# Algorithmic Behaviors Across Regions: A Geolocation Audit of YouTube Search for COVID-19 Misinformation Between the United States and South Africa

**Hayoung Jung[1], Prerna Juneja[2], Tanushree Mitra[1]**

[1]University of Washington
[2]Seattle University
hjung10@uw.edu, pjuneja@seattleu.edu, tmitra@uw.edu

## Abstract

Despite being an integral tool for finding health-related information online, YouTube has faced criticism for disseminating COVID-19 misinformation globally to its users. Yet, prior audit studies have predominantly investigated YouTube within the Global North contexts, often overlooking the Global South. To address this gap, we conducted a comprehensive 10-day geolocation-based audit on YouTube to compare the prevalence of COVID-19 misinformation in search results between the United States (US) and South Africa (SA), the countries heavily affected by the pandemic in the Global North and the Global South, respectively. For each country, we selected 3 geolocations and placed sock-puppets, or bots emulating "real" users, that collected search results for 48 search queries sorted by 4 search filters for 10 days, yielding a dataset of 915K results. We found that 31.55% of the top-10 search results contained COVID-19 misinformation. Among the top-10 search results, bots in SA faced significantly more misinformative search results than their US counterparts. Overall, our study highlights the contrasting algorithmic behaviors of YouTube search between two countries, underscoring the need for the platform to regulate algorithmic behavior consistently across different regions of the Globe.

**Warning: We caution the readers that some examples provided to better contextualize our data can be offensive.**

## 1 Introduction

*"This virus is here to stay. It is still killing and it is still changing."* — Ghebreyesus (2023), Director-General of the World Health Organization

Since March 2020, the World Health Organization (WHO) has designated COVID-19 as a global pandemic. The pandemic continues to pose a public health threat, registering 206K COVID-19 cases and 3.5K deaths worldwide in July 2024 alone (WHO 2024). As the largest video search engine, YouTube has emerged as a vital tool for finding health-related information online, particularly during outbreaks and global pandemics (Dubey et al. 2014; Bora et al. 2018; Khatri et al. 2020). However, YouTube has faced criticism for disseminating COVID-19 misinformation globally to its users, with fact-checkers calling the platform a "major conduit of fake news" (Milmo 2022). This misinforma-
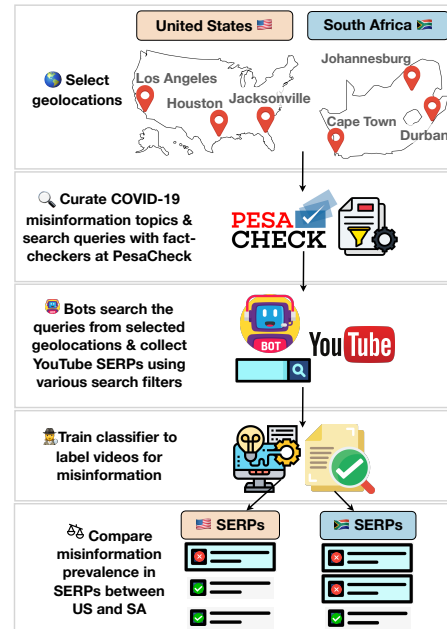


Figure 1: **Pipeline Overview.** Sock-puppet bots emulating real-world users utilized the curated search queries to gather YouTube search engine result pages (SERPs) from geolocations in the United States (US) and South Africa (SA). After training and employing a classifier to scale the video labeling process, we compared the prevalence of COVID-19 misinformation in SERPs between the two countries.

tion has undermined public health efforts worldwide, fueling vaccine hesitancy and eroding trust in health institutions (Do Nascimento et al. 2022). Thus, there is a pressing need for empirical investigation of search engine systems to ensure algorithmic accountability, safeguard global health, and promote a more responsible and trustworthy web.

In response to public pressure, YouTube collaborated with the WHO to develop a content moderation policy for COVID-19 misinformation for their platform (Google 2020). However, past reports expressed concerns that YouTube's content moderation practices are biased in favor of the Global North and neglect the misinformation challenges in the Global South, noting that YouTube does

not fact-check or remove videos in non-English languages (Nguyen and Scurato 2021; Grant 2022). While multiple previous studies have empirically investigated search engines for COVID-19 misinformation in the Global North context (Papadamou et al. 2022; Kravets et al. 2023), the Global South has received little attention. The global aspect of the pandemic presents a unique and crucial opportunity to conduct a comparative audit of COVID-19 misinformation in two different parts of the world.

Guided by an overarching research question: *What is the prevalence of COVID-19 misinformation in YouTube search results between the US and SA?*, we conducted a comprehensive 10-day geolocation-based comparative audit of YouTube search from January 30th, 2023 to February 9th, 2023. The goal of the audit was to compare the prevalence of COVID-19 misinformation in Search Engine Result Pages (SERPs) between the United States (US) and South Africa (SA), the countries heavily affected by the pandemic in the Global North and the Global South, respectively. For each country, we selected 3 geolocations and placed sock-puppets (bots emulating real users) that collected YouTube SERPs for 48 search queries belonging to 8 globally persistent COVID-19 misinformation topics, such as "Bill Gates Claims." To gain deeper insights into the platform's sorting algorithm, we sorted the search results across 4 search filters: the default "Relevance," "Upload Date," "View Count," and "Rating," resulting in 915K search results. We scored the videos based on their stance toward COVID-19 misinformation and compared the amount of misinformation present in search results between the US and SA.

We find multiple instances where bots in SA encountered statistically significantly more misinformative SERPs than bots in the US. These disparities were observed within the top-10 search results (p<0.001, $r$=0.49) and search results sorted by YouTube's "Relevance" filter (p<0.001, $r$=0.86) with medium to large effect sizes, indicating practical significance. Since 95% of the user traffic is directed towards the first page of the search results (Kaye 2013) and YouTube employs the "Relevance" filter by default, users in SA may likely encounter significantly more misinformative SERPs than users in the US. Given YouTube's established importance for finding health information, this could raise personal health risks for users in SA by potentially negatively influencing their beliefs, health practices, and decisions. Overall, our work highlights the contrasting algorithmic behaviors of YouTube's search function in two countries within the context of COVID-19 misinformation, underscoring the need for YouTube to regulate its algorithmic behaviors consistently across different regions of the Globe.

**Contributions and Implications.** To our knowledge, our study is the first large-scale geolocation-based comparative audit of YouTube search for COVID-19 misinformation across countries in the Global North and the Global South. This research explores the less-explored context of the Global South, extending the scope of previous algorithmic audits, which primarily centered on the Global North. Our work provides an elaborate understanding of YouTube's search engine across 4 distinct search filters for 8 misin-

formation topics. Through a 10-day data collection and extensive labeling process, we amassed a labeled dataset[1] of 915K search results (10,139 unique videos) and trained a classifier[2] to detect COVID-19 misinformation in YouTube videos. Overall, our work highlights the contrasting algorithmic behaviors of YouTube's search function in two countries within the context of COVID-19 misinformation, underscoring the need for YouTube to regulate its algorithmic behaviors consistently across different regions of the Globe.

## 2   Related Work

### Algorithmic Audits of Search Engines

Search engines determine what information is relevant and shape user behavior, impacting aspects such as political voting behavior (Diakopoulos et al. 2018), scientific knowledge (Papadamou et al. 2022), and beliefs (Knobloch-Westerwick et al. 2015). Despite their societal importance, search engines operate without external regulation, leaving the credibility of content unverified. Consequently, researchers have investigated search systems by conducting algorithmic audits that empirically measure and understand the conditions under which problematic content arises on the platform. Several studies have investigated search engines for misinformation (Hussein, Juneja, and Mitra 2020; Papadamou et al. 2022; Tomlein et al. 2021), conspiracy theories (Chen et al. 2023; Faddoul et al. 2020), hate speech and extremism (Albadi, Kurdi, and Mishra 2022; Ribeiro et al. 2020), and partisanship (Robertson et al. 2018). Among the various methods for auditing search engines, we utilize sock-puppet audits (programming bots to emulate real users), a methodology commonly employed in prior audits (Hussein, Juneja, and Mitra 2020; Papadamou et al. 2022; Kliman-Silver et al. 2015) for its control over experimental variables. Our study adds to the existing sock-puppet audit studies, investigating the prevalence of COVID-19 misinformation on YouTube between the US and SA.

### (Lack of) Algorithmic Audits in the Global South

Prior works have highlighted the growing need to consider the Global South in algorithmic audit research (Costanza-Chock, Raji, and Buolamwini 2022). In a recent paper, Urman, Makhortykh, and Hannák (2024) argued that most algorithmic audit research is skewed towards the Western context, noting that "countries located outside of North America and Western Europe are understudied." A limited number of studies have conducted algorithmic audits within the Global South context, examining factors such as language (Narain et al. 2023) and culture (Dammu et al. 2024). Among these audits conducted in the Global South contexts, only a few considered geolocation, with a vast majority focusing on Google Search (Le et al. 2022; Dabran-Zivan et al. 2023). We contribute to the less-researched context of the Global South and extend prior algorithmic audits on YouTube, the largest video search engine.

---

[1]https://github.com/social-comp/YouTubeAuditGeolocation-data

[2]https://huggingface.co/SocialCompUW/youtube-covid-misinfo-detect

## Search-Enabled COVID-19 Misinformation

Given the importance of search engines for finding health-related information during the COVID-19 pandemic (Khatri et al. 2020), several scholars have audited search engines for COVID-19 misinformation (Li et al. 2020; Kravets et al. 2023; Papadamou et al. 2022) and anti-vaccine content (Juneja and Mitra 2021; Papadamou et al. 2022). In addition, researchers have engineered various features and built machine learning models to automatically detect COVID-19 misinformation online (Ng, Hoffmann Pham, and Luengo-Oroz 2023; Papadamou et al. 2022; Medina Serrano, Papakyriakopoulos, and Hegelich 2020). However, most prior studies have primarily focused on the Global North, leaving a gap in audits of search engines for COVID-19 misinformation in the Global South contexts.

Existing literature regarding the Global South has several gaps, often probing search engines with the default search filter for a single day and focusing on a narrow range of topics. For example, researchers in Narain et al. (2023) focused on the language aspect of the audit, examining COVID-19 videos in 11 widely spoken languages across Africa on YouTube and collecting 562 videos for evaluation. Dabran-Zivan et al. (2023) examined geolocation and language, analyzing 3 COVID-19 conspiracy topics on Google Search across 4 languages in 10 countries for a single day, collecting 330 search results. We systematically compare the prevalence of COVID-19 misinformation on YouTube across countries in the Global North and Global South. Across geolocations in the US and SA, we examine 8 different COVID-19 misinformation topics and 4 distinct search filters over 10 days of data collection, yielding 915K search results. Unlike previous audits with single geolocations per country, we selected three per country for fine-grained comparative analysis within countries.

## 3   Audit Experiment Setup

This section presents the methodology for selecting geolocations for our audit experiments, curating globally persistent COVID-19 misinformation topics and associated search queries, and designing our experimental setup.

### Selecting the Geolocations for the Audit

We considered the US and SA because they were heavily affected by the pandemic in the Global North and the Global South, respectively, making them vulnerable to COVID-19 misinformation[3] (WHO 2024). Previous work established that Google personalizes search results across different states in the US (Kliman-Silver et al. 2015). Therefore, we chose to identify three states in each country to achieve a more robust analysis of COVID-19 misinformation in YouTube search results of the selected countries. We identified three states in the US and three provinces[4] in South Africa with the highest total confirmed cases of COVID-19 (Adeline et al. 2020; Wits University 2023), which may

| Misinformation Topics | Sample Search Queries |
| --- | --- |
| Biological Weapon | Biological Weapon, man-made virus |
| Lab Leak Theory | lab leak theory, Kungflu |
| 5G Claims | 5g and covid19 link, 5g conspiracy |
| Bill Gates Claims | bill gates exposed, bill gates vaccine chip |
| Spread of Virus | social spread, sanitize |
| Treatment of Virus | local concoctions, sesame oil |
| Population Control | population control, plandemic |
| Vaccine Content Claims | MRNA, fetal tissue research |

Table 1: The 8 globally-persistent COVID-19 misinformation topics identified by *PesaCheck*. For each topic, we provide a sample of our curated search queries.

make them susceptible to COVID-19 misinformation. To capture the highest proportion of the population in these states and provinces, we selected the largest populated city as the geolocation. For US, we selected Los Angeles (California), Houston (Texas), and Jacksonville (Florida). For South Africa, we selected Johannesburg (Gauteng), Durban (KwaZulu-Natal), and Cape Town (Western Cape). Figure 1 depicts the geolocations chosen in the US and SA.

### Curating Topics and Search Queries

**Curating COVID-19 Misinformation Topics.**   To curate globally persistent COVID-19 misinformation topics, we partnered with expert fact-checkers at *Pesacheck*.[5] As Africa's largest indigenous fact-checking organization, *PesaCheck* is affiliated with the International Fact-Checking Network (IFCN) and collaborates with expert fact checkers worldwide, making them well equipped to identify misinformation that circulates in diverse regions worldwide.

