKULTRA and Watergate conspiracies (Sunstein and Vermeule 2009), are not conspiracy theories. Differently, Sunstein and Vermeule (2009) acknowledged that conspiracy theories can be true or false, although, in their study on the causes and cures of conspiracy theories, they limited their scope to false conspiracy theories only.

Overall, these prior works shared at least three common elements: agent(s), action(s), and objective(s)/secret plot(s). We further note that conspiracy theories should not be evalu-

ated based on their veracity. Therefore, we define a conspiracy theory as follows:

“A conspiracy theory is a set of narratives designed to accuse an agent(s) (be they individuals, groups, or organizations) of committing a specific action(s), which is believed to be working towards a secretive and malevolent objective(s) (secret plot).”

Our definition has elements in common with the work of Samory and Mitra (2018), who proposed utilizing an agent-action-target triplet to extract narrative motifs from text content. However, as will be discussed in a later section, our established ground truth is not based solely on the three elements explicitly mentioned in the text, but also on a contextual understanding of the three elements.

### 3.2 Operational Definition of Online Conspiracy Narratives

While the above theoretical definition serves as a general guideline for identifying conspiracy theories, the identification of conspiratorial content within social media posts has introduced complications. Online conversations are typically informal and highly opinionated (YING and Jiang 2015), which can lead to long posts that do not provide a coherent account of what was discussed or address multiple issues deviating from the main narrative of the post, i.e., the most important idea or point that the post is trying to convey. The inherent informality and possible loss of context make it more difficult to identify conspiracy theories in social media posts. A post that attempts to debunk a conspiracy theory, for instance, should not be considered a conspiracy theory. Therefore, we propose the following coding instructions that help focus on the main narrative in the post by including three additional elements:

“Following the theoretical definition, a social media post that contains a main narrative or claim that (a) represents a known conspiracy theory or (b) suggests a secret plan, along with (c) evidence of agreement or support to some extent for the mentioned conspiracy theory or secret plan.”

## 4 Dataset

Our study focuses on the highly active Reddit communities dedicated to conspiracy theories (CTs). The selection of such communities was guided by Phadke, Samory, and Mitra (2021b), which outline a meticulous manual identification of a core group of popular CT-related subreddits, followed by a systematic search for analogous subreddits based on user engagement and contributions across the platform. For each subreddit in the list, we assessed its size by querying the number of posts created from 2005 to 2021, utilizing the Pushshift Reddit API. We then used the service’s archive<sup>4</sup> to download and extract posts published between January 2019 and December 2022 from the 14 largest subreddits. No data was collected after Reddit’s new policy change, and we adhere to the platform’s data usage guidelines.<sup>5</sup> The selected time period captures a sizeable portion of recent on-

<sup>4</sup><https://files.pushshift.io/>

<sup>5</sup>See Sec. 9 for a more detailed discussion of the data access, use, and distribution.

Subreddit	Size	Full	Clean
conspiracy	1182794	779506	201054
conspiro	92850	142	3
TruthLeaks	79764	60424	872
TopConspiracy	72389	75273	311
conspiracy_commons	54437	66905	12941
climateskeptics	51971	26091	3078
conspiracytheories	36436	51138	11379
DescentIntoTyranny	19881	11567	121
ConspiracyII	16756	15228	1059
FringeTheory	15891	14749	385
conspiracyundone	13875	9954	1420
C_S_T	13346	7126	4974
1984isreality	12857	4045	43
911truth	12518	915	199
Total	1762952	1123063	237839

Table 1: CT-focused subreddits ranked by post activity (2005-2021). “Full” shows collected posts (2019-2022), while “Clean” reflects statistics post-filtration.

line conversations regarding conspiracy theories, allowing for an analysis that reflects evolving trends during this period. Table 1 provides a summary of the targeted subreddits and their respective sizes.

The collected data from r/conspiro and r/911truth were significantly lower than their reported sizes. After further investigation, it was determined that r/conspiro was banned at the beginning of 2019, while r/911truth was quarantined, limiting its visibility to users. Subsequently, both of these subreddits were excluded from the study. Furthermore, the data retrieved from r/TopConspiracy, r/conspiracy\_commons, and r/conspiracytheories exceeded the reported figures. This can be attributed to the growing popularity of these forums in recent years, particularly considering that the data collection extended until 2022, whereas sizes are reported up to 2021. Our final dataset consists of 1,122,006 posts from 12 different subreddits.

Some posts have been removed at the time of data collection – they were either deleted by the users themselves (i.e., *self-deleted*) or by any moderation step on the platform (i.e., *banned*). The titles of these removed posts may still remain, but their post contents are no longer accessible. Thus we exclude the removed posts in the subsequent analysis. Table 1 also lists the number of posts before and after filtering the self-deleted and banned posts.

## 5 Establish Ground Truth

### 5.1 Coding Samples

We took a random sample of posts from three popular subreddits described in Sec.4, namely r/conspiracy, r/conspiracy\_commons, and r/conspiracyundone. The sampling occurred after filtering short posts (with less than 30 characters). As shown in Table 2, the final ground-truth data contains 750 coded samples as a result of our human coding

process described below.<sup>6</sup>

## 5.2 Human Coding Process

In contrast to previous works, which tend to focus on a single or a few topics, we establish a ground truth that encompasses a broad range of CT-related topics. This task requires a team of knowledgeable coders who can comprehend the context of diverse CTs. We recruited five coders with prior experience annotating social media texts (e.g., hate speech), including two Ph.D. students, one master’s student, and two undergraduates with excellent English proficiency. Among the recruited annotators, two are female. We aimed for disciplinary diversity, with two members from the Computer Science field, two from Information Systems, and one specializing in Digital Narratives.

**Coding Instruction.** Each coder underwent training to ensure that they had a comprehensive understanding of the coding guidelines. In addition to the operational definition (Sec. 3.2), we identify four major coding strategies to help coders establish their knowledge of a wide array of CTs and deal with the uncertainty and lack of context in the Reddit posts:

**An inventory of known and emerging CTs.** While our definition specifies three elements, many conspiratorial narratives did not explicitly mention all three, instead referring to commonly known CTs (e.g., 5G, NWO, QAnon) or CTs that were becoming popular (e.g., Ukraine biolab, Pizzagate) at the time of posting. Therefore, we compile a list of these CTs<sup>7</sup> and ensure that the coders have a contextual understanding of them. Example: *Be on the lookout for chemtrails in the UK today. I have a theory that something is being hidden from us.* (code: CT).

