Conspiracy Theories and Where to Find Them on TikTok

Francesco Corso[1,2], Francesco Pierri[1], Gianmarco De Francisci Morales[2]

[1]DEIB, Politecnico di Milano, Milan, Italy
[2]CENTAI, Turin, Italy
Correspondence: francesco.pierri@polimi.it, gdfm@acm.org

Abstract

TikTok has skyrocketed in popularity over recent years, especially among younger audiences. However, there are public concerns about the potential of this platform to promote and amplify harmful content. This study presents the first systematic analysis of conspiracy theories on TikTok. By leveraging the official TikTok Research API we collect a longitudinal dataset of 1.5M videos shared in the U.S. over three years. We estimate a lower bound on the prevalence of conspiratorial videos (up to 1000 new videos per month) and evaluate the effects of TikTok's Creativity Program for monetization, observing an overall increase in video duration regardless of content. Lastly, we evaluate the capabilities of state-of-the-art open-weight Large Language Models to identify conspiracy theories from audio transcriptions of videos. While these models achieve high precision in detecting harmful content (up to 96%), their overall performance remains comparable to fine-tuned traditional models such as RoBERTa. Our findings suggest that Large Language Models can serve as an effective tool for supporting content moderation strategies aimed at reducing the spread of harmful content on TikTok.

1 Introduction

Online social platforms have become integral to modern life by facilitating connections, sharing of interests, and the formation of friendships. Over the past decade, social media have also emerged as the primary source of news and information for the public.[1] TikTok, in particular, has become a key platform for online news consumption, with nearly half of TikTok users (17% of all U.S. adults) regularly relying on it for news and information.[2]

TikTok's vast user base and viral content-sharing mechanisms make it a hotbed for malicious content, by facilitating the rapid spread of harmful narratives and dangerous information (Zeng and Kaye, 2022; Weimann and Masri, 2023) while also providing monetary incentives for viral content creators, such as TikTok's Creativity Program.[3] It is thus essential to ensure a safe environment for users by monitoring and mitigating the impact of malicious actors. Indeed, researchers have recognized several threats to discourse quality, including disinformation and hate-speech campaigns, political propaganda, and the spread of conspiracy theories (Aïmeur et al., 2023; Matamoros-Fernández and Farkas, 2021; Chaudhari and Pawar, 2022).

In this work, we study the presence of online conspiracy theories on TikTok, which are defined as false or unverified narratives that allege secret, often sinister plots or events involving groups, organizations, or governments (Byford, 2011). The risks of conspiracy theories stem from their potential to warp the perception of reality for users who consume them (Uscinski et al., 2017). Online social media offer a fertile ground for the spread of conspiracies, also driven by the formation of fringe communities (Spohr, 2017; Rollo et al., 2022). These theories can range from relatively harmless speculations to more dangerous beliefs, which fuel misinformation and polarization, and can have dangerous real-world consequences.[4] The rise of Generative AI tools powered by Large Language Models (LLMs) further amplifies concerns about their potential misuse by malicious actors, as these models can generate content that may mislead and harm online users (Augenstein et al., 2024). At the same time, LLMs offer valuable opportunities for researchers and scientists to analyze large-scale social media data, identify patterns in the spread of misinformation, and understand the dynamics of conspiracy theories (Ziems et al., 2024).

[1]www.pewresearch.org accessed on 10/05/2025.
[2]www.pewresearch.org accessed on 10/05/2025.
[3]TikTok's Creativity Program accessed on 12/05/2025.
[4]www.theguardian.com accessed on 10/05/2025.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8346–8362

To date, only a few qualitative approaches have examined conspiracy theories on TikTok, including Weimann and Masri (2023), which explores far-right narratives, and Zenone and Caulfield (2022), which identifies monkeypox-related conspiracy content. Our work employs a quantitative and systematic approach to data collection and analysis to address the following research questions:

- **RQ1**: What is the prevalence of videos sharing conspiracy theories on TikTok?
- **RQ2**: Did the Creativity Program impact the supply of conspiratorial content on TikTok?
- **RQ3**: Can we leverage LLMs to detect conspiracy theories shared on TikTok?

Our contributions are as follows. To address **RQ1**, we collect a sample of 1.5M long videos (with a duration of at least 1 minute) shared on TikTok in the U.S. by approximately 1M unique users. Using a data-driven approach and a large-scale dataset of Web documents about conspiracy theories, we identify hashtags that provide strong signals for the presence of conspiratorial videos and estimate a lower bound to their prevalence on TikTok.

To answer **RQ2**, we measure a small yet significant increase in the average duration of videos shared after the introduction of the Creativity Program, which incentivizes creators to post long videos. While this finding indicates a shift in the overall content-production behavior on TikTok, we observe no specific effect on conspiratorial videos.

Finally, to address **RQ3**, we evaluate the ability of state-of-the-art, open-weight LLMs (Llama, Mistral, and Gemma) to detect conspiracy-promoting videos using audio transcriptions extracted via OpenAI Whisper. We use open-weight models to ensure a fully reproducible and controllable content moderation pipeline, independent of external inference services (Palmer et al., 2024). LLMs prove to be highly sensitive to prompt variations (Sclar et al., 2024) and, while they outperform our baseline in many configurations, they still exhibit limitations and trade-offs that must be carefully considered for real-world content moderation applications.

2 Related Work

The study of online conspiracy theories is broad and diverse, with researchers exploring this complex phenomenon from multiple perspectives. This section reviews relevant studies, emphasizing linguistic analyses, available datasets, and strategies for dealing with this phenomenon.

Douglas et al. (2019); Sutton and Douglas (2020) adopted a psychological approach to identify key traits of conspiracy theory believers, revealing common factors such as lower levels of income and education. Enders et al. (2023) examined the relationship between social media usage and the tendency to believe in conspiracy theories, finding that social networks alone are insufficient for effectively conveying conspiratorial messages. Other studies have examined the individuals who share and promote conspiratorial content, tracing their roles back to some of the earliest widely known conspiracy theories (Byford, 2011; Douglas and Sutton, 2023).

Research has also investigated the linguistic characteristics of conspiracy theories. Fong et al. (2021) and Meuer et al. (2023) analyzed the psychological nuances in the language of conspiracy theories found in online articles and Twitter posts. Additionally, several datasets have been developed to support researchers in studying conspiracy theories across social media and online-published articles (Pogorelov et al., 2021; Phillips et al., 2022; Miani et al., 2021; Golbeck et al., 2018).

On TikTok, studies have primarily taken a qualitative approach to analyzing harmful content, including conspiracy theories (Zenone and Caulfield, 2022; Li et al., 2021), by manually reviewing videos associated with specific hashtags. Weimann and Masri (2023) identified a consistent presence of far-right extremism in videos, commentary, symbolism, and imagery. Similarly, Basch et al. (2021) examined the top 100 videos related to COVID-19 vaccines and found that many discouraged vaccination. Beyond qualitative studies, a growing body of work focuses on automated moderation of harmful content in multimodal environments. Researchers have developed models that integrate text, audio, and visual signals to detect malicious online posts such as memes (Lin et al., 2024) or YouTube videos (Jo et al., 2024; Ribeiro et al., 2020), while others underline how using a combined approach can lower errors in detection pipelines (Wang et al., 2023).