Using editorial coverage, frequently fact-checked information and social media analytics tools, the expert fact-checkers identified eight globally persistent COVID-19 misinformation topics and provided a dataset of 362 fact-checked YouTube videos related to these topics. These videos originate from channels associated with 29 countries, spanning the Global North and the Global South, including Algeria, China, South Africa, and the United States. This diverse representation–18 countries from the Global North and 11 from the Global South (see Appendix Table 3)–ensures that the search queries developed from these videos are representative across different geographic regions. Refer to Appendix Table 4 for sample YouTube videos provided by *PesaCheck*. Table 1 presents the 8 topics and samples of our curated search queries, which we explain next.

**Curating Search Queries.**   We utilized three methods to curate search queries for each topic. Given that English is the only language spoken commonly in both the US and SA,[6] we focused on queries in English, allowing for a controlled cross-country comparison in the amount of COVID-19 misinformation returned by YouTube's search algorithm.

---

[3]As of May 7th, 2024, the US and SA continue to have the highest reported cases of COVID-19 in the continents of North America and Africa, respectively (WHO 2024).

[4]South Africa operates on a provincial government system.

[5]https://www.pesacheck.org

[6]English is one of eleven official languages in SA, with 10% of people speaking English natively at home and is most commonly spoken in business and commerce (Doochin 2019).

First, we used YouTube video tags in the videos provided by *Pesacheck*. Video tags are descriptive keywords that represent how content creators want their videos to be discovered (Google 2022). Misinformative videos often contain tags that describe misleading narratives, effectively serving as potential search queries to identify more misinformative videos (Juneja, Bhuiyan, and Mitra 2023). We collected 2,911 video tags from the videos and applied systematic filtering criteria used in previous work (Juneja and Mitra 2021). We excluded queries mentioning individuals or news organizations (e.g., "Obama", "Republic TV"), overly generic terms (e.g., "breaking news"), excessively specific terms (e.g., "COVID3rdWaveInMyanmar"), and irrelevant terms (e.g., "cute puppies"). Additionally, we manually merged similar queries (e.g. "man-made" and "man-made virus" into the single query "man-made virus"). These systematic filtering steps resulted in 48 tags. For example, the tags included problematic terms regarding the origin of the virus in China ("Kungflu," "China virus") and terms on self-treatment methods ("local concoction").

Second, for each topic, we identified important keywords found in video titles, descriptions, and transcripts from *PesaCheck*-provided videos. We created a document for each topic by concatenating the video metadata and applying standard preprocessing steps. To extract keywords from each topic document, we used *KeyBERT* (Grootendorst 2022), a keyword extraction tool that requires two inputs: an embedding model and a document. First, *KeyBERT* applies the embedding model to the topic document to generate the document embeddings and extract word embeddings for N-gram phrases.[7] Then, *KeyBERT* uses cosine similarity to identify keywords with the highest similarity with the document itself, providing words that best describe the entire document. We chose the Sentence Transformer *all-mpnet-base-v2* due to their best performance in generating sentence embeddings (Song et al. 2020). Since the model was not trained on COVID-19-related texts, we fine-tuned the model for domain adaptation using COVID-19 question-answering datasets (Möller et al. 2020; Lu 2022; Tang et al. 2020). See Appendix A for fine-tuning details. After fine-tuning, we used *KeyBERT* on each of the 8 topic documents to extract 5 keywords per document, resulting in 40 search queries.

Third, we included the 49 search keywords provided by *PesaCheck* because the keywords were curated by expert fact-checkers to find misinformative videos. Combining the search queries from all three methods, we obtained 137 search queries. Then, we manually removed duplicate queries, combined similar queries, and randomly selected 6 search queries per topic, resulting in a final set of 48 search queries (see Appendix Table 5). To ensure that the search queries output search results relevant to the COVID-19 pandemic, we identified keyword variations of COVID-19 and employed YouTube search operators, which are special commands that can be utilized to efficiently refine their searches on YouTube (Zarr 2020). We selected 8 keyword variations of COVID-19 using a COVID-19 Twitter dataset (Chen, Lerman, and Ferrara 2020). Each query was formatted as:

---

[7]We limited our phrases to unigram, bigram, and trigrams.

```
query (covid | corona | covid-19 | covid19 |
coronavirus | COVD | sars-cov-2 | pandemic)
```

## Experimental Design

**Overview.** To host our experiments, we used Amazon Web Services (AWS) to create all the virtual machines (VMs). We programmed Selenium bots (Patra 2020) to emulate real-world users and automate browser actions. To obscure the automated interactions of the bots, we followed the suggestions from Klimek (2021). Each bot utilized *IPRoyal* proxies (IPRoyal 2023) and validated the IP geolocation of the proxy using *IP2Location*, an IP geolocation lookup service (IP2Location 2023), to obtain personalized search results from the desired IP geolocation. For each query, we collected YouTube SERPs sorted by 4 search filters: "Relevance," "Upload Date," "View Count," and "Rating." During the data collection, we extracted the top 50 search results from each SERP. Additionally, we added wait times after every browser action and chose two evenly separate times to distribute our search queries throughout the day to avoid getting rate-limited by YouTube. We ran the audit experiment for 10 consecutive days from January 30th, 2023 to February 9th, 2023, where we simultaneously searched 24 queries at 00:00 UTC and the other 24 queries at 12:00 UTC.

To control for possible confounding factors that may affect our audit, we followed standard noise control procedures based on prior work (Kliman-Silver et al. 2015; Hussein, Juneja, and Mitra 2020; Juneja and Mitra 2021). To differentiate between noise and geolocation-based personalization in SERPs, we created identical twin bots, consisting of a treatment bot and its corresponding control bot that conducted the same actions simultaneously at each geolocation. Thus, any difference in the search results between the twin bots should be attributed to noise rather than personalization.[8] If the differences in the SERPs between two geolocations exceeded the noise, it can only be attributed to geolocation-based personalization. In our experiment, we placed twin bots at each geolocation, resulting in 12 bots.

**Validation Experiments.** Changing IP addresses using VMs (Hussein, Juneja, and Mitra 2020) and proxies (Boeker and Urman 2022) are common methods to conduct geolocation audits, especially in the US context. However, obtaining *fine-grained* coverage of IP addresses through VMs and proxy services in South Africa was challenging. The only AWS coverage in South Africa, let alone in Africa, is in Cape Town. Meanwhile, nearly all the proxy and VPN providers except *IPRoyal* were unfeasible due to the steep pricing or the lack of proxy coverage of our selected geolocations in South Africa (see Appendix Table 6 for further details). Due to challenges associated with IP addresses, we initially turned to the "geospoofing" method used in Kliman-Silver et al. (2015), in which they fed precise latitude and longitude coordinate information to automated scripts and obtained personalized search results, providing a cost-effective alternative. For brevity, we summarized the

---

[8]Note that YouTube's search engine is a black box. Even after controlling for all known sources of noise, there could be some sources of noise we are unaware of.

results from our validation experiments and left the details in the Appendix. We defined the metric for geolocation-based personalization in Appendix B. In our first validation experiment, we curated a set of search queries and validated that the queries elicited geolocation-based personalization, resulting in personalized SERPs based on geolocation (Appendix C). Using the validated search queries, we performed a second validation experiment and found that YouTube uses IP geolocation instead of geospoofed location to personalize search results (Appendix D). Thus, we conducted a third validation experiment using *IPRoyal* proxies, which validated the accuracy and consistency of the proxies in giving us the correct IP geolocations for our experiment (Appendix E).

## 4 Developing Data Annotation Scheme

Our geolocation audit experiment collected 23,020 SERPs[9] consisting of 915,440 search results (10,139 unique videos). To label these videos for COVID-19 misinformation, we underwent extensive procedures to determine what constitutes misinformation and develop the qualitative coding scheme.

### How Do We Know What Is Misinformation?

To determine what constitutes misinformation, we based our annotation heuristics on Google's COVID-19 medical misinformation policy, which has been developed in partnership with the WHO (Google 2020). We also referenced the policies of national health authorities in the US and SA, such as the Center for Disease Control and Prevention (CDC) (CDC 2023) and the South African Government's COVID-19 Fake News resources (South African Government 2023). Given the misleading nature of xenophobic terms such as "Kungflu" and "China Virus," we also incorporated them in our annotation heuristics. Additionally, we took extra precautions in annotating videos about the COVID-19 Lab Leak Theory[10] by referencing a declassified report from the US National Intelligence Council, which presented the consensus by various government agencies (ODNI 2023). The report assessed that both the Natural Origins Theory[11] and the Lab Leak Theory are plausible until more evidence comes to light. However, the report also debunked many misleading claims, such as COVID-19 being a biological weapon developed in a lab. Thus, we did not label videos as misinformative for discussing the origin theories about the virus (such videos were labeled as "On the COVID-19 origins in Wuhan, China" – see Appendix Table 10); however, we labeled videos as misinformative if they promoted debunked and misleading claims outlined in the report.

---

[9]The 12 bots scraped 48 queries across 4 search filters for 10 days, resulting in $48 \times 4 \times 10 \times 12 = 23,040$ SERPs. On the 8th day of the experimental run, a sock-puppet bot in Durban crashed due to technical errors, failing to collect 20 SERPs. We excluded the queries from this particular experimental run from the analysis.

[10]The Lab Leak Theory is contentious, arguing that COVID-19 leaked from a lab in Wuhan, China.

[11]The Natural Origins Theory contests that COVID-19 spread to humans from animals, such as bats and pangolins.

### Annotation Scale and Heuristics

Developing the qualitative coding scheme to label YouTube videos for COVID-19 misinformation was challenging, requiring multiple iterations and discussions with external researchers to refine the heuristics. In the first iteration, the first author sampled 80 videos from the audit experiment and annotated the videos. After multiple iterations analyzing each video, the author created an initial 7-point annotation scale: "Opposing COVID-19 Misinformation (-1)," "Neutral COVID-19 Information (0)," "Supporting COVID-19 Misinformation (1)", "On the COVID-19 origins in Wuhan, China (2)," "Irrelevant (3)," "Video in a language other than English (4)," and "URL not accessible (5)." Next, seven external researchers with extensive work experience in online misinformation independently annotated 13 videos and provided feedback on our annotation criteria and tasks. After discussion and incorporating their feedback, we further refined the annotation heuristics. Due to space constraints, please see Appendix Table 10 for the 7-point annotation labels, descriptions, and example videos.

## 5 Labeling YouTube Videos

After developing our data annotation scheme, we labeled the videos. Given the large amounts of data (10,139 videos), we scaled the labeling process using a machine-learning classifier for English videos. We constructed the ground-truth dataset, trained and evaluated 62 different classifiers, and separately handled videos in non-English languages.

### Creating the Ground-Truth Dataset

We obtained annotations for 3,075 videos, which were annotated by the first author and Amazon Mechanical Turk (AMT). The first author, as the expert, annotated 1,087 videos.[12] For 1,988 videos, we obtained three AMT worker annotations per video. To assess the AMT workers' agreement, we calculated the Fleiss' Kappa Score and obtained $\kappa = 0.62$, indicating "substantial agreement." We employed the majority response to assign the final label, arriving at a final label for 1,899 videos. For the remaining 89 videos, all three AMT worker responses diverged. The first author annotated the 89 videos to obtain the final annotation values. In §6, we address the steps to minimize potential harm associated with exposing AMT workers to misinformation. See Appendix F for the AMT worker training, screening, compensation, and annotation task. Overall, the ground-truth dataset consisted of 3,075 videos, 820 of which were supporting, 837 opposing, 431 neutral, 409 irrelevant, 317 non-English, 228 on COVID-19 origins, and 33 URLs not accessible. To train our classifier, we excluded videos annotated as non-English and URL not accessible.

### Training and Applying Classifier to English Videos

Using the ground-truth dataset, we trained 62 different classifiers to find the best-performing model for our task.

---

[12]To gather more annotation, the first author followed Tomlein et al. (2021) and reannotated Medina Serrano, Papakyriakopoulos, and Hegelich (2020)'s YouTube video dataset, resulting in 143 more annotations.

| Model | Acc. | F1-M | F1-W |
|---|---|---|---|
| SVM | 0.78 | 0.78 | 0.78 |
| XGB | 0.75 | 0.75 | 0.76 |
| DeBerta (base) | 0.81 | 0.81 | 0.81 |
| DeBerta (large) | **0.85** | **0.85** | **0.85** |
| GPT-4 Turbo | 0.79 | 0.79 | 0.79 |

Table 2: The best performance achieved by each model on the three-class classification task: Support Vector Machine (SVM), XGBoost (XGB), DeBerta-v3-base, DeBerta-v3-large, and GPT-4 Turbo (v1106). All model performances are evaluated on the same 10% held-out test set. See Appendix Table 11 for the performance results of all 62 trained models. Note that Acc.: Accuracy, F1-M: Macro F1-score, F1-W: Weighted F1-score.

**Consolidating From 5-Classes to 3-Classes.** Developing a classifier to detect COVID-19 misinformation in YouTube videos was difficult, requiring experimentation with several models. Initially, we trained classifiers that predicted five classes: opposing, supporting, neutral, COVID-19 origins, and irrelevant. Such a model could only achieve an accuracy of 0.71. We sought to improve our classifier's accuracy by reducing the number of classes, while maintaining our study's objective of measuring the level of COVID-19 misinformation in YouTube SERPs. Thus, we merged the classes neutral, irrelevant, and COVID-19 origins into a single category as these classes *neither* support nor oppose COVID-19 misinformation. This consolidation yielded a classification task with 3-classes: supporting misinformation, opposing misinformation, and neither.