**Rhetorical question vs. genuine inquiry.** Even without explicit language, a rhetorical question (e.g., *Does a lot of conspiracies lead right to Bill Gates? Is he the real leader of the NWO?*) may indicate an author’s support for a particular conspiracy theory. We distinguish between rhetorical and genuine questions regarding CTs. Example: *Who’s skeptical of the \$1200? What are the odds that they will force you to get the vaccine? Feels like a trap* (code: CT). Example: *Are there any live streams from Afghanistan that are not from a news source? Like people filming right now? Can’t find anything on YouTube.* (code: Not CT).

**Support/promotion vs. criticism/frustration/debunking.** CTs related to controversial subjects tend to provoke strong opinions and criticism. However, presenting critical viewpoints and negative sentiments towards controversial subjects in a post does not necessarily qualify it as a CT post, unless the post also expresses endorsement or support for a conspiracy belief. Example: *Oregon has made reading, math, and writing racist which I never thought we could be racist just for breathing! We should all embrace this and bring peace and global health!* (code: Not CT).

<sup>6</sup>Subject to Reddit’s terms, the dataset will be made available (see Sec. 9 for details of the data access, use, and distribution).

<sup>7</sup>The list of CTs, along with the annotation codebook and the dataset are available: <https://github.com/picsofab/Conspiratorial-Narratives-At-Scale>.

**Borderline cases.** Deciphering an author’s intent presents a great challenge. When a post mentions a CT, but the author’s support for the CT is highly ambiguous, the post is deemed non-CT. Example: *Slovakia Covid Testing Video, its cultic as hell and ends with ‘papers please’.* (code: Not CT).

All the annotators underwent a training process that gave them contextual understanding of CTs. This included definitions, emergence, and various examples connecting to recent events. They were also provided with instructions on how to label CTs, and were given at least four hours of hands-on practice with examples from a small dataset labeled by at least two authors on this paper. After the training, the five annotators were individually tasked to label each sample as either “Yes” (CT) or “No” (non-CT) according to our operational definition of conspiracy theory and guidelines. The annotation process can be summarized into three phases, as follows:

**1) Pilot Phase:** We evaluated the coders’ initial agreement after providing them with training on coding guidelines and a series of test cases to ensure they had adequate coding skills. Each coder independently labels 50 samples based on the codebook. The inter-rater agreement between each pair of coders ranged from 0.35 (fair agreement) to 0.80 (substantial agreement) as measured by Cohen’s Kappa, and the overall agreement, as measured by Fleiss’ Kappa, is 0.54 (moderate agreement). A meeting was held with all coders at the end of this phase to resolve conflicts and disagreements.

**2) Consolidation Phase:** While we observed a moderate overall agreement in the previous phase, there was variation in the coding of different individuals. To increase consistency among coders, we divided them into two groups based on their performance in the previous phase. Two rounds of annotations were conducted, each with 100 samples. The coders were instructed to individually label the samples and then to convene within their groups to propose labels that were accepted by the group. Cohen’s Kappa values between the two groups for each of the two rounds were 0.65 and 0.74, indicating substantial agreement and improvement over the previous step.

**3) Conclusion Phase:** Instead of relying on a simple majority vote, the final labels were decided through consensus among the coders. This is to ensure the highest possible coding quality can be reached through the final discussion. All five coders participated in meetings to resolve disagreements. During the meeting, coders defended their annotations and engaged in a discussion to reach a final consensus on the labeling of each sample. We compared the original annotations of each coder with the final agreement to assess the coders’ reliability. The remaining 500 samples were coded by two of the coders who demonstrated the highest reliability in the earlier phase, and the final labels were determined by consensus.

## 6 Machine Classification

Based on the human-annotated ground-truth samples, we develop machine classifiers to automatically classify a given post as CT or not. We extensively explore various approaches, including traditional machine-learning methods (ML), deep-learning models that incorporate large language

Subreddit	CT Count	non-CT Count	Total
conspiracy	100	204	304
conspiracy_commons	90	208	298
conspiracyundone	58	90	148
Total	248 (33%)	502 (67%)	<b>750</b>

Table 2: Ground-truth labels from human coding.

models (LLMs), and a state-of-the-art generative model, the Generative Pre-trained Transformer Machine (GPT). For both ML and LLMs experiments, we report the 5-fold cross-validation results with a 80:20 training/testing split.

## 6.1 Deep Learning Models (DLs)

Several pre-trained LLMs have demonstrated outstanding performance in a variety of NLP tasks. We leverage the capabilities of these LLMs, namely BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018), ALBERT (A Lite BERT) (Lan et al. 2019), DistilBERT (Distilled version of BERT) (Sanh et al. 2019), DeBERTa (Decoding-enhanced BERT with disentangled attention) (He et al. 2020), RoBERTa (Robustly optimized BERT approach) (Liu et al. 2019), and T5 (Text-to-Text Transfer Transformer) (Raffel et al. 2020), to develop deep-learning models by adding a sequence classification head on top with cross-entropy loss. We build our models utilizing Hugging Face architectures and optimize and fine-tune the classification parameters based on the labeled samples.<sup>8</sup>

## 6.2 Generative Models (GPT)

In this study, we utilize OpenAI’s APIs<sup>9</sup> to evaluate GPT’s performance in classifying online CT content. Recent studies highlighted the important role of prompt design when leveraging GPT (Brown et al. 2020; Wei et al. 2022). For instance, Liu et al. (2021) suggested that augmenting prompts with semantically similar examples can lead to performance improvements, while Zhong et al. (2023) discussed designing strategies that include manual (Wei et al. 2022) or template-defined (Kojima et al. 2022) steps for clarifying the procedure to attain favorable results. In our study, we design three prompting strategies, in zero- and few-shot settings:<sup>10</sup>

**1) Simple:** Asks the model to decide the text’s label using the prompt: *Decide whether the following text describes a conspiracy theory or not (yes/no). “[post text]”*.