3 Methods

3.1 Dataset

We use the TikTok Research API[5] to collect our dataset. At the time of writing, this service has

[5] https://developers.tiktok.com/products/research-api

severe restrictions on usage and availability. In particular, academics can request access to a quota of 1000 requests per day. Given the lack of reliable data returned by the API during 2018-2020 (Corso et al., 2024), we run our collection for the period 2021-2023. This process yields a total of 1 605 696 items, 1 494 831 of which are unique. These videos are created by a total of 1 178 303 unique users. In addition, we use the Research API to collect two random samples of videos (∼10k videos each) posted in the U.S. one month before and one month after the beginning of the new Creativity Program (May 3, 2023), respectively. Appendix A.3 provides more details on the data collection.

3.2 Conspiracy Hashtag Enrichment

Our objective is to identify videos featuring hashtags that explicitly indicate the conspiratorial nature of the content. To achieve this, we manually review the top 30 most frequent seeds from the LOCO conspiracy dataset (Miani et al., 2021), which includes over 90 000 conspiracy and non-conspiracy articles scraped from the Web. To minimize ambiguity and reduce noise, we exclude seeds with broader or more general semantics, such as *cancer*, *AIDS*, *coronavirus*, *5G*, and *barackobama*, as these terms often extend beyond conspiratorial contexts. For instance, the term *cancer* may appear in claims about 'cancer cures being hidden by pharmaceutical companies' but also in discussions of treatment options, awareness campaigns, and personal survivorship stories. Furthermore, this procedure aligns with our goal of establishing a reliable lower bound for the prevalence of conspiracy theory videos on TikTok. Thus, we focus on the 10 seeds (e.g., *illuminati*, *reptilians*, full list in Appendix A.2) that convey a clearer, more direct association with conspiracy theories to serve as the basis for our labeling and enrichment process.

To identify candidate hashtags associated with conspiracy theories, we start with ∼650k unique hashtags extracted from the 1.5M videos in our sample. We then filter out hashtags that appear only once in the dataset, reducing the total to 281 510. For each of these remaining hashtags, we compute the list of co-occurring hashtags and words, along with the number of videos in which the hashtag appears, obtaining the following co-occurrence matrices. A hashtag/hashtag matrix H with dimensions of $281\,510 \times 281\,510$, and a hashtag/word matrix W, with dimensions of $281\,510 \times 185\,973$, where each column represents a unique word in the corpus of video descriptions.

We compute the pairwise similarity between each seed hashtag and the rest of the hashtags, and keep the top 20 hashtags by similarity. We employ the method by Garimella et al. (2018) to calculate the pairwise similarity between hashtags:

$$sim(h_s, h_t) = \frac{\alpha \cos(W_s, W_t) + (1-\alpha)\cos(H_s, H_t)}{1+\log(df(h_t))}$$

The similarity between two hashtags (h_s, the seed hashtag, and h_t the target hashtag) is defined as the weighted sum of the cosine similarities between the respective word and hashtag co-occurrence vectors (W_i, H_i), with the mixing parameter α set to 0.3 based on previous work (Garimella et al., 2018). This sum is discounted by a function of the document frequency of the hashtag h_t, to lower the weight of those hashtags that are extremely frequent (e.g., *#fyp*, *#viral*). This process produces a total of 197 hashtags that can be used to find conspiratorial videos. Our choice of using hashtags to identify conspiratorial videos is grounded in the literature, as previous work has employed hashtags to assign distant labels in various research contexts, ranging from political polarization (Conover et al., 2011) to emotion detection (Hasan et al., 2014).

3.3 Evaluation of Conspiracy Hashtags

We manually validate each of the 197 enriched hashtags to identify those actually associated with conspiracy theories. For each hashtag, we manually inspect a sample of five videos that include it in their description to determine whether they reference conspiracy theories, and thus the hashtag should be classified as conspiratorial. This analysis identified different categories of hashtags. Out of 197 collected hashtags, we classify 92 of them as conspiratorial (**CT**), while we find 68 to be noisy or not explicitly related to conspiracy theories (**NOCT**). We find 28 so-called "dog whistling" (**DW**) (Quaranto, 2022) hashtags, i.e., hashtags that seem harmless at first but are used by conspiracy theorists as inconspicuous markers to attract other users interested in the topic being discussed. During manual validation, a hashtag is classified as "dog whistling" if the tagged videos reference conspiracy theories, but the hashtag's literal meaning differs from the context in the videos (e.g., *#radiowaves* related to weather modification and chemtrails). There are also a few cases of hashtag hijacking (Hadgu et al., 2013), in two different flavors. *Normal*, when conspiracy users use popular hashtags to boost the virality of their content, e.g., *#skylovers*, *#billieeilishmusic* (**HJ**). And *reverse*, when users add conspiratorial hashtags to disseminate debunking information about a specific conspiracy theory (**RHJ**).

Given these classes, we use a heuristic to label videos as conspiratorial if they contain at least one hashtag from the classes **CT** or **DW**, and exclude videos that have hashtags belonging to the remaining three classes (**NOCT, HJ, RHJ**). This process yields 1363 conspiracy theory videos.

We further assess the quality of this distant-labeling procedure by manually analyzing a sample of 200 random videos, 100 from the pool of videos classified as conspiratorial with the hashtag-based distant labeling, and 100 from the remaining ones. Cohen's kappa agreement (Cohen, 1960) between our manual labeling validation and the distant labels is 0.81, which indicates strong agreement.

3.4 Estimating the Number of U.S. Videos on TikTok

To address the first research question and establish a lower bound on the number of conspiracy videos on TikTok, we aim to estimate the true size of the population from which our sample was drawn. We use a simple version of the Good-Turing frequency estimator (Good, 1953) to obtain an estimate of the number of long videos on TikTok:

$$M = \frac{N}{1 - \frac{N_1}{K}}, \tag{1}$$

where N is the number of unique videos in our dataset, N_1 is the number of videos that appear only once, and K is the total number of videos in the sample. We compute this estimate for each month of the collection. This methodology accounts for the unseen portion of the population and provides a conservative yet reliable estimate of its minimum size, thus ensuring robustness in our analysis. We also employ a maximum likelihood estimation technique to verify the goodness of the Good-Turing estimator:

$$M = \frac{N}{1 - e^{\frac{-K}{M}}}, \tag{2}$$

where N is the number of unique videos collected and K is the totality of videos collected. Since Equation (2) describes a recursive equation that should converge, we compute 1000 iterations of the equation with the starting value of $M = 10$. The two estimators produce extremely similar results ($< 1\%$ difference, see Section 4).

3.5 Video Transcripts

We employ the `voice_to_text` (or **VTT**) field provided by the API as an input to the LLMs. The `voice_to_text` is a feature offered by TikTok that allows the user to visualize the dialogue or narration provided by the creator as text on the screen. It must be enabled by the content creator at the time of publishing the video.

Of the $\sim$1.5M videos collected, only around 70k have their transcripts available natively from the API as VTT. For this reason, we expand the available transcripts by using OpenAI Whisper (Radford et al., 2023) (`medium` size model). To verify the model's performance on our data, we download a set of videos from TikTok for which the `voice_to_text` field is available from the Research API, and use Whisper to extract the transcript. We then compare the model output both with the retrieved VTT and with a manual transcription of the video (for $\sim$100 videos). We use the WER (Word Error Rate) score (Wang et al., 2003) to evaluate the accuracy of the transcriptions. Both transcriptions (Whisper-generated and API provided) are fairly accurate (median WER $\sim$0.15), which makes OpenAI's Whisper a viable alternative to the native `voice_to_text` field when the latter is unavailable. We exclude the videos whose transcription is unintelligible (mostly due to having no spoken content or having overbearing music). Furthermore, we retain only English videos, i.e., those where more than 50% of their transcript contains English words. More details about the evaluation are available in Appendix A.7.