**Training and Evaluating Classifiers** To find the best-performing classifier, we systematically trained and evaluated various machine learning models, deep learning models, and LLMs on the three-class classification task, using combinations of text-based video metadata features such as titles, descriptions, transcripts, tags, and comments. Models included Support Vector Machine (Cortes 1995), XGBoost (Chen and Guestrin 2016), DeBerta-v3-base and DeBerta-v3-large (He, Gao, and Chen 2021), and GPT-4 Turbo (v1106) (OpenAI et al. 2024). We held out 10% of the ground-truth dataset as our test set and used accuracy, weighted F1-score, and macro F1-score to evaluate the model performances. The remaining 90% of the dataset was employed for training and validation. See Appendix G for the input feature descriptions and training procedure.

We evaluated and trained 62 models with different feature combinations and settings. For brevity, Table 2 presents the highest performance results achieved by each model. The DeBerta-v3-large model outperformed other models across all performance metrics on the held-out test set, obtaining a test accuracy, weighted F1, and macro F1 of 0.85. See Appendix Table 11 for the performance results of all 62 models. In Appendix H, we provide an in-depth discussion on model results and important input features for detecting COVID-19 misinformation in YouTube videos.

**Remarks.** Our best-performing model on the three-class classification task achieves a comparable or even exceeds the performance of other models from prior studies (Ng, Hoffmann Pham, and Luengo-Oroz 2023; Hou et al. 2020; Medina Serrano, Papakyriakopoulos, and Hegelich 2020; Yi Liaw et al. 2023). For example, Papadamou et al. (2022) developed a binary classifier that detected pseudoscientific videos with an accuracy of 0.79 and an F1-score of 0.74. Nevertheless, we acknowledge that this is not a perfect performance, reflecting the complex nature of identifying COVID-19 misinformation within videos and signaling the need for further research efforts. With a 15% error rate in our classifier, we took additional steps to validate the reliability of our findings due to potential labeling errors (see subsection "Validation of Results" in Appendix K).

### Handling Videos in Non-English Languages

Our classifier, trained exclusively on English videos, was applied only to annotate English videos within the remaining unlabeled portion of the dataset. Non-English videos were manually annotated separately. To identify non-English videos in the remaining dataset, we employed two tools: Google Translate's *Language Detection API* (Google 2023) and *langdetect*[13] library to predict the language of the video based on text-based metadata such as the title. For videos flagged as non-English by any of the tools, we manually verified to confirm the video's language. After identifying non-English videos in the remaining dataset, we merged them with the 317 non-English videos from the ground-truth dataset, totaling 784 confirmed non-English videos in the entire dataset. We manually annotated each video using Google Translate and referenced external researchers at our institution fluent in the respective languages.

## 6    Quantifying Misinformation Bias

To quantify the misinformation present in YouTube SERPs, we adopted the misinformation bias score metric from Hussein, Juneja, and Mitra (2020). The score determines the misinformation bias in a ranked list, giving *more weight* to the annotation labels of *higher-ranked videos* in the calculation: $\frac{\sum_{r=1}^{n}(x_r * (n-r+1))}{\frac{n*(n+1)}{2}}$, where $x$ represents the annotation label of the video, $r$ represents the rank of the video in the SERP (e.g., $r = 1$ indicates the top-most video in the SERP), and $n$ represents the total number of videos in the SERP. To conform to the video annotation scale in Hussein, Juneja, and Mitra (2020), we mapped our 3-class labels from §5 to a normalized scale of -1, 0, and 1 based on their stance towards COVID-19 misinformation. Videos that oppose COVID-19 misinformation were assigned scores of -1, while those supporting it received a score of 1. However, videos that fell into the merged category, including irrelevant, neutral, and COVID-19 origins labels, do not support nor oppose COVID-19 misinformation. Thus, they[14] re-

---

[13]https://pypi.org/project/langdetect/

[14]We also assigned a 0 score to removed videos whose URL was inaccessible (<1% of data), providing a conservative estimate of the misinformation bias in SERPs.
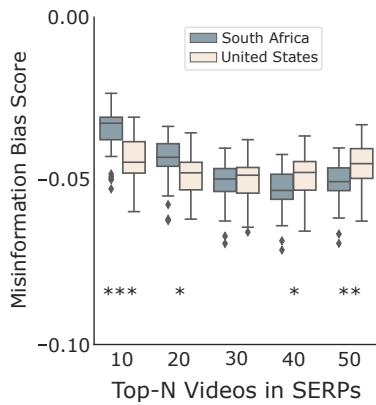
Figure 2: Distribution of the mean misinformation bias scores for the top-10 to top-50 videos in SERPs across the US and SA geolocations. These scores were computed considering the top number of videos (N) in the SERPs. Scores greater than 0 indicate that the SERPs lean toward supporting misinformation, while scores below 0 suggest SERPs lean toward opposing misinformation. Higher scores reflect a greater prevalence of misinformative videos. To compare scores between the US and SA at each level, we performed Mann-Whitney U Tests with test statistics in Appendix Table 8. Note that: $*p<0.05$; $**p<0.01$; $***p<0.001$.

ceived a score of 0. Therefore, the misinformation bias score of a SERP is a continuous value ranging from -1 (all videos oppose misinformation) to +1 (all videos support misinformation). Positive scores indicate a lean toward misinformation, negative scores indicate a lean toward opposing misinformation, and 0 suggests a set of content that neither supports nor opposes misinformation. A higher score suggests a higher prevalence of misinformation in the SERP.

# 7 Results

Here, we compare the prevalence of misinformation in YouTube SERPs across geolocations, topics, and filters. A test of normality revealed that our data is not normal. Thus, we used the non-parametric Mann-Whitney U Test for pairwise comparisons and Kruskal-Wallis tests, followed by the post-hoc Conover-Iman tests with Bonferroni adjustment for multiple comparisons. We provide details of all significance tests in Appendix I. Additionally, we include additional analyses on the misinformation bias scores in search queries and temporal trends of misinformation bias between the US and SA in Appendix J. Lastly, we conducted validation checks to reinforce the robustness of our findings, accounting for the classifier's 15% error rate and alternative treatments of the COVID-19 origins class in Appendix K.

## Misinformation Bias Across Geolocations

Figure 2 displays the distribution of the mean misinformation bias scores for the top-10 to top-50 videos in SERPs between geolocations in the US and SA. To get the mean misinformation bias scores, we computed the bias scores considering the top-N videos in the SERPs, averaging across all

queries, filters, and bots at a particular geolocation.

**Among the top-N videos, the top-10 videos in the SERPs have the highest misinformation bias scores.** We observed that the top-10 videos in the SERPs have the highest misinformation bias scores for both the US and SA (see Figure 2). This suggests that misinformative content is more prevalent in top-10 search results than lower-ranked results. Overall, the misinformation bias scores across the top-N videos in the SERPs are consistently near -0.05, indicating an general trend towards neutral-to-opposing misinformation in both the US and SA. However, we observe that nearly a third of the top-10 search results were misinformative: 31.55% of the search results supported COVID-19 misinformation, 36.03% were opposing, and 32.42% belonged to the remaining classes.

**Bots in SA encountered significantly more misinformative SERPs than bots in the US for the top-10 and top-20 videos.** For the top-10 videos, the effect size $r$ was 0.49 (see Appendix Table 8), indicating a "medium effect" (Cohen 1988). This effect size indicates that our observed differences moderately carry practical significance, suggesting that geolocation influences the algorithmic behaviors of YouTube search — in our case, resulting in statistically significantly more misinformative SERPs in SA compared to the US. We discuss the potential implications of the observed differences in §8.

**When considering top-40 and top-50 search results, bots in the US encountered significantly more misinformative SERPs than bots in SA.** This suggests that misinformative videos are still present in the US, but are downranked to lower positions in SERPs, indicating that YouTube's content moderation strategy in the US mitigates misinformation through algorithmic adjustments—downranking such content and reducing their visibility rather than outright removal. However, it is important to note that 95% of user traffic is directed towards the first page of search results (Kaye 2013), where the top-10 results reside, making it more likely for users to interact with these results than with lower-ranked results. As a result, the difference in the top-40 and top-50 results may have less impact on user exposure compared to the top-10 results, where SA users are more likely to encounter misinformation. Since video viewers are most likely to engage with the top-10 videos in SERPs, our subsequent analysis focuses on these search results.

## Misinformation Bias Within Each Country

How do the geolocations within each country influence the prevalence of misinformation in YouTube SERPs? To answer, we conducted a Kruskal-Wallis H Test to examine differences in misinformation bias scores across geolocations within each country. The test revealed a significant difference for US geolocations (KW H(2, N=30)=6.98, $p<0.01$, $\eta^2$=0.18), but no significant difference within SA. The large effect size for the US ($\eta^2$=0.18) (Tomczak et al. 2014) suggests that the observed differences carry practical significance, indicating that geolocation may influence the prevalence of misinformation in YouTube SERPs in the US. We conducted a post-hoc Conover-Iman test with Bonferroni adjustment, revealing that bots in Houston received a higher

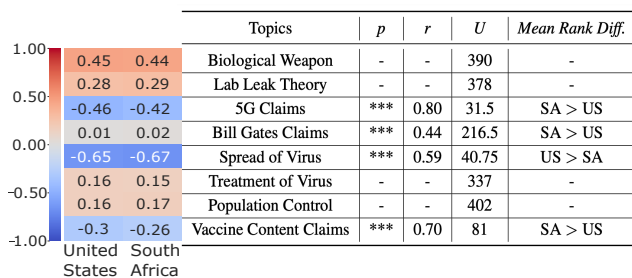| | United States | South Africa | Topics | $p$ | $r$ | $U$ | Mean Rank Diff. |
|---|---|---|---|---|---|---|---|
| | 0.45 | 0.44 | Biological Weapon | - | - | 390 | - |
| | 0.28 | 0.29 | Lab Leak Theory | - | - | 378 | - |
| | -0.46 | -0.42 | 5G Claims | *** | 0.80 | 31.5 | SA > US |
| | 0.01 | 0.02 | Bill Gates Claims | *** | 0.44 | 216.5 | SA > US |
| | -0.65 | -0.67 | Spread of Virus | *** | 0.59 | 40.75 | US > SA |
| | 0.16 | 0.15 | Treatment of Virus | - | - | 337 | - |
| | 0.16 | 0.17 | Population Control | - | - | 402 | - |
| | -0.3 | -0.26 | Vaccine Content Claims | *** | 0.70 | 81 | SA > US |

Figure 3: For each topic, we indicate the average misinformation bias scores of the top-10 search results between the US and SA (heatmap) and conduct a Mann-Whitney U Test to compare these bias scores between the two countries. We denote the p-value ($p$), Mann-Whitney effect size ($r$), $U$-value, and the mean rank difference. For example, for the "5G Claims" topic, SA > US in the "Mean Rank Diff." column indicates that bots in SA received more misinformative videos in the top-10 search results than bots in the US. Note that: *$p<0.05$; **$p<0.01$; ***$p<0.001$.



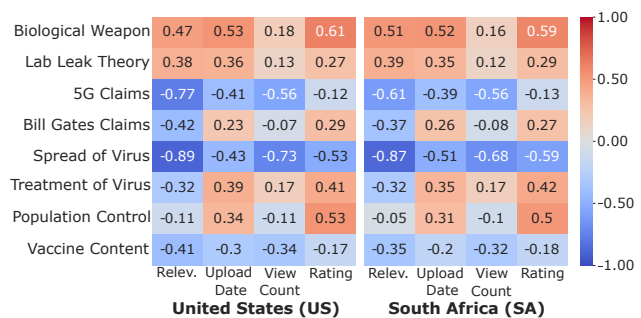| | United States (US) | | | | South Africa (SA) | | | |
|---|---|---|---|---|---|---|---|---|
| | Relev. | Upload Date | View Count | Rating | Relev. | Upload Date | View Count | Rating |
| Biological Weapon | 0.47 | 0.53 | 0.18 | 0.61 | 0.51 | 0.52 | 0.16 | 0.59 |
| Lab Leak Theory | 0.38 | 0.36 | 0.13 | 0.27 | 0.39 | 0.35 | 0.12 | 0.29 |
| 5G Claims | -0.77 | -0.41 | -0.56 | -0.12 | -0.61 | -0.39 | -0.56 | -0.13 |
| Bill Gates Claims | -0.42 | 0.23 | -0.07 | 0.29 | -0.37 | 0.26 | -0.08 | 0.27 |
| Spread of Virus | -0.89 | -0.43 | -0.73 | -0.53 | -0.87 | -0.51 | -0.68 | -0.59 |
| Treatment of Virus | -0.32 | 0.39 | 0.17 | 0.41 | -0.32 | 0.35 | 0.17 | 0.42 |
| Population Control | -0.11 | 0.34 | -0.11 | 0.53 | -0.05 | 0.31 | -0.1 | 0.5 |
| Vaccine Content | -0.41 | -0.3 | -0.34 | -0.17 | -0.35 | -0.2 | -0.32 | -0.18 |

Figure 4: Mean misinformation bias scores for the top-10 search results across all 8 topics and 4 search filters between the US and SA. Note that "Relevance" is YouTube's default sorting filter for search results. "Relev." : Relevance, "Vaccine Content": Vaccine Content Claims.

prevalence of misinformation in their search results than in Los Angeles. We discuss the implications in §8.