**2) Justification:** Asks the model to judge the text’s label and provide a justification for the label. Prompt: *Decide*

<sup>8</sup>We used the Adam optimizer and the optimal hyperparameters (batch-size:32, lr=1e-5) with 10 epochs in the experiments. All models were trained on a single GeForce GTX TITAN X 12GB GPU. The total training time took approximately 4 hours.

<sup>9</sup>Specifically, in this study, we used the model gpt-3.5-turbo, accessed in September, 2023. However, our pilot study indicates that the results produced by GPT-3.5 and GPT-4 do not have significant differences.

<sup>10</sup>Examples used in the few-shot setting are available: <https://github.com/piccolab/Conspiratorial-Narratives-At-Scale>.

*whether the following text describes a conspiracy theory or not (yes/no). Justify your answer. “[post text]”*.

**3) Step-By-Step (SBS):** Guides the model in determining the label of the text by providing step-by-step instructions designed to replicate the decision-making process instructed to human coders during annotation, with the final step asking whether the post is CT or not. Prompt: *Decide whether the following text describes a conspiracy theory or not (yes/no). First, extract the narrative or claim from the text. Second, decide if the claim is a known conspiracy theory or suggests a hidden plan. Third, decide if the text agrees with or supports the conspiracy theory or plan. Fourth, answer the question (yes/no). “[post text]”*.

To assess GPT’s performance in few-shot settings, prompts are augmented with  $n$  examples, each paired with its respective ground-truth label. In each few-shot setting, a set of examples, both from CT and non-CT labeled samples, are selected based on their similarities to the text under consideration. We experiment with  $n = 0, 1, 3, 5$  pairs, where  $n = 0$  represents a zero-shot setting, and the few-shot settings provide the model with 2, 6, and 10 examples, ordered randomly to mitigate the influence of example arrangement (Liu et al. 2021). We compute pairwise cosine similarity based on the text embeddings generated by the best-performing LLM models (Sec.6.1) in our experiments (i.e., RoBERTa, as reported in Sec.7.1). The most similar examples with respect to a given text were then extracted from the labeled data’s positive and negative samples. Throughout the experiments, the parameter max. token length is set as 1500 and the temperature is 0, which is a recommended value for classification tasks that effectively constrain randomness. Each experiment was repeated 10 times to provide robust results that accounted for GPT’s randomness.

## 6.3 Traditional Machine Learning Models (MLs)

We test commonly used supervised machine learning algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and eXtreme Gradient Boosting (XGB). The implementation of these models was executed using Python’s scikit-learn library. For XGB, we utilized the XGBoost package and conducted hyperparameter optimization through grid search. All these ML models are trained with various text embeddings, including all the LLMs described in Sec.6.1.

# 7 Results

## 7.1 Detection Performance

Table 3 presents the Precision, Recall, F1, and AUC of the Machine Learning classifications with the best text embeddings, and Deep Learning models. SVM consistently demonstrate strong performance across all text embeddings. LR exhibits high Recall, but relatively lower Precision, resulting in lower F1. Notably, RF assigns all samples as positive, thereby achieving perfect Recall but compromised Precision. In terms of text embeddings, ALBERT, the most compact model among those tested, exhibits relatively lower performance. In contrast, DistilBERT outper-

Model	Precision	Recall	F1	AUC
ML Models				
DT+RoBERTa	57.6%	52.1%	0.497	0.605
RF+RoBERTa	35%	100%	0.519	0.5
LR+RoBERTa	42.4%	79%	0.553	0.609
KNN+T5	54%	69.6%	0.608	0.689
SVM+T5	79.1%	54.7%	0.647	0.735
XGB+T5	72.5%	66.2%	0.692	0.763
DL Models				
T5	33.8%	96.5%	0.5	0.5
ALBERT	58.9%	54.3%	0.565	0.673
DistilBERT	63.9%	69.1%	0.664	0.745
BERT	66.7%	70%	0.673	0.752
DeBERTa	66.1%	71.5%	0.687	0.762
RoBERTa	70%	73.8%	<b>0.714</b>	<b>0.787</b>

Table 3: Performance metrics of the ML and DL classification models. Best performances in bold.

forms both BERT and DeBERTa, despite its smaller size owing to distillation. RoBERTa and T5 consistently deliver superior performance across all evaluated metrics.

Among the Deep Learning models, RoBERTa emerges as the top performer, recording the highest F1, AUC, and Precision. Following closely are DeBERTa, BERT, and DistilBERT, which produce comparable results. ALBERT, although exhibiting lower performance, outperforms T5, which surprisingly ranks as the least effective model. We posit that T5’s underperformance may be attributed to the need for a larger training dataset and further exploration of hyperparameter optimization.

## 7.2 Analysis of BERT-Based Models’ Results

To assess the precision of BERT-based models, particularly the top-performing RoBERTa (with an AUC of 0.787), we conduct a qualitative error analysis. While RoBERTa exhibited superior performance compared to other models, it inevitably encountered challenges in accurately labeling certain samples. This section presents instances of both false positives and false negatives. Note that in our experiment, GPT misclassified all the following examples.

**False Positive** We observed that the RoBERTa model has a tendency to label a given sample as positive when it contains a known CT, irrespective of the post author’s intended sentiment.

**I1** FP (Label: No; RoBERTa: Yes) *“This sub has become cancer. It is a dumping ground for Qanaon retards and pro Biden shills. Biden and Trump are equally evil. Trump is the vax daddy. Biden diddles kids if you deny either youre blissfully ignorant.”*

**False Negative** We observed that the RoBERTa model has limitations in identifying new and emerging narratives. For example, the author in I2 below constructs a novel narrative on gene alteration implying it is a made-up plan.

**I2** FN (Label: Yes; RoBERTa: No) *“I cant believe Im watching the last battle of current humanity. If vaccines*

Setting	Precision	Recall	F1	AUC
Simple				
0-shot	<b>72.50%</b>	72.60%	<b>0.726</b>	<b>0.795</b>
1-shot	65.60%	76.90%	0.708	0.785
3-shot	67.10%	77.00%	0.717	0.792
5-shot	67.40%	73.00%	0.7	0.777
Justification				
0-shot	69.11%	75.81%	0.723	<b>0.795</b>
1-shot	66.40%	75.20%	0.705	0.782
3-shot	61.50%	83.00%	0.706	0.786
5-shot	58.50%	<b>86.10%</b>	0.695	0.778
SBS				
0-shot	66.00%	64.00%	0.653	0.736
1-shot	67.00%	76.00%	0.713	0.788
3-shot	63.20%	79.20%	0.703	0.782
5-shot	60.80%	81.00%	0.694	0.776

Table 4: Performance metrics of GPT under different settings. Best performances in bold.