3.6 Large Language Models and Prompting Strategies

Our goal is to assess whether open-weight Large Language Models can detect conspiratorial content. We measure the classification performance of some of the most recent open models: LLama3 (8B)(Dubey et al., 2024), Mistral (7B) (Jiang et al., 2023), and Gemma (7B) (Team et al., 2024). We chose these models as they have a comparable number of parameters and size of the context window. Furthermore, they allow full reproducibility of the results, and their computational requirements are modest if compared to larger models. In addition to evaluating single LLMs as classifiers, we consider an ensemble decision using a majority voting strategy on the outputs of the three models. Our choice of relying on text-based models was driven by practical considerations. Since the TikTok Research API provides access to audio transcriptions but not to visual content, this makes transcription-based processing a more scalable strategy for real-world applications, also because multimodal models are more resource-intensive than text-only models (Fahad et al., 2025). We design a zero-shot binary classification task where items about conspiracy theories represent the positive class. Zero-shot prompting implies that the models are not given any examples before seeing the actual item to classify. We do not consider few-shot approaches, as prior literature has shown they do not provide any improvement in this task (Diab et al., 2024). We use prompts inspired by the ones described by Diab et al. (2024), which were devised to classify Reddit posts from `r/conspiracy` and other conspiracy-related subreddits. We use three different variants, as follows:

- **Simple**: *'Decide whether the given transcription of a video talks about a conspiracy theory or not (if yes output = 1/else output=0). Provide just your output, no justification.'*
- **With Definition**: We append at the beginning of the simple prompt the definition of conspiracy theory proposed by Douglas and Sutton (2023) (fully listed in appendix A.8). The definition is preceded by the following additional sentence: *'Given this definition of conspiracy theory:'*
- **Step-by-step**: we drop the last sentence of the simple prompt (*'Provide just your output, no justification'*) and we append the following text to elicit the chain-of-thought process (Wei et al., 2022) from the model: *'First, extract the narrative or claim from the text. Second, decide if the claim talks about a conspiracy theory. Third, answer the question (if yes output = 1/else output=0)'*

Each model thus is given the following input structure: 'Transcription:'{Video Transcription}. {Prompt}. We describe the input as a 'transcription of a video' rather than 'textual content of a post' to be faithful to its nature. We also use numerical binary labels instead of strings to have a 'ready-to-use' classification output by the model. Recall that the ground truth labels for the videos are obtained through the hashtag enrichment process combined with the *distant supervision* approach validated via a manual inspection of $\sim$1000 videos, as described in Section 3.3. To test the system in a variety of settings representative of the operational conditions that might be encountered in practice, we use LLMs to classify conspiracy theory videos in three configurations of the data sources available:

- **`C1`: Balanced dataset with distant labeling:** An ideal case with wide availability of positive-class videos ($1/2$ of the dataset);
- **`C2`: Unbalanced dataset with distant labeling:** A more realistic setting with a minority of positive examples ($1/8$) and an increased representation of the negative class;
- **`C3`: Unbalanced dataset with manual labeling:** The setting is the same as `C2`, but the labels of the positive class are obtained manually.

The balanced dataset `C1` consists of a total of 1666 entries, with 887 belonging to the positive class (conspiracy) and 779 assigned to the negative class (not conspiracy). The unbalanced datasets contain the same negative-class items as in the balanced setting. For `C2`, the positive class is a random sample of 100 positive examples from `C1`. Instead, in `C3` the 100 examples of the positive class are manually labeled, rather than using the hashtag-based distant supervision. Specifically, we watch the videos to assess whether they explicitly mention conspiracy theories, e.g., a video with a voice-over (whether a real human voice or a generated one) that describes how NASA lied about the 1969 moon landing. To account for randomness in the inference process, we run our pipeline with 25 different input seeds (see Appendix A.2). We keep the `temperature` parameter of the models fixed to 0 to enhance reproducibility (Törnberg, 2024). All these experiments were run on a local machine with a 16-core CPU and a Nvidia A6000 GPU with 48 GB of VRAM. The time required to run the experiments is approximately 10 hours.

4 Results

4.1 Lower Bound on the Number of Conspiracy Videos on TikTok

Figure 1 presents the estimated monthly count of long videos shared on TikTok in the U.S., derived via the Good-Turing estimator described in Section 3. Over the past three years, the number of videos posted on the platform has grown steadily, increasing by two orders of magnitude.

For each month in our dataset, we compute the number of conspiracy videos and determine the

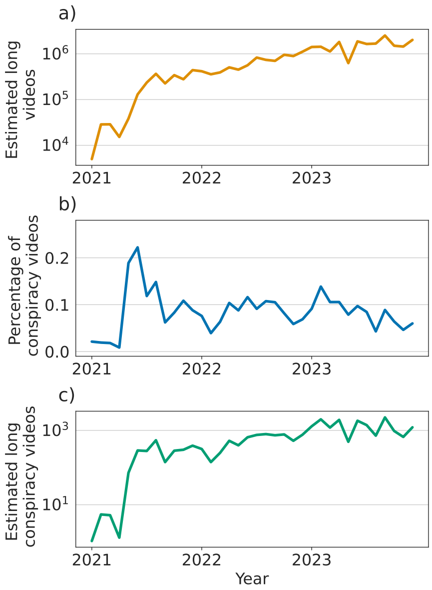

Figure 1: (a) Estimated number of long videos on TikTok U.S. per month. (b) Percentage of conspiracy videos. (c) Estimated number of long conspiracy videos on TikTok U.S.

percentage of conspiratorial content using distant labeling. Combining this information with estimates of the total number of long videos posted on TikTok in the U.S., we can estimate the number of conspiratorial videos. Figures 1b and 1c show the temporal trends of these volumes. In Q3 2021, there was a spike in the upload of conspiracy videos, peaking at 0.2% of our overall sample (N=542), likely driven by the platform's growth and an increase in the total number of posted videos. This percentage declines in the following months and stabilizes around 0.1%, even as the absolute number of conspiracy videos grows steadily before leveling off, reaching an estimated one thousand conspiracy videos per month in 2023.

Based on these findings, we conclude that conspiracy-related videos are present on TikTok, with a prevalence reaching up to one in every five hundred uploaded videos during specific periods.

4.2 Impact of the Creativity Program on Conspiratorial Content

On February 20, 2023, TikTok announced a new Creativity Program for creators, who would be eligible to earn money from their content if they met specific requirements, such as posting videos longer than 1 minute. These changes took effect starting from May 3 in the U.S. and other countries. We already observed in Section 4.1 an overall increment in the quantity of long conspiratorial videos during the period of analysis. We now investigate whether this effect might be due to the introduction of the Creativity Program. To this aim, we collect an additional dataset of 10k videos of any duration sampled uniformly at random (i) before and (ii) after the introduction of the Creativity Program, retrieving the necessary video metadata to extract the length of the content (since the API does not provide this information). This data is instrumental to measure changes in the whole population of videos, including short ones (duration < 1 min). Figure 2 shows the distribution of the durations of collected videos in the two periods. The number of long videos in the sample increases from N=39 (0.39%) to N=103 (1.03%), probably due to the introduction of the Creativity Program on May 3. The application of two statistical tests (two-sided Mann-Whitney and Chi-square) confirms that the two distributions are statistically different ($p < 0.001$ in both cases). This analysis indicates that the new Creativity Program influenced platform-wide behavior globally rather than specifically affecting the production of conspiratorial content, as the fraction of conspiracy-related videos remained stable.