## Misinformation Bias in Topics

Figure 3 shows the mean misinformation bias scores and the Mann-Whitney U Test results between the US and SA across 8 topics. The scores in the heatmap were computed by averaging across each topic's constituting queries, search filters, bots within a country, and the 10-day experimental period. Figure 4 shows the misinformation bias scores across the topics, search filters, and countries.

**Bots in SA encountered significantly more misinformative SERPs than bots in the US for 3 topics.** In Figure 3, the topics "5G claims" ($r$=0.80), "Bill Gates Claims" ($r$=0.44), and "Vaccine Content Claims" ($r$=0.59) exhibited statistically significant differences in misinformation bias between the US and SA, indicating medium to large effect sizes. These effect sizes imply practical significance, suggesting that users in SA may encounter significantly more misinformative SERPs than users in the US for these topics. Of particular concern is the presence of these differences in topics related to public health, such as "Vaccine Content Claims." In contrast, bots in the US encountered significantly more misinformative SERPs than bots in the SA for a single topic—"Spread of Virus." We discuss the implications of these findings in §8.

**Topics "Biological Weapon" and "Lab Leak Theory" exhibited positive scores across all search filters in both countries** (see Figure 4). Given that positive scores indicate that the SERPs lean toward misinformative content, this finding suggests the problematic nature of these topics, with bots exposed to misinformative SERPs regardless of their geolocations or the search filters used.

**Topics "5G Claims," "Spread of Virus," and "Vaccine Content Claims" displayed negative scores across all search filters in both countries.** This suggests that SERPs for these topics generally oppose misinformation in both

the US and SA. This result may be attributed to YouTube's targeted content moderation efforts on these topics (Hern 2020), potentially highlighting some successes in combating misinformation on its platform.

**Three topics showed negative scores in SERPs sorted by "Relevance" and "View Count," but positive scores when sorted by "Upload Date" and "Rating."** We observe this pattern for "Bill Gates Claims, "Treatment of Virus," and "Population Control" topics, indicating that bots encounter more misinformative SERPs when searching for newly uploaded or highly rated videos for such topics.

## Misinformation Bias in Search Filters

Figure 5 depicts the mean misinformation bias scores and the Mann-Whitney U Test results between the US and SA across 4 search filters.

**Both "Upload Date" and "Rating" filters showed positive scores in both countries.** This indicates that SERPs sorted for highly rated (more likes than dislikes) or newly uploaded videos lean towards misinformative content. Our findings suggest that highly rated and newly uploaded videos tend to feature misinformation. This observation suggests that YouTube may not prioritize its content moderation efforts toward recently uploaded or highly rated videos.

**In contrast, "Relevance" and "View Count" filters showed negative scores in both countries.** This suggests that SERPs sorted by relevant or most-viewed videos lean towards opposing misinformation. This may suggest that YouTube's content moderation efforts prioritize relevant and highly viewed videos, which are likely to be surfaced by the search engine or have gained viewer attention.

**Bots in SA received significantly more misinformative SERPs than bots in the US when sorting the results by "Relevance."** As shown in Figure 5, the effect size of $r$=0.86 indicates a large effect, suggesting that users in SA may encounter more misinformative SERPs than their US counterparts when using the default "Relevance" filter.

**However, bots in the US received significantly more misinformative SERPs than those in SA when sorting the results by "Rating."** The effect size $r$=0.48 indicates a medium effect. While this effect size is not as pronounced as

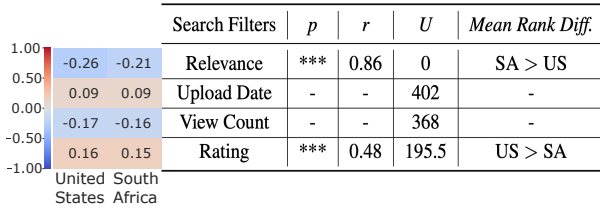| | | Search Filters | *p* | *r* | *U* | *Mean Rank Diff.* |
|---|---|---|---|---|---|---|
| -0.26 | -0.21 | Relevance | *** | 0.86 | 0 | SA > US |
| 0.09 | 0.09 | Upload Date | - | - | 402 | - |
| -0.17 | -0.16 | View Count | - | - | 368 | - |
| 0.16 | 0.15 | Rating | *** | 0.48 | 195.5 | US > SA |

United States  South Africa

Figure 5: For each search filter, we indicate the average misinformation bias scores of the top-10 search results between the US and SA (heatmap) and conduct a Mann-Whitney U Test to compare these bias scores between the two countries. *p<0.05; **p<0.01; ***p<0.001.

when sorting by "Relevance" (*r=0.86*), it suggests that the observed differences have moderate practical significance. It is important to note that YouTube's default filter is "Relevance," suggesting that users are more likely to engage with SERPs sorted by this criterion. Consequently, the observed differences with the "Rating" filter may have a lesser impact on users compared to the "Relevance" filter, where bots in SA received significantly more misinformative SERPs than bots in the US.

# 8 Discussion

**COVID-19 Misinformation on YouTube:** Auditing YouTube's search engine for COVID-19 misinformation is urgently needed to ensure algorithmic accountability, protect public health, and develop more responsible web. Despite being designated a global pandemic in 2020, COVID-19 remains a public health threat (Colarossi 2024). While YouTube has been a valuable tool for finding health information during the pandemic, it has also disseminated harmful misinformation, impeding public health efforts and vaccine hesitancy (Do Nascimento et al. 2022).

Our audit, conducted in early 2023, uncovered that 31.55% of the top-10 search results contained COVID-19 misinformation. This percentage aligns with a 2020 study (Li et al. 2020), which identified that over 25% of the most-viewed YouTube videos contain misleading information. Despite the pandemic being declared in 2020, our findings indicate that COVID-19 misinformation remains pervasive in YouTube search results three years later.

Notably, topics like "Biological Weapon" and "Lab Leak Theory" were consistently contaminated with COVID-19 misinformation across all search filters in both countries (Figure 4). These results indicate the problematic and global nature of these misinformation topics, highlighting the urgent need for YouTube to enhance its content moderation in these areas. On the other hand, topics "5G Claims," "Spread of Virus," and "Vaccine Content Claims" contained opposing search results across all filters in both countries. This finding may be attributed to YouTube's content moderation policies and its commitment to reducing misinformative content relating to 5G conspiracy, vaccines, and information that contradicts medical authorities (Hern 2020; Pruitt-Young 2021; Google 2020). Overall, our study suggests that while YouTube may be successful in moderating content on

some topics, it still has considerable work to do in moderating COVID-19 misinformation in its search engine.

**Misinformation in Emerging and Highly-Rated Content:** We observed an alarming trend in both countries for certain filters such as "Upload Date" and "Rating." As illustrated in Figure 5, the misinformation bias scores were positive for these filters, indicating that emerging and highly rated videos tend to feature misinformation. This is problematic, as users may place trust in videos highly rated by others, and users seeking the latest information about the pandemic may be at heightened risk of encountering misinformation. On August 25th, 2023, the European Union's Digital Services Act went into effect, placing responsibility on online search engines, including YouTube, for the misinformative content on the platforms (European Commission 2023). In light of these regulations, our study provides valuable insights for YouTube to enhance its content moderation practices, especially in addressing emerging or highly rated content that may contribute to the spread of COVID-19 misinformation and, consequently, impact public health.

**Contrasting Algorithmic Behaviors of YouTube Search in Different Geolocations:** We observe several instances of statistically significant disparities, with bots in SA encountering significantly more misinformative SERPs than bots in the US on YouTube, particularly within the top-10 search results (p<0.001, *r=0.49*) and search results sorted by the "Relevance" filter (p<0.001, *r=0.86*). Since users are likely to engage with these search results (see §7), this suggests that users in SA may encounter significantly more misinformative SERPs than users in the US. Considering the established importance of YouTube for finding health-related information (Khatri et al. 2020), exposing users in SA to more misinformation in their daily YouTube usage may increase their personal health risks by potentially negatively influencing their beliefs, health practices, and decisions.

Furthermore, our results indicate that bots in SA were exposed to significantly more misinformative SERPs compared to bots in the US for the topics "5G Claims," "Bill Gates Claims," and "Vaccine Content Claims" (Figure 3). Notably, South Africa experienced distressing incidents during the pandemic, such as the burning of 5G towers fueled by a conspiracy theory linking them to the spread of COVID-19 (Reuters 2021). Concurrently, South Africa grappled with many public health challenges, including a vaccine supply shortage and a deep distrust of medical authorities contributing to vaccine hesitancy (Chutel and Fisher 2021). Given these circumstances, the heightened exposure of SA users to misinformation compared to US users on these topics could potentially exacerbate societal issues in SA, including vaccine hesitancy and undermining public health efforts.

Within the US, our analysis revealed that bots in Houston, Texas, received significantly more misinformative SERPs than those in Los Angeles, California. Prior work has shown that conservatives were more susceptible to believing falsehoods and conspiracy beliefs (Garrett and Bond 2021). Since Houston is in a conservative-leaning state (Texas), promoting significantly more misinformative search results to conservative-leaning users could exacerbate harm by rein-

forcing or fostering beliefs in COVID-19 misinformation.

However, political leaning alone is unlikely to explain the differences observed within the US. Other factors, such as regional search behaviors, public health policies, and online media use policies may also play a role. The observed geographic differences within the US, as well as between the US and SA, highlight important societal and policy implications for platforms like YouTube. These platforms should assess whether current moderation policies sufficiently address geographic disparities in the prevalence of COVID-19 misinformation in search results. To mitigate these disparities, platforms could analyze region-specific search behaviors and trends to adjust rankings or visibility for content likely to contain misinformation. Additionally, regionally tailored counter-misinformation efforts—such as promoting public health campaigns in high-risk areas—could help curb misinformation across diverse regions.

**The Challenges and Importance of Conducting Audits in the Global South:** While algorithmic audits have been predominantly focused on the Global North contexts, there has been limited exploration into audits concerning the Global South. Our research explores the less-researched Global South context by conducting the first large-scale geolocation-based audit of YouTube search for COVID-19 misinformation between the US and SA.

However, we must acknowledge the challenge posed by the limited availability of affordable proxies, VPNs, and virtual machines in Global South countries. This makes large-scale audit experiments in the Global South either cost-prohibitive or, in some cases, logistically impossible. Addressing this issue underscores the pressing need for enhanced technological access and affordability to facilitate more *extensive* research initiatives within the Global South, such as conducting audits in multiple countries in the region.

Disparate algorithmic behaviors based on geolocation raise concerns regarding fairness, especially when returning significantly more misinformative SERPs in one region over another. Our findings serve as an important call to action for YouTube to regulate its algorithmic behavior consistently to ensure equitable and accurate information dissemination across different regions. This proactive approach is crucial to facilitate the development of a more responsible and fair video search engine that preserves factual and accurate information irrespective of the user's geolocation.

Historically, past audits have successfully generated awareness about these issues, creating pressure on platforms to improve their algorithm (Raji and Buolamwini 2019). Similarly, we hope our work drives meaningful changes on YouTube and inspires future audits in the Global South.

## 9 Limitations and Future Directions

Our work is not without limitations. Our audit used only English search queries to maintain consistency in search query language across both countries. Future studies can audit YouTube's search results by utilizing queries in diverse regional and local languages within the countries. In addition, our study focuses on YouTube's search engine; future studies can investigate the impacts of geolocation on the recommendation systems on the home page and video page. As the COVID-19 pandemic remains a public health threat globally, our study focuses on COVID-19 misinformation on YouTube. However, other domains, such as climate change misinformation, warrants attention—our methods, including the experimental setup, can be adapted to investigate misinformation beyond COVID-19.

In our study, we employ sock-puppet bots to emulate the actions of real users, which avoids contaminating real users' search and recommendation results (explained further in §6). However, it is important to note that while we can analyze what happens when a hypothetical user searches on YouTube, we cannot determine the prevalence of misinformation for *real* users in South Africa compared to the United States. To address this limitation, future work could ethically and carefully conduct *crowdsourced* audits (Juneja, Bhuiyan, and Mitra 2023) to gather real user data and compare misinformation prevalence in these different geographic regions.

Given the challenges of conducting audits in the Global South, our audit focuses on a country from the Global North and one from the Global South. Future works should conduct audits in multiple countries within the Global South. Our work uses the misinformation bias score (Hussein, Juneja, and Mitra 2020), capturing the amount of misinformation and the video's rank but not its relevance. Future works can use metrics like the Convex Aggregated Measure (Lioma, Simonsen, and Larsen 2017), which uses relevance and credibility in ranked lists.

Although we showed in §5 that our classifier performs comparably to YouTube models tackling related issues, we acknowledge that our classifier has an error rate of 15%, potentially influencing downstream misinformation analysis. We partly address this limitation by conducting additional analyses, which provided further evidence of the reliability of our findings despite the error rate of our classifier (see subsection "Validation of Results" in Appendix K).

## 10 Conclusion

In this study, we conducted an audit on YouTube comparing the prevalence of COVID-19 misinformation in SERPs between the US and SA. Our findings revealed that overall, 31.55% of the top-10 search results contained COVID-19 misinformation. Among these top-10 search results, we found that bots in SA encountered significantly more misinformative SERPs than bots in the US. Overall, our study highlights the potential need to regulate algorithmic behavior consistently across different regions and calls for future research efforts to consider contexts in the Global South.