*are gene therapy or some gene altering something, well be the last human beings of old world. New generations all will have different genes. Mutated, damaged, who knows?And if microchipdigital currency replaces current monetary system, its end of old system too.”*

## 7.3 Analysis of GPT’s Results

Table 4 provides an overview of GPT’s performance across various experimental settings, each experiment was repeated 10 times to alleviate GPT’s randomness. The Standard Deviation across all metrics ranges between 0.002 and 0.042. Notably, the “Simple” prompt setting slightly outperforms both the “Justification” and “Step-By-Step” settings. However, the absence of reasoning in this setting casts doubt on the final verdict, particularly considering the alternation of labels when justification was requested.

Furthermore, our findings reveal that additional context, presented in few-shot setting, has a negligible effect on F1 and tends to sacrifice Precision in favor of Recall. This is consistent with previous works (Zhao et al. 2021; Reynolds and McDonell 2021). For instance, Reynolds and McDonell (2021) argued that in-context examples help in task identification rather than learning, particularly in reasoning-intensive tasks. This finding potentially interprets SBS results, where additional examples clarify detailed instructions and contribute to performance enhancement.

The results of the Justification settings offer a pathway for further investigating whether GPT’s labels were produced with proper rationales. To see this, we conducted a theme analysis on GPT’s results under the Justification-zero-shot setting (which yielded the best AUC).

We grouped all the annotated samples into true/false positive/negative (TP/FP/TN/FN) according to GPT’s predicted labels, and then manually reviewed all generated justifications. We summarize the five most common patterns in GPT’s justifications with illustrative examples (labeled as S1–S8). Note that the prediction of our best classifier are

consistent with the ground-truth labels in all of the following examples.

**Correct reasoning** GPT performs well in detecting conspiratorial text relating to well-known conspiracy theories such as COVID-19, UFOs, the New World Order, elites, climate change, and CTs seen in the political domain. The justifications demonstrated its ability to connect existing resources to recognize CTs in a given post. In some instances, it is able to identify less common CTs (S1) and distinguish benign CT-related text (S2).

**S1** TP (Label: Yes; GPT: Yes) “*Dr. Strangelove was right, except with chlorine instead of fluoride. Semen is not supposed to smell like bleach, chlorine and chlorides are getting absorbed when you bathe in/drink/inhale it. So your ‘precious bodily fluids’ is indeed being attacked.*” **GPT reason:** “*The text suggests that there is a deliberate attack on people’s bodily fluids through the absorption of chlorine and chlorides when bathing, drinking, and inhaling. It also references Dr. Strangelove, a fictional movie character known for his conspiracy theories. The claim that semen is not supposed to smell like bleach adds to the conspiratorial nature of the text.*”

**S2** TN (Label: No; GPT: No) “*Does Epstein have any links to a government intelligence agency? I’ve seen a lot of talk about this but not any actual links.*” **GPT reason:** “*The text is a question seeking information about whether there are any known links between Jeffrey Epstein and a government intelligence agency. It does not make any specific claims or promote a conspiracy theory; instead, it appears to be an inquiry seeking clarification or evidence regarding a topic of interest.*”

**Fail: Criticism** In some instances, GPT overinterpreted the text as conspiratorial based on keywords such as “media” (S3) that may have appeared in other CTs. We observed, however, that these cases were not intended to promote any CTs, but rather to provoke ideas, share opinions, or provide criticisms in a different context.

**S3** FP (Label: No; GPT: Yes) “*The corruption in science and the censorship in the media is one and the same: peer review. The notion that ideas derive their truth from the opinion of a minority able to control the public discourse, rather the idea’s factuality, or merit. That’s why science has been unable to produce major breakthroughs for over a century. Original thought never comes through consensus, but from fringe minds.*” **GPT reason:** “*The text suggests that there is corruption in science and censorship in the media, and attributes this to peer review. It claims that a minority is able to control public discourse and that science has been unable to produce major breakthroughs for over a century because original thought does not come from consensus but from fringe minds. These claims lack substantial evidence and rely on the assumption that there is a deliberate effort to suppress original ideas and control the scientific and media landscape. This aligns with the characteristics of a conspiracy theory, which typically involves the belief in secret plots or hidden agendas by powerful groups.*”

**Fail: Sarcasm** We also discovered that GPT was incapable of recognizing sarcasm and falsely interpreted the author of a post as believing in a situation that the author intended to ridicule (S4).

**S4** FP (Label: No; GPT: Yes) “*I think we should just humor Q and MAGA and agree with them that DJT is still president and still calling the shots via FEMA etc. What harm could it do? Next election when Kamala and AOC are elected we will tell the the same thing that DJT is actually the real president. Why not?*” **GPT reason:** “*The author believes that Donald Trump is still in charge and making decisions through FEMA (Federal Emergency Management Agency) despite Joe Biden being the current president. This belief goes against the accepted reality of the current political situation and presents an alternative narrative without any substantial evidence. The idea that Trump is secretly controlling the government and will continue to do so even after the next election is not supported by factual information.*”

**Fail: Non-existent connection (causal hallucination)** In cases where the text is limited (i.e., a few words or a single sentence), GPT appears to relate the content subjects to other distant materials, such as suspecting an action or event as conspiratorial based on the position of a mentioned person (S5) or source (S6), despite the fact that no such connection exists in the text, and no evidence for the authors’ intent to support any conspiratorial beliefs.