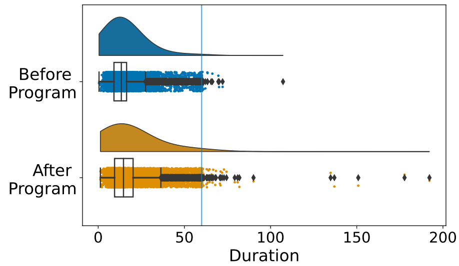

Figure 2: Distribution of the duration in seconds of videos created before (N=42) and after (N=103) the beginning of the Creativity Program (May 3, 2023). The vertical blue line indicates 60 seconds. Medians: before = 13.4 s, after = 14.7 s

4.3 Detecting Conspiracy Theories with LLMs

This section evaluates the performance of different models on identifying conspiratorial videos by measuring their precision (P) and recall (R). The last two configurations (`C2` and `C3`), which represent unbalanced settings, are particularly relevant in the context of content moderation, as they highlight the trade-off between false positives (and thus precision) and recall.

As a baseline model, we use RoBERTa base (Liu et al., 2019) and fine-tune it on both the balanced (`C1`) ($P = R = 0.83 \pm 0.01$ on validation) and

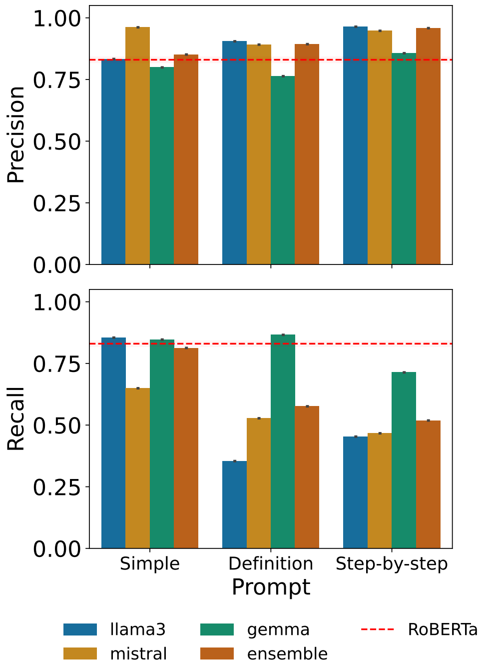

Figure 3: Aggregated results on the positive class (conspiracy videos) of the balanced classification experiment with distant labels (C1), for the three prompts and the three models plus the ensemble, across all seeds. The red line is the result of the fine-tuning, 95% C.I. for RoBERTa.

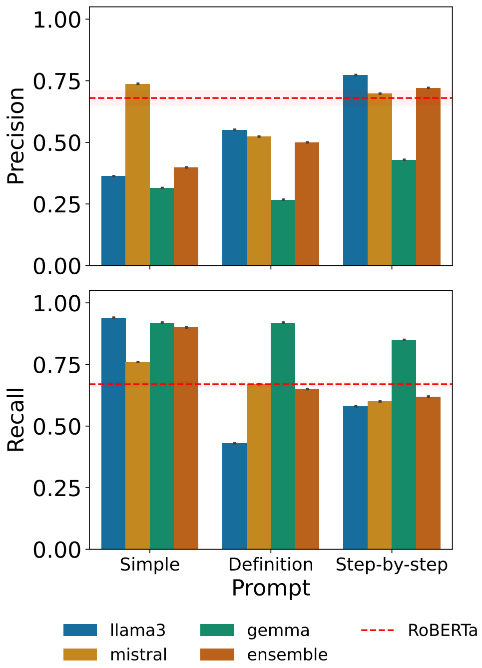

Figure 4: Aggregated results on the positive class (conspiracy videos) of the unbalanced classification experiment with manual labels (C3), for the three prompts and the three models plus the ensemble, across all seeds. The red line is the result of the fine-tuning, 95% C.I. for RoBERTa.

unbalanced datasets (C3) ($P = 0.68 \pm 0.04$, $R = 0.67 \pm 0.01$ on validation) with a 0.72-0.25 train-test split. Appendix A.4 provides further details.

Figures 3 and 4 report the results for the classification task, i.e., detecting the positive (conspiracy) class, specifically for the first (C1) and third (C3) configurations. Additional results for case C2 are included in Appendix A.5 and are similar to those for C3. In both configurations, the three models are surprisingly consistent across all the 25 runs. There is little to no variation in the classification results across the two datasets and the different prompts.

Balanced setting (C1). Here, the different prompts show an interesting pattern (Figure 3). The best-performing model, precision-wise, is Llama3 with an almost perfect score (0.96), with the Step-by-step prompt. Most models performed better (+1 to +12 pp.) than the baseline, except for Gemma with the Simple and Definition prompts. Overall, the Step-by-step prompt results in the highest precision, although the difference with the Simple prompt is relatively small. We see instead an opposite tendency for what concerns the recall: the Simple prompt is the best performing overall, with the highest score of 0.87 by Gemma with the Definition prompt. This last configuration, together with Gemma and LLama3 with the Simple prompt, are the only ones that perform better than the baseline in terms of recall, while the other models perform significantly worse than RoBERTa (-1 to -49 pp.).

Unbalanced setting with manual labels (C3). Here we observe a significant decrease in overall performance (Figure 4). However, the trends identified in the previous configuration remain valid. The model with the highest precision is again Llama3, with a score of almost 0.77 with the Step-by-step prompt. The same model is the one with the highest recall (0.94), although with the Simple prompt. Unfortunately, most of the models perform significantly worse than the baseline in terms of recall (-1 to -49 pp.). On this dataset, the best overall tradeoff is given by Mistral with the Simple prompt, as it improves over the baseline in both metrics.

Effect of text length. Finally, we investigate the relation between classification performance and text length. Figure 5 charts the classification results for case C1 for the different quartiles of the text length distribution. In these plots, we are more interested in the trends rather than the absolute values of the points of the curve (which is better shown in Figures 3 and 4). Considering precision, the longer

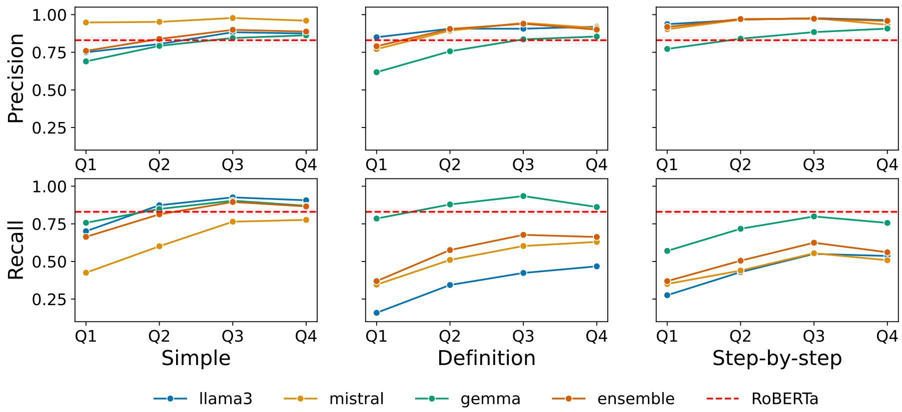

Figure 5: Aggregated results on the positive class (conspiracy videos) of the balanced classification experiment with distant labels (**C1**), for the three prompts and the three models plus the ensemble, across all seeds, split by quartiles on the distribution of the length of the transcriptions in words. The red line is the result of the fine-tuning, 95% C.I. for RoBERTa. Q1 = 210 words, Q2 = 325 words, Q3 = 472 words, Q4 = 1919 words.

the transcript, the higher the metric. This trend is shared among all the models we consider. A similar behavior is present for recall, where the metric increases steadily for longer texts, with even a steeper slope than precision. However, the improvement stops or even reverts when going from **Q3** to **Q4**. For the negative class, RoBERTa is the most precise model in case **C1**, while in **C2** and **C3** LLama3 and Gemma had almost perfect precision (0.99) while being slightly lower in recall (0.98). We report the full results of the classification process for the negative (not conspiracy) class and additional results for **C3** in Appendix A.5.