## Acknowledgements

# References

Abdelfatah, R.; and Arablouei, R. 2024. The story of a scientist who tried to stand for the truth and avoid Covid politics. https://www.npr.org/2024/09/24/nx-s1-5121198/the-story-of-a-scientist-who-tried-to-stand-for-the-truth-and-avoid-covid-politics. Accessed: 2024-09-24.

Adeline, S.; et al. 2020. Coronavirus update: View charts of U.S. cases and deaths. https://www.npr.org/sections/health-shots/2020/09/01/816707182/map-tracking-the-spread-of-the-coronavirus-in-the-u-s. Accessed: 2023-06-03.

Albadi, N.; Kurdi, M.; and Mishra, S. 2022. Deradicalizing YouTube: Characterization, Detection, and Personalization of Religiously Intolerant Arabic Videos. *Proc. CSCW*.

Boeker, M.; and Urman, A. 2022. An Empirical Investigation of Personalization Factors on TikTok. In *Proc. WWW*, 2298–2309. ACM. ISBN 9781450390965.

Bojanowski, P.; et al. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146.

Bora, K.; et al. 2018. Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. *Pathogens and global health*.

CDC. 2023. Coronavirus Disease 2019. https://www.cdc.gov/coronavirus/2019-ncov/index.html. Accessed: 2023-06-03.

Chen, A. Y.; et al. 2023. Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. *Science Advances*, 9(35): eadd8080.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill*, 6(2): e19273.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proc. KDD*. ACM. ISBN 9781450342322.

Chrome DevTools. 2023. Chrome DevTools. https://www.selenium.dev/documentation/webdriver/bidirectional/chrome_devtools/. Accessed: 2023-06-03.

Chutel, L.; and Fisher, M. 2021. The Next Challenge to Vaccinating Africa: Overcoming Skepticism. https://www.nytimes.com/2021/12/01/world/africa/coranavirus-vaccine-hesitancy-africa.html. Accessed: 2023-06-04.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Academic press.

Colarossi, J. 2024. Is COVID-19 Still a Pandemic? https://www.bu.edu/articles/2024/is-covid-19-still-a-pandemic. Accessed: 2024-05-14.

Cortes, C. 1995. Support-Vector Networks. *Mach. Learn.*

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proc. FAccT*, 1571–1583. ACM. ISBN 9781450393522.

Dabran-Zivan; et al. 2023. "Is COVID-19 a hoax?": auditing the quality of COVID-19 conspiracy-related information and misinformation in Google search results in four languages. *Internet Research*, 33(5): 1774–1801.

Dammu, P. P. S.; Jung, H.; Singh, A.; Choudhury, M.; and Mitra, T. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20339–20369. Association for Computational Linguistics.

Diakopoulos, N.; et al. 2018. I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, 22: 320–341.

Do Nascimento, I. J. B.; et al. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9): 544.

Doochin, D. 2019. What Are The Languages Spoken In South Africa? https://www.babbel.com/en/magazine/languages-in-south-africa. Acccessed: 2023-09-26.

Dubey, D.; et al. 2014. Analysis of YouTube as a source of information for West Nile virus infection. *Clinical medicine & research*.

European Commission. 2023. Digital Services Act. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng. Accessed: 2023-10-11.

Faddoul, M.; et al. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv:2003.03318*.

Garrett, R. K.; and Bond, R. M. 2021. Conservatives' susceptibility to political misperceptions. *Science Advances*, 7(23): eabf1234.

Ghebreyesus, T. A. 2023. WHO Director-General's opening remarks at the media briefing. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023. Accessed: 2024-05-14.

Google. 2020. COVID-19 misinformation policy. https://support.google.com/youtube/answer/9891785. Accessed: 2023-06-03.

Google. 2022. Add tags to your YouTube videos. https://support.google.com/youtube/answer/146402. Accessed: 2022-07-31.

Google. 2023. Detecting languages (Advanced). https://cloud.google.com/translate/docs/advanced/detecting-language-v3. Accessed: 2023-10-07.

Grant, N. 2022. YouTube May Have Misinformation Blind Spots, Researchers Say. https://www.nytimes.com/2022/11/05/technology/youtube-misinformation.html. Accessed: 2023-06-03.

Grootendorst, M. 2022. KeyBERT. https://maartengr.github.io/KeyBERT/. Accessed: 2022-07-15.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hern, A. 2020. YouTube moves to limit spread of false coronavirus 5G theory. https://www.theguardian.com/world/2020/apr/05/youtube-to-suppress-content-spreading-coronavirus-5g-conspiracy-theory. Accessed: 2023-06-04.

Hou, R.; et al. 2020. Towards Automatic Detection of Misinformation in Online Medical Videos. In *2019 ICMI*, 235–243. Association for Computing Machinery. ISBN 9781450368605.

Hussein, E.; Juneja, P.; and Mitra, T. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. CSCW*.

IP2Location. 2023. IP Address to IP Location. https://www.ip2location.com/. Accessed: 2023-06-03.

IPRoyal. 2023. Premium Quality Proxies. https://iproyal.com/. Accessed: 2023-06-03.

Juneja, P.; Bhuiyan, M. M.; and Mitra, T. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. In *Proc. CHI*. ACM. ISBN 9781450394215.

Juneja, P.; and Mitra, T. 2021. Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation. In *Proc. CHI*. ACM. ISBN 9781450380966.

Kaye, L. 2013. 95 percent of web traffic goes to sites on Page 1 of Google SERPs. https://www.brafton.com/news/95-percent-of-web-traffic-goes-to-sites-on-page-1-of-google-serps-study/. Accessed: 2023-06-03.

Khatri, P.; et al. 2020. YouTube as source of information on 2019 novel coronavirus outbreak: a cross sectional study of English and Mandarin content. *Travel Medicine and Infectious Disease*, 35.

Kliman-Silver, C.; et al. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proc. of the 2015 IMC*, 121–127. ACM. ISBN 9781450338486.

Klimek, L. 2021. 12 Ways to hide your Bot Automation from Detection. https://piprogramming.org/articles/How-to-make-Selenium-undetectable-and-stealth--7-Ways-to-hide-your-Bot-Automation-from-Detection.html. Accessed: 2023-06-03.

Knobloch-Westerwick, S.; et al. 2015. Science Exemplars in the Eye of the Beholder: How Exposure to Online Science Information Affects Attitudes. *Science Communication*, 37(5): 575–601.

Kravets, D.; Ryzhova, A.; Toepfl, F.; and Beseler, A. 2023. Different platforms, different plots? The Kremlin-controlled search engine Yandex as a resource for Russia's informational influence in Belarus during the COVID-19 pandemic. *Journalism*, 24(12).

Le, B.; et al. 2022. A crowdsourcing methodology to measure algorithmic bias in black-box systems: A case study with COVID-related searches. In *BIAS*. Springer.

Li, H. O.-Y.; et al. 2020. YouTube as a source of information on COVID-19: a pandemic of misinformation? *BMJ Global Health*.

Lioma, C.; Simonsen, J. G.; and Larsen, B. 2017. Evaluation Measures for Relevance and Credibility in Ranked Lists. In *Proc. of the ACM SIGIR ICTIR*, 91–98. ACM. ISBN 9781450344906.

Lu, X. H. 2022. A collection of COVID-19 question-answer pairs and transformer baselines for evaluating QA models.

Medina Serrano, J. C.; Papakyriakopoulos, O.; and Hegelich, S. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: ACL.

Mikolov, T.; et al. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Milmo, D. 2022. YouTube is major conduit of fake news, factcheckers say. https://www.theguardian.com/technology/2022/jan/12/youtube-is-major-conduit-of-fake-news-factcheckers-say. Accessed: 2023-06-03.

Mishra, S.; and Chatterjee, P. 2024. Exploring ChatGPT for Toxicity Detection in GitHub. In *Proc. of the 2024 ACM/IEEE 44th ICSE: NIER*. ACM. ISBN 9798400705007.

Möller, T.; et al. 2020. COVID-QA: A Question Answering Dataset for COVID-19. In *Proc. NLP-COVID19*. ACL.

Narain, K.; et al. 2023. COVID-19 information on YouTube: analysis of quality and reliability of videos in eleven widely spoken languages across Africa. *Global Health, Epidemiology and Genomics*.

Ng, Y. M. M.; Hoffmann Pham, K.; and Luengo-Oroz, M. 2023. Exploring YouTube's Recommendation System in the Context of COVID-19 Vaccines: Computational and Comparative Analysis of Video Trajectories. *J Med Internet Res*, 25: e49061.

Nguyen, N.; and Scurato, C. 2021. Facebook and YouTube's refusal to moderate misinformation in global languages harms communities of color. https://prismreports.org/2021/11/02/facebook-and-youtubes-refusal-to-moderate-misinformation-in-global-languages-harms-communities-of-color/. Accessed: 2023-06-03.

ODNI. 2023. Updated Assessment On COVID Origins. https://www.dni.gov/files/ODNI/documents/assessments/Declassified-Assessment-on-COVID-19-Origins.pdf. Accessed: 2023-05-30.

OpenAI. 2024. Best practices for prompt engineering. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api. Accessed: 2024-08-30.

OpenAI; et al. 2024. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Papadamou, K.; et al. 2022. "It Is Just a Flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *Proc. ICWSM*, 16(1): 723–734.

Park, C. Y.; Li, S. S.; Jung, H.; Volkova, S.; Mitra, T.; Jurgens, D.; and Tsvetkov, Y. 2024. ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 16659–16695. Association for Computational Linguistics.

Patra, D. 2020. selenium-stealth. https://pypi.org/project/selenium-stealth/. Accessed: 2023-06-03.

Pruitt-Young, S. 2021. YouTube Is Banning All Content That Spreads Vaccine Misinformation. https://www.npr.org/2021/09/29/1041493544/youtube-vaccine-misinformation-ban. Accessed: 2023-06-04.

Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Reuters. 2021. 5G COVID-19 conspiracy theory baseless and fake, S.Africa's telecoms regulator says. https://www.reuters.com/article/uk-safrica-telecoms-5g/5g-covid-19-conspiracy-theory-baseless-and-fake-s-africas-telecoms-regulator-says-idUSKBN29G2B0. Accessed: 2023-06-04.

Ribeiro, M. H.; et al. 2020. Auditing radicalization pathways on YouTube. In *Proc. FAT\**, 131–141. ACM. ISBN 9781450369367.

Robertson, R. E.; et al. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. CSCW*.

Scialom, T.; Chakrabarty, T.; and Muresan, S. 2022. Fine-tuned Language Models are Continual Learners. In *Proc. of the 2022 Conference on EMNLP*, 6107–6122. ACL.

Song, K.; et al. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proc. NIPS*. Curran Associates Inc. ISBN 9781713829546.

South African Government. 2023. Fake news - Coronavirus. https://www.gov.za/covid-19/resources/fake-news-coronavirus-covid-19. Accessed: 2023-06-03.

Tang, R.; et al. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. arXiv:2004.11339.

Tomczak, M.; et al. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sports Sciences*.

Tomlein, M.; et al. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *Proc. RecSys*, 1–11. ACM. ISBN 9781450384582.

Urman, A.; Makhortykh, M.; and Hannák, A. 2024. WEIRD Audits? Research Trends, Linguistic and Geographical Disparities in the Algorithm Audits of Online Platforms – A Systematic Literature Review. arXiv:2401.11194.

WHO. 2024. WHO Coronavirus Dashboard. https://data.who.int/dashboards/covid19/cases. Accessed: 2024-05-03.

Wits University. 2023. Provincial Breakdown of COVID-19 in South Africa. https://www.covid19sa.org/provincial-breakdown. Accessed: 2023-06-03.

Yi Liaw, S.; et al. 2023. YouNICon: YouTube's CommuNIty of Conspiracy Videos. *Proc. ICWSM*, 17(1): 1102–1111.

Zarr, O. 2020. 50+ YouTube Search Operators For YouTube Advanced Search. https://seosly.com/blog/youtube-search-operators/. Accessed: 2023-06-03.

# Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our work fills a critical gap in algorithmic audit research, providing the first large-scale geolocation-based comparative audit of YouTube search for COVID-19 misinformation between a country in the Global North and one in the South. Our work highlights fairness concerns related to the differing algorithmic behaviors based on geolocation and serves as an important call to action for YouTube to regulate its algorithmic behaviors consistently across all regions.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. We carefully reviewed the claims in the abstract and Introduction to ensure that they accurately reflect our contributions and project scope.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Please refer to Sections 3-5. We justify the appropriateness of our methods, such as our search query curation and experimental design, by referencing prior works with similar tasks/objectives. For our analysis results, we justify the use of the misinformation bias score metric based on its usage in previous studies quantifying misinformation in search engine result pages. Our classifier's performance is quantitatively validated by comparing the classifier labels with human annotations, and we contextualize this performance by referencing prior works (see subsection "Remarks" under Section 5). Overall, we justify our methodological approach throughout the paper.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. Please refer to Section 3 and Appendix Tables 3-4, where we describe the characteristics and sample data points from the *PesaCheck* dataset, which we used to derive the set of search queries for the audit experiment.

   (e) Did you describe the limitations of your work? Yes. We discuss the limitations of our work in Section 9 "Limitations and Future Directions."