**S5** FP (Label: No; GPT: Yes) “*Michael Flynn Calls for Trump to Declare Martial Law.*” **GPT reason:** “*The statement suggests that Michael Flynn is calling for Trump to declare martial law, which implies a potential conspiracy theory. The idea of a former high-ranking military officer advocating for martial law can be seen as a controversial and potentially conspiratorial claim. However, without further context or evidence, it is difficult to determine the veracity of the statement.*”

**S6** FP (Label: No; GPT: Yes) “*Reporter Who Broke Clinton-Lynch Tarmac Story And Received Significant Death Threats Found Dead. See Infowars (not linking to it in Reddit).*” **GPT reason:** “*The statement suggests that the reporter who broke the Clinton-Lynch Tarmac story and received death threats has been found dead. However, without any further evidence or verifiable sources, it is difficult to determine the truthfulness of this claim. The mention of Infowars also raises suspicions, as Infowars is known for promoting conspiracy theories.*”

**Fail: Implicit connection or rhetorical question** In contrast to its tendency to overinterpret or misuse subjects that appear in the text (e.g., keywords, well-known officials, and media outlets), GPT is incapable of recognizing more nuanced communication cues, such as an implicit attribution (S7) or a rhetorical question (S8), which are frequently observed in online texts. In the next two examples, we provide human reason due to the subtlety of these cues.

**S7** FN (Label: Yes; GPT: No) “*Turkey rocked by 7M Quake hours after France attack; World leaders Outraged at Macron!*” **GPT reason:** “*The given text does not*



provide any evidence or support for a conspiracy theory. It mentions two separate events, a 7M earthquake in Turkey and an attack in France, along with world leaders being outraged at Macron. While the text may imply a connection between these events, it does not present any conspiracy theories or elaborate on any hidden agendas or secret plots. Therefore, it cannot be considered a conspiracy theory.” Human reason: “The author connects two unrelated events, which indicates that the first is caused by the second. This argument is supported by the last sentence of the post, which states that world leaders are outraged at France accusing them of committing the earthquake.”

**S8** FN (Label: Yes; GPT: No) “How does marxism/communism take over so many countries? It happened to Europe, the UK, Russia and is now happening to the US. How are they able to do this to every country while so many people don’t even realize it?” GPT reason: “The statement does not present any conspiracy theories. It discusses the spread of Marxism/Communism to various countries and questions how it has been able to happen without many people realizing it. While it may touch upon political ideologies and their influence, it does not propose any hidden or secretive plots, making it not a conspiracy theory.”

Human reason: “The author addresses the spread of Marxism/communism in a systematic method. The question is rhetoric and implies an affirmative tone. This aligns with the known CT ‘Cultural Marxism’, which talks about the spread of Marxism in Europe, and extends it to the US.”

Based on our analysis of GPT’s classification reasoning, it was found that, in certain cases, GPT can accurately identify the narrative elements that support its decision to classify a text as either CT or non-CT. However, there are instances where it misuses information both within and outside the text, and fails to recognize the subtle dialectic cues often present in informal communications. While GPT has been shown useful in other domains, such as data augmentation (Møller et al. 2023), our study suggests that it should be used with caution in the context of CTs.

## 7.4 Prevalence of CT Narratives in CT-Subreddits

We investigate the prevalence of CT narratives within the most active online conspiracy subreddits as described in Sec. 4. We employ the same filtering procedures outlined in Sec.5.1, which include removing self-deleted, banned, and short posts. We then use our best-trained classifier, i.e. RoBERTa, to classify each of the posts into CT or non-CT. Table 5 lists the positive ratios, i.e., proportions of conspiratorial narratives, for each of the 12 subreddits. We estimate the upper and lower bounds of the positive ratios based on the Precision and Recall of the classifier, as well as the detected positive ratios and sample size. Specifically, the upper bound is estimated by assuming that all detected positives are true positives, and the lower bound is estimated by assuming the detected positives contain no false nega-

tives. We observe a wide range of positive ratios, from 20% to 46.5%. It is important to note that these ratios may be affected by the distinct focus of each subreddit as well as the moderation in place. The top three subreddits with the highest positive ratios are r/1984isreality (46.5%), r/conspiracyundone (41.9%), and r/TopConspiracy (40.5%). The ratio of all subreddit posts, 31.3%, corresponds closely to the ratio observed in the annotated subset, 33%.

Subreddit	Posts	Pos. Ratio	Upper Bound	Lower Bound
conspiracy	201054	0.312	0.422	0.218
TruthLeaks	872	0.279	0.377	0.195
TopConspiracy	311	0.405	0.547	0.284
conspiracy_commons	12941	0.321	0.434	0.225
climateskeptics	3078	0.235	0.318	0.165
conspiracytheories	11379	0.337	0.455	0.236
DescentIntoTyranny	121	0.273	0.369	0.191
ConspiracyII	1059	0.355	0.480	0.249
FringeTheory	385	0.200	0.270	0.140
conspiracyundone	1420	0.419	0.566	0.293
C.S_T	4974	0.318	0.430	0.223
1984isreality	43	0.465	0.628	0.326
Overall	237637	0.313	0.423	0.219

Table 5: Classification results of posts in each subreddit. The second column shows the number of posts (CT and non-CT) considered in this study after filtering.

We further analyze the differences of CT and non-CT posts in terms of audience reactions they received. We use two measures for the reactions: (a) number of comments, and (b) karma scores, defined as the number of upvotes minus the number of downvotes, or zero if the difference is negative. We compare the distributions of the two measures in our dataset. Fig. 1 shows the empirical cumulative distribution function, or eCDF, for the two measures. As both measures have skewed distributions, the Mann-Whitney U test is used to determine whether the two distributions are significantly different. Overall, we found that CT posts tend to receive more comments than non-CT ( $p < 10^{-6}$ ) and greater karma scores ( $p < 10^{-6}$ ).

These findings indicate that a CT post typically receives greater engagement and a higher ratio of upvotes to downvotes. On the one hand, this reflects the nature of these subreddits, where users tend to be more interested in CTs; on the other hand, it raises concerns that posts with CTs, with more interactions received, are likely to be promoted by platform algorithms, and users who post conspiratorial narratives may accumulate karma scores to spread their content more easily than those who do not (e.g., actively debunk CTs).

## 7.5 Cross-Domain Test

To assess the efficacy of our method on samples from different domains, we employ our top-performing model, RoBERTa, on a Twitter dataset obtained from (Phillips, Ng, and Carley 2022) and conduct a comparative analysis. In their study, a two-step approach was adopted: the first step detects the presence of a CT in a tweet, while the second step

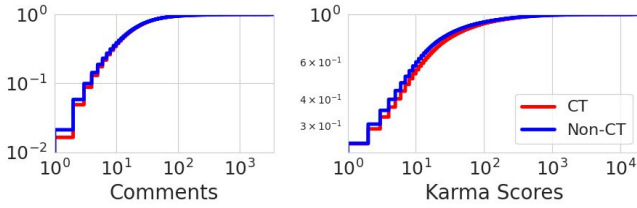


Figure 1: Disparities of comments (left) and karma scores (right) distributions between CT and non-CT posts, represented by the eCDF.

classifies the tweet into one of three categories—Against, Neutral, and Supportive. In our work, we consolidated the first step and the first two categories into non-CT, making the Supportive classification directly comparable to CT content only. The results are presented in Table 6.