To conclude, note that the ensemble model proves to be a balanced choice, with a better trade-off between precision and recall, but at the cost of increased deployment complexity and runtime, since each item needs to be labeled by each of the three different models. In Section 4.4, we present a qualitative error analysis, focusing on videos that were unanimously misclassified by all models. Specifically, we provide summaries for two false positives and two false negatives, along with interpretations of the possible errors.

4.4 Error Analysis

We provide examples of errors (**False Positive**: the models interpreted the transcription as conspiratorial when it is not, **False Negative**: the transcription mentioned conspiracy theories but the models did not capture it) committed by the models unanimously, meaning that all three of them misclassified the item. We provide a manual interpretation rather than relying on the model-generated explanations, as these are neither reproducible nor reflective of the model's true reasoning process (Saba, 2023).

First, we show two **False Negative** summarizing their transcription:

- **Dome-Earth:** The speaker rejects the flat-earth theory but claims the Earth is under a dome, making references to the Torah and older versions of the Bible that describe the creation of a "firmament" supporting their dome theory. They argue that NASA's images of Earth are digitally altered and claim NASA admits this.

- **Big Foot:** The speaker mentions hearing distinct knocking sounds on wood, a behavior often associated with Bigfoot sightings. This knocking occurred in a quiet neighborhood where few people were outside after dark. The speaker links the current experience to a previous instance in the same area where they believe they encountered Bigfoot. Notes that local dogs often bark uncontrollably at night, adding to the sense of something unusual occurring.

Both examples explicitly reference a conspiracy theory; however, the models failed to detect it. A possible explanation lies in the tone of the content. In the first example, the rejection of the conspiracy theory (flat-earth) in favor of another theory (dome-earth) may have influenced the models' output, as its sequential processing of the text might have led it to misinterpret the context by interpreting the early rejection as a debunking video. The second example explicitly mentions "Big Foot", yet the models might have categorized the content as mystery or creepy-themed material rather than identifying it as a conspiracy theory.

We then provide two **False Positive** examples:

- **God and racism:** The speaker asserts that racism in humans is a character flaw. However, they distinguish this from divine sovereignty, claiming God has the prerogative to show preferential treatment. One speaker suggests this preference could be interpreted as "racism," while the other counters that it was divine prerogative, not racism.

- **Woke Culture:** The speaker references a previous video about "fake woke" culture, criticizing what they perceive as excessive political correctness, particularly on the West Coast. They mention a comedic skit, which they believe humorously illustrates their point about the absurdities of extreme "woke" ideologies. The skit appears to parody extreme "woke" ideologies by exaggerating contradictions, hypocrisy, and rigid social rules.

These examples were both mistakenly classified as containing conspiracy theories. In the first example, we can suppose the models were driven to label the sample as positive due to the presence of topics such as "God" and "Divine Prerogatives". In the second example, instead, the models may fail to capture the humorous context of the transcript and instead focus on the literal political meaning of the text and the possible conspiracies associated with it (e.g., "Woke Agenda").

5 Discussion and Conclusion

We collected a representative random sample of approximately 1.5 million videos posted on TikTok in the U.S. via the official Research API. From this dataset, we identified relevant hashtags associated with prominent and widely known conspiracy theories, such as *illuminati* and *chemtrails*. Through a combination of hashtag enrichment and manual validation, we identified over a thousand conspiracy-related videos and established a lower bound on the total number of conspiracy theory videos available on the platform. We then analyzed the correlation between the introduction of TikTok's Creativity Program and the overall increase in the duration of videos posted by users, finding that while conspiracy videos became longer after May 3, 2023 (the starting date for the program), so did non-conspiracy ones. We investigated the capabilities of open-weight and widely available Large Language Models to classify conspiracy videos given their content in textual format, i.e., the transcript of the speech. We explored three classification pipelines that represent different scenarios developers and regulators might encounter. First, we employed a hashtag-based distant supervision labeling approach, which eliminates the need for manual video inspection. Next, we replicated this pipeline in a more realistic setting, where conspiracy videos constitute a minority, highlighting the impact of class imbalance. Finally, we refined the pipeline by incorporating manual labeling of the collected conspiracy videos to enhance classification accuracy.

Both classical methods (RoBERTa) and LLMs have their tradeoffs. Zero-shot prompting removes the cost of training, thus needing only the computational power for inference tasks. More importantly, it obviates the need to label examples (whether manually or via heuristics such as distant supervision). Instead, fine-tuning RoBERTa requires the computational and storage capabilities to handle the training and validation of the model. Still, the inference via LLMs is computationally more expensive than with RoBERTa. These results should inform decision-making strategies when faced with the deployment of similar systems, e.g., to decide how much effort to allocate to data labeling. Additionally, the experimental settings we used produced results that are in line with other work in the literature (Atreja et al., 2024; Ziems et al., 2024), especially for what concerns the use of definitions and explanations in the prompts and the correlation with increased performance. We underline that in our setup, some configurations surpass the baseline state-of-the-art models, which is a result in contrast with current literature (Plaza-del Arco et al., 2024). The findings and methodologies presented in this paper have significant implications for both researchers and social media platforms. By identifying conspiracy theory videos, our framework can aid in developing more effective moderation strategies and policies to mitigate the diffusion of harmful content. Moreover, the insights gained from analyzing the impact of TikTok's Creativity Program offer valuable perspectives on how platform policies influence content creation trends. Lastly, LLMs' effectiveness in automating labeling highlights their potential for scalable, data-driven content moderation.

Limitations

As with any empirical work, this study is subject to some limitations. First, we rely on the TikTok Research API, which functions as an opaque system. As we lack control over its internal mechanisms, our data may be subject to biases or misrepresentations. Additionally, our estimates depend on the initial seed hashtags used in the enrichment process, meaning that different seeds could yield different results. Moreover, our analyses assume that the meaning of conspiratorial hashtags is mostly static across the period of the study. Our analysis of the Creativity Program's effect is correlational, as the possible presence of hidden confounders prevents us from making causal inferences as in other studies (Lenti et al., 2025). While we provide a detailed and reproducible setup when utilizing LLMs, some degree of uncertainty inherent to these models remains unavoidable. Also, the models we chose to analyze can have weaker performance compared to larger versions. In the trade-off between performance and reproducibility, together with computational requirements, we chose to prioritize the latter as it allowed us to study the possibility of using these smaller and less costly models to approach the task of content analysis. Finally, we do not distinguish between the various conspiracy theories present within our collected sample.

Future research should critically examine potential biases in the data provided by the TikTok Research API and compare it with datasets collected directly from the platform. Enhancing the current work could involve incorporating multimodal data, such as video keyframes, into the classification pipeline or employing anguage models that are natively multimodal (Liu et al., 2023). This approach may improve the robustness of the analysis. Similarly, extending the study to other online media platforms and evaluating newly released and bigger models would broaden its scope and applicability. Additionally, employing Retrieval-Augmented Generation (RAG) to provide contextual information in prompts for LLMs could improve the model's comprehension and accuracy. Further investigation should also focus on a more detailed quantitative analysis of the behavioral patterns of conspiracy content creators, tracking how these behaviors evolve over time. Such studies could shed light on the dynamics of conspiracy theory propagation across platforms.