   (f) Did you discuss any potential negative societal impacts of your work? Yes. Refer to Section 11 "Ethics Statement" where we discuss the potential negative impacts of our work and the steps we took to mitigate them.

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

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. In Section 11 "Ethics Statement," we followed best practices to mitigate potential harm to human annotators. We ethically sourced proxies from consenting individuals and properly deleted all IP addresses after the experiment. To ensure responsible release, we will release the classifier on Hugging-Face and data on a Github repository (see footnote 1). For reproducibility, we thoroughly document the significance testing details in Appendix Section I.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, I read the ethics review guidelines and ensured that our paper conforms to them to the best of our capabilities.

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

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

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

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

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

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

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

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

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

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

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

4. Additionally, if you ran machine learning experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, in the "Data Set" entry accompanying the submission, we provide in-depth instruction and data needed to reproduce our main experimental results. We will also make this public via Github Repository.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. In the Appendix Section G "Classifier Training Details," we discuss the training details, such as feature descriptions, training procedures, prompt design considerations, etc.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal

cluster, or cloud provider)? Yes. In the Appendix Section G "Classifier Training Details," we discuss the type of GPU employed to train the deep learning models. Most of our training did not require intensive computing or an external cloud provider.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. Please refer to the subsection "Training and Evaluating Classifiers" (under Section 5 "Labeling YouTube Videos"), where we quantitatively validate our classifier by comparing machine-generated and human-generated labels. We contextualize our model performance with the existing COVID-19 misinformation detection literature. Further, in Appendix Section K subsection "Validation of Results," we validate the reliability of our findings due to potential labeling errors by the classifier.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. In the Section "Limitations and Future Directions," we acknowledge the error rate of our classifier, which may potentially influence our downstream analysis. To mitigate such impacts, we conducted further validation in the Appendix Section K subsection "Validation of Results," which provides further evidence of the reliability of our findings despite the misclassification error rate.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. In the subsection "Training and Evaluating Classifiers," we cite the creators of all the machine learning models employed in our study. Additionally, we cite the creators of the COVID-19 QA datasets and *KeyBERT* in the Section "Curating Topics and Search Queries."

(b) Did you mention the license of the assets? No, we did not due to space constraints. The *all-mpnet-base-v2* model,[15] Selenium Python Library,[16] and COVID-19 QA dataset from Möller et al. (2020) are under the apache-2.0 license, and the remaining models, libraries, and tools employed in the study are released under the MIT license.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we included our dataset in the "Data Set" entry accompanying the submission.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Since we collected publicly available data on YouTube, explicit consent from people whose data we are curating is not required. Analyzing this retrospective data did not constitute human subjects research, and thus, informed consent is not required from the video creators.

---

[15]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[16]https://www.selenium.dev/documentation/about/copyright/

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, given our data's offensive and misleading nature, we explicitly provide a content warning after the abstract.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? No, we did not due to space constraints. We will publish the provided dataset onto a publicly-facing Github repository (F, A) in .csv format (I). The dataset will be released under the MIT license (R).

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, in the "Data Set" entry accompanying the submission, we provide a detailed README.MD file that provides an overview and data description of the datasets. We will also release the same README.MD file in the public Github repository.

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? Yes, Figures 9-13 contain the screenshots and instruction texts given to participants.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we described the potential participant risks in the Section "Ethics Statement." Following the best practices from a prior audit work (Papadamou et al. 2022), we took extensive steps to mitigate participant risks and ensure informed consent when designing our task.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. We included the estimated hourly wage paid to participants and the amount spent per participant annotation in Appendix Section F.

(d) Did you discuss how data is stored, shared, and deidentified? Yes. We describe how we processed the annotations in the Appendix Subsection "YouTube Annotation Task." We employed a majority agreement to arrive at a final label for a YouTube video. Since only the final labels are included in our dataset, no identifiable information about the annotators is shared.

## Ethical Statement

We took several steps to minimize potential harm from our audit experiments. First, we refrained from recruiting real-world users in the data collection phase of our audit experiments. According to Hussein, Juneja, and Mitra (2020), searching misinformative queries could contaminate users' searches and recommendations, which may have long-term consequences regarding what videos are recommended to participants by YouTube's algorithm. Thus, in our audit experiments, we employed sock-puppet bots that emulated the actions of actual users. Sec-

ond, to set up our bots' IP address geolocations, we used proxies from *IPRoyal*, which ethically source their proxies from consenting individuals who willingly share their bandwidth in exchange for compensation. Furthermore, we did not use the IP addresses to identify the individuals sharing their bandwidth and deleted all IP address information after the experiment.

Additionally, we acknowledge the potential harm of exposing AMT workers to misinformative content on YouTube. We took four measures to mitigate such risks and impacts to annotators. First, we explicitly detailed the labeling task instructions at the beginning of the Qualification Test, providing an early warning to the annotators regarding possible misinformative content in the video they are about to watch. We also added the same instructions and warning in the actual task itself. Second, to clarify and help annotators understand what misinformation is in their annotation task, we explicitly provided several examples of videos containing COVID-19 misinformation and also included a reference to Google's COVID-19 medical misinformation policy (Google 2020), which lists out debunked misinformation claims surrounding the COVID-19 pandemic. Third, to address any potential distress, concerns, or confusion, we included our lab email in the annotation instructions to provide a clear and easy way for annotators to contact us. Fourth, annotators could quit the task anytime and get paid for their work.

| Countries | |
|---|---|
| Australia | Algeria |
| Canada | Bangladesh |
| Hungary | China |
| Ireland | Ethiopia |
| Italy | India |
| Japan | Iraq |
| Republic of Korea | Jamaica |
| New Zealand | Kenya |
| Singapore | Malaysia |
| United Kingdom | Morocco |
| United States | Nigeria |
| Pakistan | Philippines |
| Qatar | South Africa |
| Sri Lanka | Uganda |
| United Arab Emirates | |

Table 3: *PesaCheck*'s fact-checked dataset contains 362 videos, which originate from channels based in 29 different countries. The 29 countries span both the Global North and the Global South, with 18 of the 29 countries being associated with the Global South.

## A   Model Fine-tuning

We selected the Sentence Transformer model *all-mpnet-base-v2* due to their best performance generating sentence embeddings across various tasks (Song et al. 2020). However, *all-mpnet-base-v2* was not trained on COVID-19-related texts. Scialom, Chakrabarty, and Muresan (2022)

demonstrated that language models can be "continual learners" of new domains via self-supervised training. Thus, we continued training the model using a self-supervised contrastive learning framework (i.e. multiple-negative-ranking-loss) on COVID-19 question-answering (QA) datasets (Möller et al. 2020; Lu 2022; Tang et al. 2020), totaling 3,296 QA pairs, to generate more meaningful sentence embeddings on the COVID-19-related texts. We randomly partitioned the dataset into 80% training, 10% validation, and 10% testing sets. To avoid overfitting, we trained the model until the training loss decreased further relative to the validation loss. Our best model used an Adam optimizer with a learning rate of 2e-5, 3 epochs, 32 batch size, and a warmup step for 10% of total steps. We found that our fine-tuning worked. The finetuned model had a test loss of 0.7466 and the original pre-trained model had a test loss of 1.7722, indicating that the finetuned model generated more meaningful sentence embeddings on COVID-19 pandemic-related texts.

## B   Measuring Geolocation-based Personalization

To measure geolocation-based personalization, we used the Jaccard index to calculate the similarity of two SERPs. Previous audits used the Jaccard index to measure personalization in web searches (Kliman-Silver et al. 2015; Juneja and Mitra 2021). The Jaccard index determines the similarity between two lists – 1 indicates that both lists have identical elements, while 0 indicates that both have completely different elements.

We formalize the metric to compute the geolocation-based personalization between two geolocations, location $x$ and location $y$. We indicate the SERPs collected by a treatment bot at location $x$ as $SERP1_x$ and its corresponding twin bot as $SERP2_x$. First, we establish the noise level at each geolocation by calculating the Jaccard index of SERPs collected by the identical twins at the same geolocation (Equation 1). Ideally, these twin bots should have a Jaccard index of 1 because these are identical "users" searching for the same query at the same time and geolocation; however, even after controlling for all known sources of noise, there may be some sources of noise that we are not aware of. Second, we take the higher noise level (i.e. lower Jaccard's index) between the geolocations to establish a baseline noise and obtain a more conservative estimate of geolocation-based personalization between the two geolocations (Equation 2). Third, we compute the difference in the SERPs between the two geolocations by calculating the Jaccard's index of SERPs collected by the treatment bot at each geolocation (Equation 3). Finally, we calculate the difference between the baseline noise level and Equation 3 – the additional difference beyond the baseline noise can then be attributed to geolocation-based personalization, denoted as `GBP` in Equation 4.

The `GBP` metric ranges from 0 to 1 and indicates how much of the differences between the SERPs from location $x$ and location $y$ can be attributed to geolocation-based personalization. A zero indicates that the difference between the treatments did not exceed the baseline noise, while a posi-

| Misinformation Topics | Sample YouTube Videos (Video Title and ID) |
|---|---|
| Biological Weapon | PG-18 Reasons & evidence that I infer COVID-19 as the biological weapon made by Chinese Government (*Video ID: 2uDpARUBPEE*) |
| Lab Leak Theory | Russian Flu of 1977 Linked to Lab Leak & Live Vaccine Trials Startling Similarities to COVID-19 (*Video ID: us8gd_XY2fY*) |
| 5G Claims | Why The Coronavirus Test Is Likely The Bill Gates 5G Google Chip (*Video ID: 02RVwXRsAHg*) |
| Bill Gates Claims | Rev Chris Okotie: THE COVID 19 MYSTERY (*Video ID: 8uPbf7T0KbQ*) |
| Spread of Virus | How to Stay Safe and Prevent the Spread of COVID 19 (*Video ID: 8SDneT-SzB5E*) |
| Treatment of Virus | COVID-19 SELF TREATMENT: How to determine which remedies to use (*Video ID: 92DWaf6qmWM*) |
| Population Control | W.H.O WHISTLEBLOWER: Depopulation, total control, Perpetual fear and perpetual vaccination. (*Video ID: vD6IcaAFORM*) |
| Vaccine Content Claims | COVID 19 VACCINES ARE A GIFT FROM ABORTED BABIES? CIVIL RIGHTS FOR CATHOLICS? (*Video ID: kRTv8G2VI1M*) |

Table 4: The 8 globally-persistent COVID-19 misinformation topics identified by *PesaCheck*. For each topic, we provide a sample YouTube video from *PesaCheck*'s dataset.

tive value indicates that the difference in the SERPs can be attributed to geolocation-based personalization.

$$\text{Noise}(x) = \text{Jaccard}(\text{SERP1}_x, \text{SERP2}_x) \quad (1)$$

$$\text{Baseline}(x, y) = \min(\text{Noise}(x), \text{Noise}(y)) \quad (2)$$

$$\text{Diff.}(x, y) = \text{Jaccard}(\text{SERP1}_x, \text{SERP1}_y) \quad (3)$$

$$\text{GBP}(x, y) = \text{Baseline}(x, y) - \text{Diff.}(x, y) \quad (4)$$

## C   Curating and Validating Search Queries

To conduct all our validation experiments, we curated a set of 18 search queries[17] based prior work (Kliman-Silver et al. 2015) that we thought were likely to elicit geolocation-based personalization. We also incorporated 5 random search queries curated from Section 3, giving us 23 search queries total. To validate that the search queries elicit geolocation-based personalization, we conducted a validation experiment in Seattle, Washington, US, and Cape Town, Western Cape, SA, placing a twin bot in each geolocation. We ran the experiment from July 17th to 19th, 2022 at three evenly distributed times of 00:00 UTC, 08:00 UTC, and 16:00 UTC. Using the GBP metric from Appendix B, the average GBP was 0.18, indicating that 18% of the search results differed between Cape Town, South Africa, and Seattle, Washington, even after accounting for noise. This validated that our curated search queries elicited geolocation-based personalization on YouTube. We used the search queries in the subsequent validation experiments

---

[17]Such queries were: Day in the life, City vlogs, Travel, Hiking, University, Food, School, Places nearby, Things to do nearby, news near me, Hospital, Bank, Government, Politics, Businesses, Stores, Groceries, and Bars.

## D   Validating Geospoofing Method

With a validated set of search queries, we tested whether YouTube personalizes search results based on GPS coordinates provided rather than IP addresses by comparing the search results of bots with different IP addresses but with the same provided GPS coordinates. To override the bot's current geolocation with our provided GPS coordinates, we used Chrome DevTool (Chrome DevTools 2023) to integrate the geolocation emulation feature with our Selenium bots. We created twin bots in Cape Town (SA), Paris (France), and North California (US) regions on AWS, giving us 6 bots total. We provided all bots with the same GPS coordinate in Paris, France. As such, twin bots in varying countries should have different IP geolocations, but the same geospoofed GPS coordinates. If YouTube used GPS coordinates to personalize SERPs, we would see no difference in SERPs after accounting for noise because the bots have the same GPS coordinates, resulting in a GBP metric value of 0. However, if YouTube used IP addresses, the GBP metric value would be greater than 0. From July 25th to July 27th, 2022, we ran the validation experiment at three evenly distributed times 00:00 UTC, 08:00 UTC, and 16:00 UTC. The average GBP value between North California and Paris was 0.15, while the average GBP value between Cape Town and Paris was 0.11. Our results indicate that YouTube uses IP addresses over the provided geospoofed GPS coordinates to personalize their SERPs. Given this result, we shifted towards designing our experiment using proxies.