It is worth noting that our classifier achieved a high F1 score despite not receiving any training or fine-tuning on this Twitter dataset. In fact, the score was only slightly lower than the authors’ best model which was specifically trained using the same dataset. Furthermore, our classifier demonstrated an impressive precision rate of 91%. However, it was expected to have a lower recall rate, given that the posting styles on the two platforms differ significantly. This suggests that our approach, which integrates attitudes towards a conspiracy theory into a unified classifier, makes it possible to identify conspiratorial posts more precisely, even across different social media platforms and topics.

Model	Precision	Recall	F1
Phillips, Ng, and Carley (2022)	71.1%	78.5%	0.746
Our RoBERTa	91%	56%	0.690

Table 6: Cross-Domain Comparison of CT Classifications

## 8 Discussions

This research examined the pervasive problem of detecting conspiracy theories in online discussions. We developed a comprehensive and general classification scheme that incorporates the theoretical and operational definitions of conspiracy theories, as well as deep learning and large language models, to distinguish conspiratorial narratives effectively. Our approach utilized human-labeled ground truth to train a BERT-based classifier. This classifier was subsequently used to examine the ratios of conspiratorial narratives in the most active conspiracy-related Reddit forums. Our research revealed that only one-third of these forums’ posts were classified as containing conspiracy theory narratives. This finding challenged previously held assumptions regarding the prevalence of conspiratorial narratives in such communities (Phadke, Samory, and Mitra 2021a; Samory and Mitra 2018; Klein, Clutton, and Dunn 2019; Engel et al. 2022; Pappasavva et al. 2021; Bessi et al. 2015; Zollo et al. 2017).

Our research also revealed that posts that promote CT narratives tend to receive more comments and higher karma scores, indicating that CT promoters may gain additional advantages that enable them to promote their content more

broadly. Based on our findings, platforms should consider a different promotion mechanism that takes into account the distinct nature of the online communities, and more sophisticated techniques should be integrated into content moderation to reduce the visibility of conspiratorial content.

In addition, our analysis showed that our classifier performs comparably to GPT, despite being trained on extensive text corpora and computational resources. Our qualitative analysis of GPT’s classification decision justifications revealed several alarming flaws, indicating its limited capacity for enhancing data quality or detection performance in the context of CT.

This study has some limitations. Firstly, the classification criteria aimed to detect both established CTs and content with the potential to evolve into such theories. However, the inherent ambiguity of authors’ intentions in online discussions, particularly in relation to conspiracy beliefs, made accurate categorization challenging. Second, despite our efforts to ensure consistency in annotation, different annotators may interpret specific samples differently, potentially introducing subjectivity into the annotation process. Lastly, while the extent of GPT’s uncontrollable randomness that drives the diverse outcomes may not significantly affect the average performance evaluation, it may pose a reproducibility challenge. Future research should consider methods that account for the variability of GPT’s responses in order to systematically improve its ability to perform tasks requiring a substantial amount of contextual understanding.

## 9 Ethical Considerations

**Data Collection and Management.** The data used in this study were collected through the Pushshift Reddit API. We used the Pushshift API to access publicly available subreddits and posts that users chose to make public on Reddit. We also collected publicly viewable metrics related to publicly posted content (e.g., a post’s “comment” count and “score” value). All of the information we collected has been shared publicly on the platform with an unrestricted audience. Starting June 19, 2023, access to data via third-party services was limited per Reddit’s introduction of new Data API Terms.<sup>11</sup> As of the time this paper was written, Pushshift only provided restricted access.<sup>12</sup> We adhere to Reddit’s new policy and the platform’s data usage guidelines,<sup>13</sup> which state that the data may only be used for research purposes, and we will not redistribute our data or any derivative products or services based on our data (e.g. models trained using Reddit data) without further permission from Reddit. Prior to publishing our work, we will seek Reddit’s permission for data usage and possible academic sharing.

**User Anonymity and Privacy.** We had no direct interaction with Reddit users and gathered no private information about them. As outlined in Sec. 4, we excluded the self-deleted or banned posts in our analysis. Consequently, our

<sup>11</sup><https://www.redditinc.com/policies/data-api-terms>

<sup>12</sup>[https://www.reddit.com/r/pushshift/comments/13w6j20/advancing\\_communityled\\_moderation\\_an\\_update\\_on/](https://www.reddit.com/r/pushshift/comments/13w6j20/advancing_communityled_moderation_an_update_on/)

<sup>13</sup>[https://support.reddithelp.com/hc/en-us/articles/14945211791892#h\\_01H69EJB9GRHCMPZMKFQTNQKY0](https://support.reddithelp.com/hc/en-us/articles/14945211791892#h_01H69EJB9GRHCMPZMKFQTNQKY0)

study may be biased and may have missed some of the most harmful narratives. Our study results were either presented anonymously or in a summary format, with no user-specific information disclosed. While we included a few Reddit posts as examples, we took measures to ensure the anonymity of these sources. Specifically, we conducted a comprehensive search (using the keyword “Reddit” as well as the post’s title and content on a popular search engine) to ensure that the user information associated with the presented examples cannot be easily recovered.

**Content Credibility.** Several examples of conspiracy narratives are presented in this paper for illustrative purposes, which poses a risk that some readers may consider them credible. Even though some CTs turn out to be true, we emphasize that a vast number of CTs are not credible (Sunstein and Vermeule 2009) – they are unproven, misleading, or lack empirical evidence. We caution our readers to read these examples with skepticism and critical thinking.

**Annotation Complexity and Subjectivity.** In accordance with our proposed coding scheme, a comprehensive, topic-independent annotation dataset has been generated, which can be utilized by researchers in the future. However, we acknowledge that the annotation process is inherently subjective, and human errors may arise due to its complexity. For instance, determining the attitude of a post-author can be challenging, particularly when authors deliberately conceal their attitudes. Despite striving to reach a consensus among our well-trained annotators, it is acknowledged that other researchers may arrive at different decisions.