Ethical considerations

TikTok users have explicitly agreed to the Terms of Service, which include the acknowledgment and approval of the transfer of personal data through the API.[6] Throughout the data collection process, we ensured the anonymization of user data to protect individual privacy. All results presented are aggregated, i.e., they reflect collective trends rather than identifiable individual behaviors. Moreover, no personal information is given as input to the Large Language Models we employed. As requested by the Research API Terms of Service, we have implemented robust data security measures to prevent unauthorized access and ensure that all data handling procedures align with current best practices in data protection. We cannot make our full dataset public due to the TikTok API usage policies. We provide access to the code to replicate our results and the dehydrated datasets (Video ID - Label) in the repository associated with the paper.[7] Additionally, while our study highlights the potential of LLMs for detecting conspiratorial content, their current limitations also present an ethical concern, as adversaries may exploit these weaknesses to evade moderation techniques. The models we employed are free to use and attribute ownership of the output to the user (Mistral:Apache2.0[8], Llama and Gemma have proprietary custom licenses.[9],[10])

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30. (Cited on 1)

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv preprint arXiv:2406.11980*. (Cited on 9)

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863. (Cited on 1)

[6] TikTok privacy policy accessed on 12/01/2025

[7] Anonymyzed repository: `https://anonymous.4open.science/r/ct_tt-FC7E`

[8] Mistral License

[9] Gemma License

[10] Llama License

Corey H Basch, Zoe Meleo-Erwin, Joseph Fera, Christie Jaime, and Charles E Basch. 2021. A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on tiktok. *Human vaccines & immunotherapeutics*, 17(8):2373–2377. (Cited on 2)

J. Byford. 2011. *Conspiracy Theories: A Critical Introduction*. Springer. Google-Books-ID: m5Er9ELOwQkC. (Cited on 1, 2)

Deptii D. Chaudhari and Ambika V. Pawar. 2022. A Systematic Comparison of Machine Learning and NLP Techniques to Unveil Propaganda in Social Media:. *Journal of Information Technology Research*, 15(1):1–14. Number: 1. (Cited on 1)

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. (Cited on 4)

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96. (Cited on 3)

Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2024. What we can learn from tiktok through its research api. In *Companion Publication of the 16th ACM Web Science Conference*, pages 110–114. (Cited on 3, 14)

Ahmad Diab, Rr. Nefriana, and Yu-Ru Lin. 2024. Classifying conspiratorial narratives at scale: False alarms and erroneous connections. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):340–353. (Cited on 5)

Karen M. Douglas and Robbie M. Sutton. 2023. What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication. *Annual Review of Psychology*, 74(1):271–298. _eprint: https://doi.org/10.1146/annurev-psych-032420-031329. (Cited on 2, 5, 16)

Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding Conspiracy Theories. *Political Psychology*, 40(S1):3–35. (Cited on 2)

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. (Cited on 4)

Adam M. Enders, Joseph E. Uscinski, Michelle I. Seelig, Casey A. Klofstad, Stefan Wuchty, John R. Funchion, Manohar N. Murthi, Kamal Premaratne, and Justin Stoler. 2023. The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. *Political Behavior*, 45(2):781–804. (Cited on 2)

SK Ahammad Fahad, Daniel Wang Zhengkui, Ng Pai Chet, Nicholas Wong, Aik Beng Ng, and Simon See. 2025. Advancements and applications of multimodal large language models: Integration, challenges, and future directions. In *AI-Driven: Social Media Analytics and Cybersecurity*, pages 309–336. Springer. (Cited on 5)

Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander Van Der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on twitter. *Group Processes & Intergroup Relations*, 24(4):606–623. (Cited on 2)

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27. (Cited on 3)

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake News vs Satire: A Dataset and Analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, pages 17–21, New York, NY, USA. Association for Computing Machinery. (Cited on 2)

Irving J Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264. (Cited on 4)

Asmelash Teka Hadgu, Kiran Garimella, and Ingmar Weber. 2013. Political hashtag hijacking in the us. In *Proceedings of the 22nd international conference on world wide web*, pages 55–56. (Cited on 3)

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*, volume 34, page 100. (Cited on 3)

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*. (Cited on 4)

Claire Wonjeong Jo, Magdalena Wojcieszak, et al. 2024. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*. (Cited on 2)

Jacopo Lenti, Luca Maria Aiello, Corrado Monti, and Gianmarco De Francisci Morales. 2025. Causal modeling of climate activism on reddit. In *Proceedings of the ACM on Web Conference 2025*, pages 590–600. (Cited on 10)

Yachao Li, Mengfei Guan, Paige Hammond, and Lane E Berrey. 2021. Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in the COVID-19 information hub. *Health Education Research*, 36(3):261–271. (Cited on 2)

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 2359–2370, New York, NY, USA. Association for Computing Machinery. (Cited on 2)

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916. (Cited on 10)

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. *CoRR*, abs/1907.11692. (Cited on 6, 14)

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22(2):205–224. (Cited on 1)

Marcel Meuer, Aileen Oeberst, and Roland Imhoff. 2023. How do conspiratorial explanations differ from non-conspiratorial explanations? A content analysis of real-world online articles. *European Journal of Social Psychology*, 53(2):288–306. (Cited on 2)

Alessandro Miani, Thomas T. Hills, and Adrian Bangerter. 2021. LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, pages 1–24. (Cited on 2, 3, 14)

Alexis Palmer, Noah A Smith, and Arthur Spirling. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, 4(1):2–3. (Cited on 2)

Samantha C. Phillips, Lynnette Hui Xian Ng, and Kathleen M. Carley. 2022. Hoaxes and Hidden agendas: A Twitter Conspiracy Theory Dataset: Data Paper. In *Companion Proceedings of the Web Conference 2022*, WWW '22, pages 876–880, New York, NY, USA. Association for Computing Machinery. (Cited on 2)

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds. exploring model label variation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 19–30, Torino, Italia. ELRA and ICCL. (Cited on 9)

Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, OASIS '21, pages 21–25, New York, NY, USA. Association for Computing Machinery. (Cited on 2)

Anne Quaranto. 2022. Dog whistles, covertly coded speech, and the practices that enable them. *Synthese*, 200(4):330. (Cited on 3)

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR. (Cited on 4)

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141. (Cited on 2)

Cesare Rollo, Gianmarco De Francisci Morales, Corrado Monti, and André Panisson. 2022. Communities, gateways, and bridges: Measuring attention flow in the reddit political sphere. In *International Conference on Social Informatics*, pages 3–19. Springer. (Cited on 1)

Walid S Saba. 2023. Stochastic llms do not understand language: towards symbolic, explainable and ontologically based llms. In *International Conference on Conceptual Modeling*, pages 3–19. Springer. (Cited on 8)

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. (Cited on 2)

Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review*, 34(3):150–160. (Cited on 1)

Robbie M Sutton and Karen M Douglas. 2020. Conspiracy theories and the conspiracy mindset: implications for political ideology. *Current Opinion in Behavioral Sciences*, 34:118–122. (Cited on 2)

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*. (Cited on 4)

Petter Törnberg. 2024. Best practices for text annotation with large language models. *Sociologica*, 18(2):67–85. (Cited on 5)

Joseph E. Uscinski, Karen Douglas, and Stephan Lewandowsky. 2017. Climate Change Conspiracy Theories. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press. (Cited on 1)

Wenxuan Wang, Jingyuan Huang, Chang Chen, Jiazhen Gu, Jianping Zhang, Weibin Wu, Pinjia He, and Michael Lyu. 2023. Validating multimedia content moderation software via semantic fusion. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 576–588. (Cited on 2)

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, pages 577–582. IEEE. (Cited on 4)

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837. (Cited on 5)

Gabriel Weimann and Natalie Masri. 2023. Research note: Spreading hate on tiktok. *Studies in conflict & terrorism*, 46(5):752–765. (Cited on 1, 2)

Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on tiktok. *Policy & Internet*, 14(1):79–95. (Cited on 1)

Marco Zenone and Timothy Caulfield. 2022. Using data from a short video social media platform to identify emergent monkeypox conspiracy theories. *JAMA Network Open*, 5(10):e2236993–e2236993. (Cited on 2)

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291. (Cited on 1, 9)

A Appendix

A.1 Seed hashtags

Table 1 presents the 10 seed keywords used for the hashtag enrichment process, which include nine terms from LOCO dataset (Miani et al., 2021) along with the keyword *conspiracy*, accompanied by brief descriptions.