## E   Validating Proxies for Accurate IP Location

After obtaining unsuccessful validation results from the geospoofing method, we conducted a third experiment to validate that *IPRoyal* provides the bots with accurate IP geolocations. We created 2 sets of twin bots in Cape Town,

| Misinformation Topics | Search Queries |
|---|---|
| Biological Weapon | Biological Weapon, CCP virus, man-made virus, China virus, fact chinesemade virus, revelations gravitas china virus coverup |
| Lab Leak Theory | lab leak theory, WHO coverup, china scientist create virus, North Carolina lab in US, Kungflu, evidence virus lab |
| 5G Claims | 5g bad effect, 5g conspiracy, 5g and covid19 link, the dangers of 5g radiation, why 5g testing cause corona, dangers 5g |
| Bill Gates Claims | destroy Africa, depopulate the world, vaccine testing africa, bill gates vaccine chip, bill gates exposed, happened vaccine devil |
| Spread of Virus | sanitize, dogs and cats, Covid 19 Spread And Precautions, precaution for pets, spread of covid by a bat, social spread |
| Treatment of Virus | sesame oil, garlic, herbs, local concoctions, treatment covid19 government, dealing vulnerable population |
| Population Control | population control, mass murder, plandemic, nuremberg code, brainwashing, vaccines depopulation |
| Vaccine Content Claims | MRNA, hek-293 cells, fetal tissue research, abortion used in vaccine, organ harvesting, conscience vaccines abortion |

Table 5: The final set of 48 search queries spanning 8 misinformation topics employed in our audit study. For each topic, we utilized 6 search queries associated with the topic.

South Africa on AWS, giving us 4 bots total. A set of twin bots would search regularly, while the other set of twin bots would search using the *IPRoyal* proxies set to Cape Town, South Africa. Since we are comparing bots in the same geolocation, we should ideally see no geolocation-based personalization. Additionally, we used IP2Location to keep track of IP addresses (IP2Location 2023), enabling us to validate whether the *IPRoyal* proxies are associated with our desired IP geolocation. Since the 2 sets of twin bots were at the same IP geolocation, our GBP metric value should be close to 0. From December 22nd to December 24th, 2022, we ran the validation experiment at two evenly distributed times of 00:00 UTC and 12:00 UTC. The average geolocation-based personalization value was -0.02,[18] validating that the bots had no geolocation-based personalization. Additionally, our IP2Location data also validates the consistency and accuracy of *IPRoyal* in providing proxies with valid IP geolocations in Cape Town.

## F   Amazon Mechanical Turk Job

This section describes how we obtained annotations from Amazon Mechanical Turk (AMT) workers. We implemented a rigorous screening process for AMT workers to ensure high-quality annotations. In the following sections, we briefly provide details on our screening process and the annotation task.

---

[18]Note that, our formalized metric from Appendix B may result in a slight negative value since we take the more restrictive noise to establish our baseline noise. Regardless, any slightly negative value indicates that the differences between the treatments did not exceed the baseline noise.

### Training and Screening Workers

To get high-quality annotations, we trained and screened AMT workers by adding two qualification requirements. First, to find workers with a proven track record of high-quality work, we required AMT workers to have at least 1000 approved tasks with at least a 99% approval rating on the AMT platform. Second, we required the AMT workers to get a full score of 100 on our Qualification Test. We introduced the test to train and ensure that AMT workers attempting our annotation task understood our annotation scheme well. The Qualification Test consisted of 5 questions, with one eligibility question asking them to confirm whether they are affiliated with the authors' university. The other four questions asked AMT workers to annotate YouTube videos whose annotation labels were known in advance— these videos were already annotated by the authors (see Figure 12 for a sample Qualification Test question). To ensure that the AMT workers understood the task and annotation scheme, we gave detailed instructions and described each annotation value in detail with various examples of YouTube videos in the Qualification Test (Figures 9, 10, 11). For each video example, we provided the video metadata (e.g. video title, embedded Video URL) and explained why a particular annotation label was assigned to the video (see Figure 9).

We took two steps to ensure the instructions, test questions, and overall annotation task were clear and comprehensive. First, we posted the Qualification Test and the annotation task on r/mturk,[19] a subreddit community of AMT workers, and Turker Nation, an unofficial Slack channel of AMT workers. Second, we conducted a pilot run by posting 13 annotation tasks and the aforementioned screening

---

[19]https://www.reddit.com/r/mturk/

| Proxy Provider | Price ($) | Proxy Geolocations in South Africa (City) |
| --- | --- | --- |
| Bright Data | $30 / GB of traffic | Benoni, Bloemfontein, Boksburg, Cape Town, Centurion, Ceres, Durban, Johannesburg, Pietermaritzburg, Polokwane, Port Elizabeth, Pretoria, Queenstown, Rabieridge, Tembisa, Tzaneen, Vanderbijlpark |
| Smart Proxy | $10 / GB of traffic | N/A |
| OxyLab | $15 / GB of traffic | Cape Town |
| IPRoyal | $2–4 / GB of traffic | Cape Town, Johannesburg, Durban, Pretoria |
| NordVPN | $11.99 per month | N/A |
| ExpressVPN | $12.95 per month | N/A |

Table 6: Proxy providers with services available in South Africa. We list each proxy provider's price (per gigabyte of traffic data or monthly subscription) and the proxy geolocations available in South Africa. Some proxy providers do not have a dedicated pool of proxies in specific geolocations/cities in South Africa, which we denoted as "N/A" in the column "Proxy Geolocations in South Africa (City)." Except for *IPRoyal*, all proxy providers were expensive (e.g., >$10 / GB of traffic) or did not have proxies available in our selected geolocations in SA.

requirements. Our task received positive feedback from the AMT community, which indicates acknowledgment of our task design. One respondent reported *"Seems interesting. Hope to get to work on it when it's ready,"* while another stated *"The test was fairly easy and constructed well."* After obtaining positive feedback from the AMT communities and a successful pilot run, we released our AMT annotation task titled "YouTube Video Labeling (COVID-19)." We paid the AMT workers the United States federal minimum wage ($7.25 per hour). When refining our annotation heuristics in §4, we kept track of the time it took the external researchers to annotate the 13 videos, which ranged from 26 seconds to 42 minutes in video duration. We found that it took, on average, 1.5 minutes to annotate a video. However, we erred on the side of overestimation and used 2 minutes as the baseline to annotate a video. We prorated the federal minimum wage using the 2-minute annotation time, thus paying each annotator $0.24 per annotation. Our compensation is much higher than prior works (Papadamou et al. 2022; Albadi, Kurdi, and Mishra 2022; Yi Liaw et al. 2023); for example, Papadamou et al. (2022) paid $0.03 per annotation for labeling YouTube videos for pseudoscientific content.

### YouTube Annotation Task

In our annotation task, we required each AMT worker to assign an annotation label to the video and provide a short rationale behind selecting the label. We required a minimum of 10 characters for our rationale. We released 1,988 videos on AMT in batches of 15 at a time. Three different AMT workers annotated each video. Throughout the annotation process, we sampled videos in each batch to ensure quality work and reasonable rationales from the AMT workers.[20] Using the collected annotations, we calculated the Fleiss' Kappa Score ($\kappa$) to assess the annotators' agreement. We obtained $\kappa = 0.62$, which is considered "substantial agreement." Our $\kappa$ value of 0.62 is a much higher agreement score compared

[20] We found that two AMT workers submitted identical annotations and rationale across the tasks; we rejected their work.

to a close prior work; Papadamou et al. (2022) asked three AMT workers to annotate YouTube videos for pseudoscientific content, including COVID-19 conspiracy theories, using 3 labels, resulting in a Fleiss' $\kappa$ score of 0.14.

To determine the final label of the YouTube video, we chose the annotation label with at least a majority agreement among the workers. We arrived at a final label for 1,899 videos through a majority agreement among the workers. For the remaining 89 videos, all three AMT worker responses differed. The first author, as the expert, annotated these YouTube videos to arrive at the final annotation label. Figure 13 shows the interface of our AMT annotation task.

## G Classifier Training Details

This section describes the selected input features and training procedures of the chosen classification models.

### Feature Descriptions.

We considered the following input features for our classifier.
**Video Title:** The title of the video.
**Video Description:** A small description regarding the content of the video.
**Video Transcript:** Transcript contains the video's actual content, which is either subtitles uploaded by the content creator or auto-generated by YouTube. This feature often contains the main themes discussed by the creator/uploader of the video.
**Video Tags:** As mentioned in §3, video tags are descriptive keywords representing how content creators want their videos to be discovered. The content creators can specify relevant tags to associate with their videos during the upload process.
**Video Comments:** Medina Serrano, Papakyriakopoulos, and Hegelich (2020) found that YouTube comments was a highly predictive feature for detecting COVID-19 misinformation in YouTube videos. Thus, we selected the top 100 comments associated with each video.

| Model | Hyperparameter Search Space |
|-------|------------------------------|
| SVM | C:  [0.1, 1, 10, 100, 1000]<br>$\gamma$:  [1, 0.1, 0.01, 0.001, 0.0001]<br>Kernel: [rbf] |
| XGBoost | max_depth:  [2, 3, 4, 5, 6, 7, 8, 9, 10]<br>n_estimators: [60, 100, 140, 180, 220]<br>learning_rate: [0.5, 0.1, 0.01, 0.05, 0.001] |

Table 7: Hyperparameter search space for each ML model. Grid search with 5-fold cross-validation was used to find optimal hyperparameters with highest average accuracy.

## Training Procedures

Here, we outline the various training procedures and hyperparameters employed to train various classification models.

**Support Vector Machines and XGBoost.** For traditional machine learning algorithms such as SVMs and XGBoost, we applied standard preprocessing procedures on the selected features, such as removing stopwords and whitespace. To create the feature vectors, we tried out 4 different vectorizers: count, TF-IDF, FastText (Bojanowski et al. 2017), and Word2Vec (Mikolov et al. 2013). To find the best set of hyperparameters, we employed a grid search strategy with 5-fold cross-validation on our training dataset, exploring each model's defined hyperparameter search space (see Table 7) and recording the set of hyperparameters that resulted in the highest average cross-validation accuracy. We experimented with various combinations of models, vectorizers, and feature sets.

**DeBerta-v3-base and DeBerta-v3-large.** From training our SVM and XGBoost models, we found that video comments as an input feature resulted in poor performance for our task (see Appendix H). Thus, we concatenated the video's title, description, transcript, and tags as our input text. We intentionally excluded including the video comment as a feature due to their poor performance in predicting COVID-19 misinformation in videos. The concatenated input text was then truncated, retaining the first 1,024 tokens. In addition to the 10% held-out test set, we set aside 10% of the ground-truth dataset to serve as our validation set, leaving 80% of the dataset as our training set. To train the model, we used the Adam optimizer and cross-entropy loss as our loss function. To avoid overfitting, we employed early stopping, in which we trained the model until the validation loss no longer improved over several iterations. We conducted extensive hyperparameter tuning and arrived at each model's final set of hyperparameters. For the DeBerta-v3-base, we employed a training batch size of 8, a learning rate of 1e-5, and a weight decay of 1e-3. For the DeBerta-v3-large, we employed a training batch size of 4, a learning rate of 5e-6, and a weight decay of 1e-4. These models were trained on a single NVIDIA A40 GPU.

**GPT-4 Turbo.** We describe the prompt design considerations and tuning process to evaluate GPT-4 Turbo's performance in detecting COVID-19 misinformation based on YouTube video metadata. Since we employed an OpenAI's

GPT-4 model, we designed our prompts based on OpenAI's recommendations on prompt-engineering (OpenAI 2024) and prior work (Mishra and Chatterjee 2024).

First, prior research showed that models demonstrate improved performance when compelled to reason and justify their decisions (OpenAI 2024; Dammu et al. 2024). Therefore, we mandated GPT-4 to provide direct excerpts from the YouTube video metadata and concise justifications regarding their selected label.

Second, OpenAI (2024) recommended that asking the model to adopt a "persona" in their system can lead to better results from LLMs. Thus, for our task, we prompted GPT-4 to adopt a persona as a "public health expert" with comprehensive knowledge of the COVID-19 pandemic and its misinformation.

Third, we experimented with both zero-shot prompting and few-shot prompting. Zero-shot prompts involve presenting the task to the LLM, only including the label descriptions without accompanying examples or training. Meanwhile, few-shot prompts involve including task-specific examples within the prompt, enabling the pre-trained LLM to condition on the illustrative examples rather than updating its weights. We provided five few-shot examples, each containing video metadata, an assigned label, and reasoning for the label. These few-shot examples were the same as the ones provided to AMT workers on their annotation tasks.

Fourth, we experimented with varying temperature levels to find our task's most ideal parameter setting. Ranging from 0 to 2, the temperature parameter influences how models generate text. When utilizing lower values for temperature, such as 0, the generated text becomes deterministic, selecting more consistent and coherent outputs. Previous works (Mishra and Chatterjee 2024; Park et al. 2024) also experimented with various temperature values, such as 0.2 and 0.7, for their text annotation tasks and found that lower temperature values (e.g. 0.2) resulted in the best performance. Likewise, we selected relatively lower temperature values: 0, 0.2, and 0.7 for our experiments. We used default settings for other parameters. See Table 12 for our zero-shot prompt and Table 13 for our few-shot prompt.