**Potential Use/Abuse of Work.** This study aimed to understand the feasibility and limitations of identifying online conspiracy narratives. While our work contributed to ways of countering the spread of conspiracy theories and promoting a more healthy public discourse, there is a possibility that individuals who intend to disseminate CTs may take advantage of our study outcome to bypass automatic detection mechanisms. By acknowledging this limitation, we underscore the importance of ongoing research and vigilance in refining automated detection methods to safeguard against potential abuse of new research outcomes.

## Acknowledgements

The authors would like to acknowledge support from AFOSR, ONR, Minerva, NSF #2318461, Collaboratory Against Hate Research and Action Center, and Pitt Cyber Institute’s PCAG awards. The research was also partially supported by Pitt’s CRC resources (RRID:SCR 022735 through NIH #S10OD028483). Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

## References

- Albertson, B.; and Guiler, K. 2020. Conspiracy theories, election rigging, and support for democratic norms. *Research & Politics*, 7(3): 2053168020959859.
- Allington, D.; Duffy, B.; Wessely, S.; Dhavan, N.; and Rubin, J. 2021. Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological medicine*, 51(10): 1763–1769.
- Bessi, A.; Coletto, M.; Davidescu, G. A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2): e0118093.
- Bleakley, P. 2023. Panic, pizza and mainstreaming the alt-right: A social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology*, 71(3): 509–525.
- Bond, B. E.; and Neville-Shepard, R. 2023. The rise of presidential eschatology: conspiracy theories, religion, and the January 6th insurrection. *ABS 2023*, 67(5): 681–696.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS 2020*, 33: 1877–1901.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Engel, K.; Hua, Y.; Zeng, T.; and Naaman, M. 2022. Characterizing reddit participation of users who engage in the qanon conspiracy theories. *CSCW 2022*, 6(CSCW1): 1–22.
- Farhart, C. E.; Douglas-Durham, E.; Trujillo, K. L.; and Vitoriol, J. A. 2022. Vax attacks: How conspiracy theory belief undermines vaccine support. *Progress in Molecular Biology and Translational Science*, 188(1): 135–169.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Hoseini, M.; Melo, P.; Benevenuto, F.; Feldmann, A.; and Zannettou, S. 2023. On the globalization of the QAnon conspiracy theory through Telegram. In *WebSci 2023*, 75–85.
- Introne, J.; Korsunskaya, A.; Krsova, L.; and Zhang, Z. 2020. Mapping the narrative ecosystem of conspiracy theories in online anti-vaccination discussions. In *SMSociety 2020*, 184–192.
- Kim, S.; and Kim, J. 2023. The information ecosystem of conspiracy theory: Examining the QAnon narrative on Facebook. *CSCW 2023*, 7(CSCW1): 1–24.
- Klein, C.; Clutton, P.; and Dunn, A. G. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit’s conspiracy theory forum. *PloS one*, 14(11): e0225098.
- Klein, C.; Clutton, P.; and Polito, V. 2018. Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in psychology*, 189.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS 2022*, 35: 22199–22213.
- Kou, Y.; Gui, X.; Chen, Y.; and Pine, K. 2017. Conspiracy talk on social media: collective sensemaking during a public health crisis. *CSCW 2017*, 1(CSCW): 1–21.
- Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

- Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Miani, A.; Hills, T.; and Bangerter, A. 2021. LOCO: The 88-million-word language of conspiracy corpus. *Behavior research methods*, 1–24.
- Miani, A.; Hills, T.; and Bangerter, A. 2022. Interconnectedness and (in) coherence as a signature of conspiracy worldviews. *Science Advances*, 8(43): eabq3668.
- Møller, A. G.; Dalsgaard, J. A.; Pera, A.; and Aiello, L. M. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.
- Mompelat, L.; Tian, Z.; Kessler, A.; Luetgen, M.; Rajanala, A.; Kübler, S.; and Seelig, M. 2022. How “loco” is the LOCO corpus? annotating the language of conspiracy theories. In *LAW-XVI 2022*, 111–119.
- Nerghes, A.; Kerkhof, P.; and Hellsten, I. 2018. Early public responses to the Zika-virus on YouTube: Prevalence of and differences between conspiracy theory and informational videos. In *WebSci 2018*, 127–134.
- Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and Cristofaro, E. D. 2021. “Is it a coincidence?”: An exploratory study of QAnon on Voat. In *WWW 2021*, 460–471.
- Pasquetto, I. V.; Olivieri, A. F.; Tacchetti, L.; Riotta, G.; and Spada, A. 2022. Disinformation as Infrastructure: Making and maintaining the QAnon conspiracy on Italian digital media. *CSCW 2022*, 6(CSCW1): 1–31.
- Phadke, S.; Samory, M.; and Mitra, T. 2021a. Characterizing social imaginaries and self-disclosures of dissonance in online conspiracy discussion communities. *CSCW 2021*, 5(CSCW2): 1–35.
- Phadke, S.; Samory, M.; and Mitra, T. 2021b. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *CSCW 2021*, 4(CSCW3): 1–30.
- Phillips, S. C.; Ng, L. H. X.; and Carley, K. M. 2022. Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In *Companion Proceedings of the Web Conference 2022*, 876–880.
- Platt, A.; Brown, J.; and Venske, A. 2022. Toward Detecting Conspiracy Language in Misinformation Documents. In *SIGMIS-CPR 2022*, 1–4.
- Pogorelov, K.; Schroeder, D. T.; Filkuková, P.; Brenner, S.; and Langguth, J. 2021. Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *HT OASIS 2021*, 21–25.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR 2020*, 21(1): 5485–5551.
- Reynolds, L.; and McDonell, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI EA 2021*, 1–7.
- Romer, D.; and Jamieson, K. H. 2020. Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social science & medicine*, 263: 113356.
- Samory, M.; and Mitra, T. 2018. ‘The Government Spies Using Our Webcams’ The Language of Conspiracy Theories in Online Discussions. *CSCW 2018*, 2(CSCW): 1–24.
- Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Shahsavari, S.; Holur, P.; Wang, T.; Tangherlini, T. R.; and Roychowdhury, V. 2020. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *JCSS 2020*, 3(2): 279–317.
- Sunstein, C. R.; and Vermeule, A. 2009. Conspiracy theories: Causes and cures. *Journal of political philosophy*, 17(2): 202–227.
- Swami, V.; and Furnham, A. 2014. 12 Political paranoia and conspiracy theories. *Power, politics, and paranoia: Why people are suspicious of their leaders*, 218.
- Tangherlini, T. R.; Shahsavari, S.; Shahbazi, B.; Ebrahimzadeh, E.; and Roychowdhury, V. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLoS one*, 15(6): e0233879.
- Van Prooijen, J.-W.; and Douglas, K. M. 2017. Conspiracy theories as part of history: The role of societal crisis situations. *Memory studies*, 10(3): 323–333.
- Vegetti, F.; and Littvay, L. 2022. Belief in conspiracy theories and attitudes toward political violence. *IPSR/RISP 2022*, 52(1): 18–32.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS 2022*, 35: 24824–24837.
- Wood, M. J.; and Douglas, K. M. 2015. Online communication as a window to conspiracist worldviews. *Frontiers in psychology*, 6: 836.
- YING, D.; and Jiang, J. 2015. Towards opinion summarization from online forums. In . *ACL*.
- Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML 2021*, 12697–12706. PMLR.
- Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.
- Zollo, F.; Bessi, A.; Del Vicario, M.; Scala, A.; Caldarelli, G.; Shekhtman, L.; Havlin, S.; and Quattrocioni, W. 2017. Debunking in a world of tribes. *PLoS one*, 12(7): e0181821.
- Zonis, M.; and Joseph, C. M. 1994. Conspiracy thinking in the Middle East. *Political Psychology*, 443–459.