A.2 Randomness in LLMs experiments

We assess the model's consistency by using different seeds during the classification process. Setting a seed ensures the reproducibility of our results. Here we list the seeds used in our classification process:
0, 4, 7, 8, 10, 12, 17, 25, 26, 33, 37, 42, 49, 50, 57, 69, 73, 100, 123, 420, 444, 666, 777, 1111, 1999

A.3 Data Collection Process

We collect our data by stratifying samples by week to mitigate issues observed in previous research, particularly the non-uniform distribution of data returned by the random sample endpoint (Corso et al., 2024). We focus on collecting long videos, i.e., those with duration ≥ 1 minute, as they are expected to provide sufficient contextual information for effective analysis by LLMs. We perform requests to the `/video/query` endpoint for each week from January 2021 to December 2023, retaining only videos published in the U.S. and setting the parameter `is_random = True`. Figure 6 reports the time series of collected unique videos. There is a visible change in the number of collected videos starting from July 2021, when TikTok started allowing videos up to three minutes in length to be uploaded. We conduct exploratory analyses of the videos' temporal features to build on previous work and evaluate the effectiveness of our sampling technique. Figure 7 shows the results of these evaluations. Sub-figure **a** shows the number of videos posted for each given day of the month: all days are represented, with some expected variability. Sub-figures **b,c** depict similar behavior to what has been already described, with a majority of the videos created on Saturdays and the normal circadian world rhythm (Corso et al., 2024). Finally, sub-figure **d** reveals that the first minute of the hour presents a slightly higher number of videos, possibly due to the scheduling of videos by creator accounts.

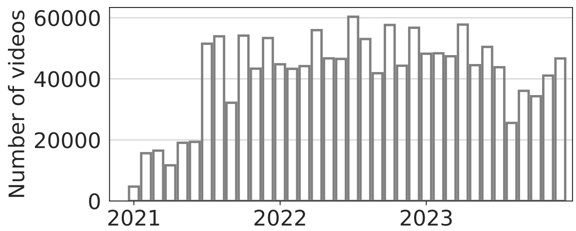

Figure 6: Monthly number of unique long videos (duration ≥ 1 minute) collected from 2021 to 2023.

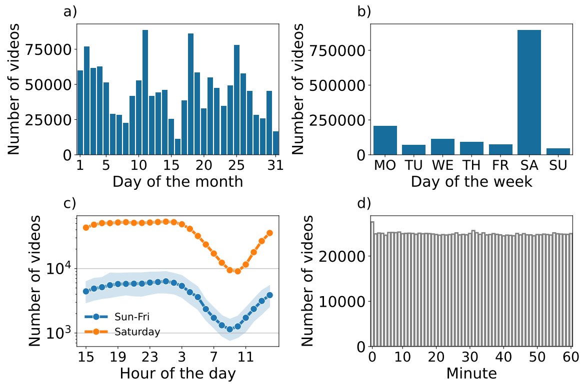

Figure 7: Number of videos posted (a) for each day of the month, (b) for each day of the week, (c) for each hour of the day (UTC), and (d) for each minute of the hour. The API shows a clear bias at the daily level, with a non-uniform distribution over different days of the month, and a disproportionate number of videos shared on Saturdays.

A.4 Fine-tuning RoBERTa

As a baseline, we use RoBERTa (Liu et al., 2019), available on HuggingFace. We adopt the default training pipeline provided by the `tranformers` library and fine-tune the model by using both balanced and unbalanced datasets with distant labels. We use a training-test split of 0.75-0.25 for 5 epochs with a learning rate of $\alpha = 2 \times 10^{-5}$ and we instruct the default trainer to keep the best model at the end.

A.5 Additional Classification Results

Figure 8 reports the trends for Precision and Recall in `C3`. Table 2 lists the classification results for the negative class. Figure 11 shows the classification results for the positive class with the unbalanced dataset and distant labels. The results are similar to the ones for the configuration `C1` analyzed in Section 4. Generally, giving longer texts increases the metrics. However, in some cases, long texts (**Q4**) perform worse than slightly shorter ones (**Q3**) in Precision and Recall.

Table 1: Seeds from LOCO used as hashtags for the enrichment process along with a brief description of the theory.

Seed	Description
conspiracy	General seed of the topic.
flatearth	The flat Earth conspiracy theory posits that the Earth is not a sphere but a flat, disc-shaped plane, contrary to centuries of scientific evidence and exploration.
qanon	The QAnon conspiracy theory alleges that a secret cabal of Satan-worshipping pedophiles, including prominent politicians and celebrities, is controlling the world and that former President Donald Trump is working to dismantle this group.
newworldorder	The New World Order conspiracy theory asserts that a clandestine group of powerful elites is conspiring to establish a single, global government that would control all aspects of life and suppress individual freedoms.
chemtrails	The chemtrails conspiracy theory claims that the trails left by aircraft in the sky, known as contrails, are actually chemicals being deliberately sprayed for nefarious purposes, such as weather manipulation, population control, or biological warfare.
mindcontrol	The mind control conspiracy theory suggests that governments, corporations, or other powerful entities are using advanced technologies or psychological techniques to manipulate and control the thoughts, behaviors, and perceptions of individuals.
reptilian	The reptilian conspiracy theory posits that shape-shifting reptilian aliens, disguised as humans, are controlling the world by infiltrating positions of power in government, media, and industry.
bigfoot	The Bigfoot conspiracy theory suggests that the existence of Bigfoot, a large, ape-like creature said to inhabit remote forests, is being deliberately covered up by governments or other authorities. Proponents believe that evidence of Bigfoot's existence is suppressed to prevent public knowledge
illuminati	The Illuminati conspiracy theory asserts that a secret society called the Illuminati is covertly orchestrating global events and manipulating governments, economies, and media to establish a New World Order.
ufo	The UFO conspiracy theory claims that governments, particularly the U.S. government, are hiding evidence of unidentified flying objects (UFOs) and extraterrestrial life from the public.

Table 2: Classification results for the negative class (not conspiracy videos) , for the different LLM configurations with **C1** balanced dataset with distant labels (Left), **C2** unbalanced with distant labels (Center), and **C3** unbalanced with manual labels (Right). SP = Simple Prompt, DP = Definition Prompt, SBS = Step-by-step prompt.