## 10 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?

Answer: Yes, and we verified that our research 1) does not violate privacy norms or unfair profiling (Sec.9, under “User Anonymity and Privacy”), 2) is not tied to any socio-economic factors, 3) does not imply disrespect to societies or cultures (Sec.9, under “Content Credibility”).

- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: Yes, and as described in the abstract, we introduce a comprehensive scheme (Sec.3 and 5) for classifying conspiracy theories on Reddit (Sec.4), we leverage BERT-based models to build conspiracy theory classifiers and compare their performance to Generative Pre-trained Transformer (GPT) (Sec.6 and 7) while identifying its shortcomings (Sec.7.2). We also present the first large-scale classification study of conspiracy-related Reddit forums (Sec.7.3)

- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made?

Answer: Yes. We discussed the limitations of existing methods (Sec.1 and 2) and proposed a classification scheme followed by hired and trained human annotators (Sec.3, and 5). We targeted the most popular online forums. We employed a diverse set of machine learning techniques, including both traditional and advanced deep learning methods, along with the GPT (Sec.6). In particular, these models have been extensively explored in prior research, showcasing their remarkable expressiveness and aptitude for understanding textual language. This solid foundation in the literature substantiates our rationale for incorporating these models into our study.

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions?

Answer: Yes, and it can be found in “Ethical Consideration” (Sec.9) section, under “Data Collection and Management”.

- (e) Did you describe the limitations of your work?

Answer: Yes, and it can be found in the “Discussion” (Sec.8) and “Ethical Consideration” (Sec.9) sections.

- (f) Did you discuss any potential negative societal impacts of your work?

Answer: Yes, and it can be found in “Ethical Consideration” (Sec.9) section, under “Content Credibility” and “Potential Impacts of Work”.

- (g) Did you discuss any potential misuse of your work?

Answer: Yes, and it can be found in “Ethical Consideration” (Sec.9) section, under “Potential Impacts of Work”.

- (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?

Answer: Yes, and it can be found in “Ethical Consideration” (Sec.9) section, under “User Anonymity and Privacy” and “Potential Use/Abuse of Work”.

- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?

Answer: Yes, and we adhere to the guidelines, as demonstrated in Sec.9.

### 2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results?

Answer: NA

- (b) Have you provided justifications for all theoretical results?

Answer: NA

- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results?

Answer: NA

- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study?

Answer: NA

- (e) Did you address potential biases or limitations in your theoretical framework?

Answer: NA

- (f) Have you related your theoretical results to the existing literature in social science?

Answer: NA

- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain?

Answer: NA

### 3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results?

Answer: NA

- (b) Did you include complete proofs of all theoretical results?

Answer: 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)?

Answer: No, because access to data via third-party services was limited per Reddit’s introduction of new Data API (Please refer to Sec.9, under “Data Collection and Management”). We adhere to Reddit’s new policy and the platform’s data usage guidelines.

- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?  
 Answer: Yes, and the details can be found in Sec.6.
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?  
 Answer: No, due to space constraints. While we conducted thorough experiments encompassing various Machine Learning models with diverse text embeddings, we were constrained by space and thus could only present the best result for each tested setting in Sec.7. However, most of the experiments listed in the paper utilized default parameters, and our hyperparameter search was limited, as stated in Sec.6.
- (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)?  
 Answer: Yes, and it can be found in Sec.6.
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made?  
 Answer: Yes, and it can be found in Sec.7.
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance?  
 Answer: Yes, and it can be found in Sec.7.2.
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?  
 Answer: Yes, we have provided proper citations for all the used assets and models.
- (b) Did you mention the license of the assets?  
 Answer: No, because the images and tables featured in the paper were generated by the authors. Any software and reused codes utilized in this study are open source.
- (c) Did you include any new assets in the supplemental material or as a URL?  
 Answer: No, because we do not provide supplement materials. Publishing our assets is subject to Reddit’s permission (please refer to answer 4.a in this checklist and Sec.9).
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating?  
 Answer: Yes, and it can be found in Sec.9 under “Data Collection and Management”.
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?  
 Answer: Yes, and it can be found in Sec.9 under “Data Collection and Management”.
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR?  
 Answer: No, because we do not share the dataset at the time (please refer to answer 4.a in this checklist and Sec.9). Our dataset includes metadata that adheres to the principles of FAIR. It will be made available to the public following approval from Reddit.
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset?  
 Answer: Yes, and it will be published along with the dataset (subject to Reddit’s permission, refer to Sec.9).
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?  
 Answer: NA
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals?  
 Answer: NA
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?  
 Answer: NA
- (d) Did you discuss how data is stored, shared, and de-identified?  
 Answer: NA