Configuration	**Balanced with distant labels**						**Unbalanced with distant labels**						**Unbalanced with manual labels**					
Measure	Precision			Recall			Precision			Recall			Precision			Recall		
Prompt	SP	DP	SBS	SP	DP	SBS	SP	DP	SBS	SP	DP	SBS	SP	DP	SBS	SP	DP	SBS
LLama3	0.84	0.58	0.62	0.81	0.96	0.98	0.98	0.92	0.98	0.81	0.96	0.98	0.99	0.93	0.95	0.80	0.96	0.98
Mistral	0.72	0.63	0.62	0.97	0.81	0.90	0.95	0.94	0.94	0.97	0.80	0.91	0.97	0.96	0.95	0.97	0.81	0.91
Gemma	0.83	0.84	0.74	0.77	0.67	0.86	0.97	0.98	0.97	0.77	0.69	0.87	0.99	0.99	0.98	0.76	0.66	0.86
Ensemble	0.80	0.66	0.65	0.84	0.92	0.98	0.97	0.93	0.94	0.84	0.93	0.98	0.99	0.95	0.95	0.84	0.92	0.97
RoBERTa		0.88 ± 0.02			0.86 ± 0.03			0.95 ± 0.01			0.97 ± 0.01			0.95 ± 0.01			0.97 ± 0.01	

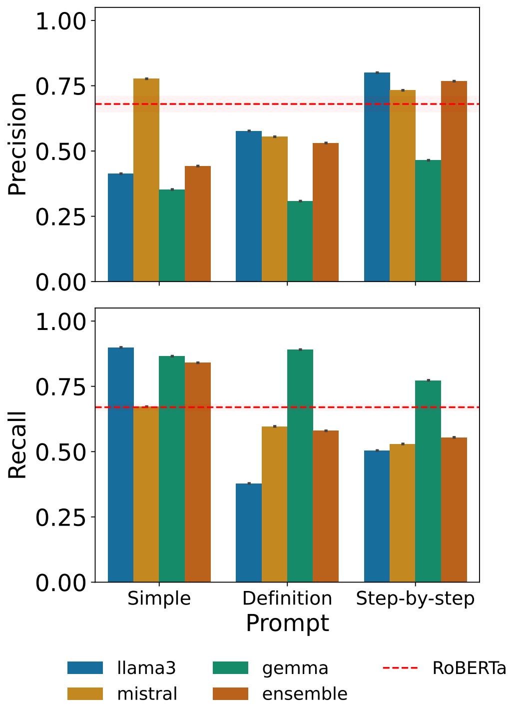

Figure 8: Aggregated results on the positive class of the unbalanced classification experiment with distant labels (`C2`), for the three prompts and the three models plus the ensemble, across all seeds. The red line is the result of the fine-tuning, C.I 95% for RoBERTa (the same used for case `C3`).

A.6 Evaluation of Whisper accuracy

Some videos (<5%) have unintelligible content or spoken content in languages different from English. We discard these videos from our sample via simple text filtering based on the obtained transcription from the TikTok Research API. We then manually code the transcriptions of a set of 100 videos drawn randomly from our sample. This process consisted in watching every video of the sample and manually transcribing the content of the audio.

A.7 Transcription Word Lengths

Figure 10 shows the per-class distributions of the length (in words) of the transcripts extracted by Whisper. There is a small yet significant statistical difference between the two distributions (two-sided Mann-Whitney, $p < 0.001$). Section 4.3 investigates the correlation between the length of the transcript and the performance of LLMs.

A.8 Definition of Conspiracy Theory

We use the definition of conspiracy theory provided by Douglas and Sutton (2023) in the *Definition Prompt*. *'A conspiracy theory is a belief that two or more actors have coordinated in secret to achieve*

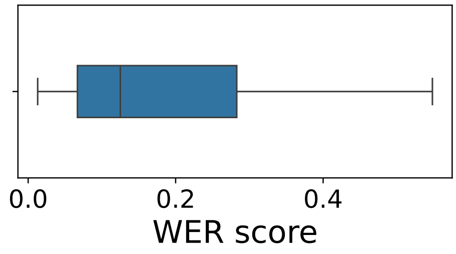

Figure 9: Distribution of WER scores for videos transcribed by Whisper compared to manual coding.

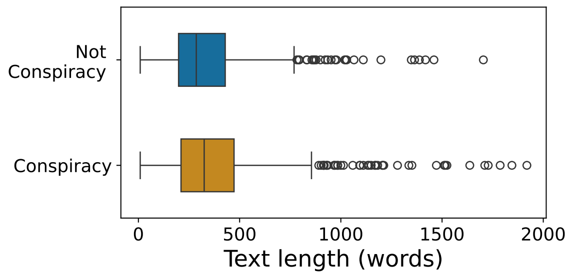

Figure 10: Distributions of the length of the video transcripts for the positive (conspiracy) and negative (not conspiracy) class in the balanced dataset. Medians: Not Conspiracy = 286 words, Conspiracy = 325 words

an outcome and that their conspiracy is of public interest but not public knowledge. Conspiracy theories (a) are oppositional, which means they oppose publicly accepted understandings of events; (b) describe malevolent or forbidden acts; (c) ascribe agency to individuals and groups rather than to impersonal or systemic forces; (d) are epistemically risky, meaning that though they are not necessarily false or implausible, taken collectively they are more prone to falsity than other types of belief; and (e) are social constructs that are not merely adopted by individuals but are shared with social objectives in mind, and they have the potential not only to represent and interpret reality but also to fashion new social realities'.

Table 3: Sizes of the various datasets used to answer the three research questions, and duration of the videos.

Research Question	Total Videos	Unique Videos	CT Videos	NoCT Videos	Duration
RQ1	1 605 696	1 494 831	1300		>1 min
RQ2	24 366	24 366			Any
RQ3 - C1		1666	887	779	>1 min
RQ3 - C2		879	100	779	>1 min
RQ3 - C3		879	100	779	>1 min

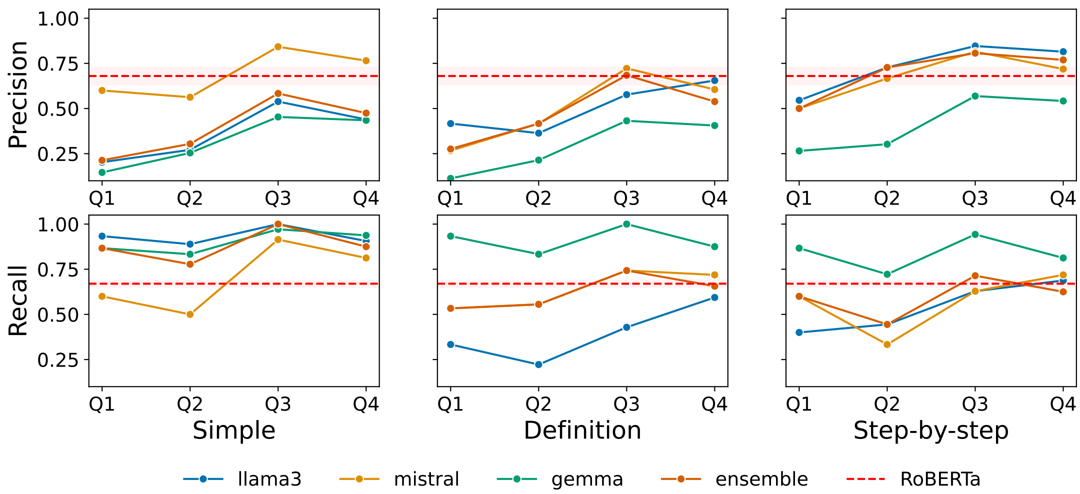

Figure 11: Aggregated results on the positive class (conspiracy videos) of the unbalanced classification experiment with manual labels (**C3**), for the three prompts and the three models plus the ensemble, across all seeds, split by quartiles on the distribution of the length of the transcriptions in words. The red line is the result of the fine-tuning, 95% C.I. for RoBERTa. Q1 = 210 words, Q2 = 325 words, Q3 = 472 words, Q4 = 1919 words.