Given the poor performance of video comments as an input feature when training SVM and XGBoost for our task (see Appendix H), we supplied the model with video metadata, such as the video title, description, transcript, and tags, within the prompts. To manage the cost of GPT-4's API based on token usage, we took measures to efficiently control expenses by truncating the video metadata. We truncated the video metadata in our prompt, restricting it to the first 500 tokens per video metadata supplied. This also included the video metadata included in the few-shot examples. To check whether truncating the video metadata impacts GPT-4's performance, we also evaluated how GPT-4 performs when given prompts with full metadata vs. prompts with truncated metadata.

## H  Classifier Performance Results

This section describes the classifier performance results and analysis to understand which input features contribute most to detecting COVID-19 misinformation in YouTube videos.

## Performance Results

In total, we trained and evaluated a total of 62 different trained classifiers. To assess the models consistently, we evaluated the model over the same 10% held-out test set (273 videos) from our ground-truth dataset, providing insights into the model's generalizability and unbiased performance over unseen data. We employed three metrics to capture performance: accuracy, weighted F1-score, and macro F1-score. Table 11 reports the performance of all classifiers. Our best-performing model was DeBerta-v3-large (Index 54 in Table 11), which scored 0.85 across accuracy, weighted F1-score, and macro F1-score. Among traditional machine learning classifiers, we found that SVMs performed the best when using titles, descriptions, transcripts, and tags as the feature set with a TFIDF vectorizer (Index 33 of Table 11), scoring 0.78 across the performance metrics.

Notably, GPT-4 Turbo performed the best when provided with a zero-shot prompt comprising truncated title, description, transcript, and tags as metadata, along with a temperature of 0.2 (Index 57 of Table 11), achieving 0.79 across the performance metrics. Despite not being given any few-shot examples or extensively fine-tuned on the task, zero-shot GPT-4 outperformed the best-performing traditional machine-learning algorithms (i.e., the aforementioned SVM model). In the same setting (e.g. zero-shot, temperature=0), we also evaluated prompts where we provided truncated video metadata to the first 500 tokens (Index 56) vs. complete video metadata in the prompt (Index 55). We found that truncating the video metadata did not negatively impact GPT-4's performance; in fact, our results suggest that prompts with truncated video metadata improved GPT-4's performance (0.78 across performance metrics) compared to that of complete video metadata (0.76 across performance metrics). This may be attributed to the complex and long prompts when using full video metadata. Thus, the improved performance when using truncated metadata aligns with OpenAI (2024)'s recommendation of utilizing shorter, concise details to simplify the prompt.

**Which video features are important in detecting COVID-19 misinformation?** To understand which input feature contributes most to the classification of COVID-19 misinformation videos, we systematically evaluated each feature individually and in combination with other features. Due to the computational and financial costs of employing GPT-4 Turbo and DeBerta-v3, we opted for SVM with FastText as our vectorizer. The training procedure for SVMs remained consistent with the approach outlined earlier.

Performance results for each individual feature and combination of features are presented in Index 0-13 of Table 11. Among the individual features, we observed that video titles yielded the highest performance (0.64 accuracy), suggesting that they are an informative input feature. Similarly, descriptions (0.6 accuracy) and transcripts (0.64 accuracy) demonstrated comparable performance. In contrast to the findings of Medina Serrano, Papakyriakopoulos, and Hegelich (2020), where they identified video comments as highly predictive in detecting COVID-19 misinformation in YouTube videos, our results revealed video comment as the

| Top-N Videos | $p$ | $r$ | $U$ | Mean Rank Diff. |
|---|---|---|---|---|
| 10 | *** | 0.49 | 191 | SA > US |
| 20 | * | 0.30 | 293 | SA > US |
| 30 | - | - | 429 | - |
| 40 | * | 0.31 | 285 | US > SA |
| 50 | ** | 0.36 | 256 | US > SA |

Table 8: Results of the Mann-Whitney U Tests to compare the misinformation bias scores between US and SA at each level considering the top-N videos in the SERPs. For each test, we denote the p-value ($p$), Mann-Whitney effect size ($r$), U-value ($U$), and the mean rank difference. For example, when considering the top-10 videos, SA > US indicates that bots in SA received more misinformative videos in the top-10 search results than bots in the US. Note that: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. Results with no significance was denoted with -.

least effective input feature, achieving an accuracy of only 0.49. Combining title, description, and transcript (Index 9) or title, description, transcript, and tags (Index 12) yielded the best performance of 0.69 accuracy. This further indicates that video comments are not informative for our classification task.

## I  Significance Testing Details

For replicability, we outline the inputs, variables, and numbers involved in the Mann-Whitney U and Kruskal-Wallis H tests.

**Mann-Whitney U Test for Top-N Videos in SERPs.**

As shown in Figure 2 and Table 8, we performed a Mann-Whitney U Test to compare the misinformation bias scores between US and SA geolocations at each level considering the top-N videos in the SERPs. The independent variable was the geolocations in the US vs. SA, while the dependent variable was the misinformation bias scores. The misinformation bias score was calculated by considering the top 10 to top 50 videos in the SERPs. The misinformation bias scores were averaged across all search queries, search filters, and the twin bots associated with each geolocation. Each geolocation resulted in 10 averaged misinformation bias scores, each score value corresponding to a particular day during our experimental run. Given that we had 3 geolocations per country, we combined the averaged scores for each country, resulting in 30 scores for the US and 30 scores for SA. We conducted the Mann-Whitney U Test to compare the misinformation bias scores between the US and SA at each level of top-N videos in the SERPs. We record the p-value, Mann-Whitney effect size $r$, U-value, and mean rank difference in Table 8.
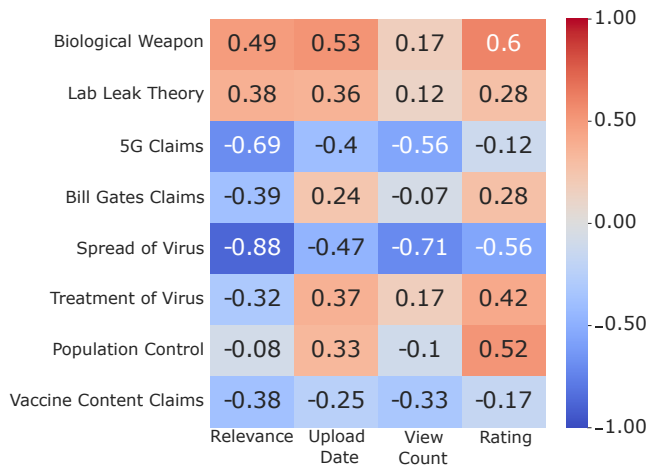
Figure 6: The mean misinformation bias scores for all 8 topics across 4 search filters based on all SERPs collected in the US and SA. Note that the bias scores were calculated considering the top-10 search results.

## Kruskal-Wallis H Test for Within-Country Comparison

In the Results section (§7), we performed Kruskal-Wallis H Test to compare the misinformation bias scores within the geolocations in a given country. The independent variable was the 3 selected geolocations within a country, while the dependent variable was the misinformation bias scores. The misinformation bias scores were calculated by considering the top-10 search results in the SERPs. For each geolocation, we calculated the average misinformation bias scores across the search queries, search filters, and twin bots associated with that geolocation. This resulted in 10 mean misinformation bias scores per geolocation, where each score value corresponded to a particular experimental run. To assess the differences among the geolocations within the country, we conducted the Kruskal-Wallis H Test using the mean misinformation bias scores from the three geolocations within a country, resulting in a total of 30 samples (10 samples each from three geolocations). We also computed the Kruskal-Wallis Test effect size (Tomczak et al. 2014), denoted as $\eta^2$. Following a significant result from the Kruskal-Wallis H Test, we conducted a post-hoc Conover-Iman Test with Bonferroni adjustment using the same inputs to examine the pairwise differences between geolocation. We repeated this procedure for each country.

## Mann-Whitney U Test for Topic-Wise Comparisons

In Figure 3, we performed Mann-Whitney U Tests to compare the misinformation bias scores between the US and SA geolocations for each of the 8 topics. The independent variable was the geolocations in the US and SA, while the dependent variable was the misinformation bias score, considering the top-10 search results in the SERPs. For a particular topic, we averaged the misinformation bias scores across the topic's constituting search queries, search filters, and twin bots associated with each geolocation. Each geolocation resulted in 10 averaged misinformation bias scores for a particular topic; each score value corresponding to a particular day during our experimental run. Given that we had 3 geolocations per country, we combined the averaged scores for each country, resulting in 30 scores for the US and 30 scores for SA for the particular topic. Subsequently, we conducted a Mann-Whitney U Test for each topic to compare the misinformation bias scores between the US and SA. We record the p-value, Mann-Whitney effect size $r$, U-value, and mean rank difference in Figure 3.
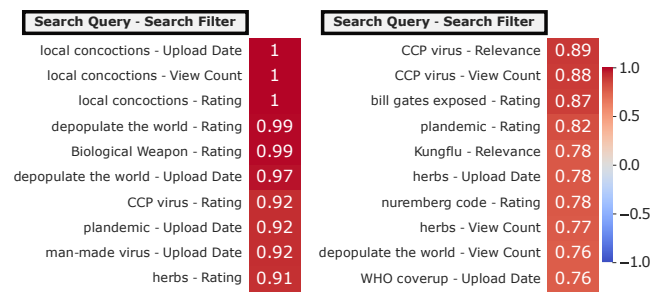


Figure 7: Top 20 search query-filter combinations when sorted by the mean misinformation bias score. These query-filter combinations are the most problematic ones, containing the highest amount of misinformation in the search results.

## Mann-Whitney U Test for Filter-Wise Comparisons

In Figure 5, we performed Mann-Whitney U Tests to compare the misinformation bias scores between the US and SA geolocations for each of the 4 search filters. The independent variable was the geolocations in the US and SA, while the dependent variable was the misinformation bias score, considering the top-10 search results in the SERPs. For a particular search filter, we averaged the misinformation bias scores across the search queries and twin bots associated with each geolocation. Each geolocation resulted in 10 averaged misinformation bias scores for a particular filter, each score value corresponding to a particular day during our experimental run. Given that we had 3 geolocations per country, we combined the averaged scores for each country, resulting in 30 scores for the US and 30 scores for SA for a particular filter. Subsequently, we conducted a Mann-Whitney U Test for each filter to compare the misinformation bias scores between the US and SA. We record the p-value, Mann-Whitney effect size $r$, U-value, and mean rank difference in Figure 5.

## J    Additional Analysis

In this section, we analyze the misinformation bias in search queries and temporal trends of the misinformation bias scores between bots in the US and SA. We also provide Figure 6, which shows the heatmap of the mean misinformation bias scores for all 8 topics across 4 filters considering all SERPs collected in the US and SA.
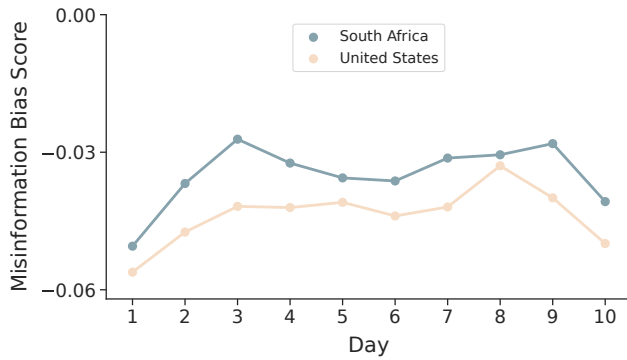
Figure 8: Mean misinformation bias scores of SERPs between bots in the US and SA for each day of the experiment run. Over the 10 days, bots in the SA consistently maintained a higher bias score than bots in the US.

## Misinformation Bias in Search Queries

Figure 7 shows the top 20 search queries and filter combinations with the highest misinformation bias scores. Surprisingly, 10 search query-filter combinations have a very high misinformation bias score ($> 0.9$). The search query "local concoctions" yields a bias score 1 for three filter types, indicating that the search results are completely plagued with misinformative content. Most of the search queries in Figure 7 have a negative connotation, i.e. the search queries themselves have a bias (e.g. the queries "plandemic" and "depopulate the world" indicates an intent to search for misinformative, conspiratorial content). This observation reveals that if you search for misinformation, you will get high amounts of misinformative search results. This indicates how current search engines work; they curate and recommend content without necessarily considering the content's veracity. The most troublesome observation is the presence of high misinformation bias for misleading, xenophobic queries, "CCP virus" (0.88-0.92) and "Kungflu" (0.78). This suggests that YouTube frequently returns highly misinformative search results for such misleading terms, potentially perpetuating their usage.

## Temporal Trends

Figure 8 shows the mean misinformation bias scores of SERPs between bots in the US vs. SA across the 10 experimental days. Each point is the average of all search queries, search filters, and bots associated with the country for a particular experimental day. Throughout the experimental days, we observed that the misinformation bias scores for SA were consistently higher than the scores for the US. This indicates that, overall, bots in SA routinely received more misinformative SERPs than bots in the US during our study.

## K Validations and Robustness Analysis

In this section, we conducted validation checks to reinforce the robustness of our findings.

| Condition | p | r | Mean Rank Diff. |
|---|---|---|---|
| Main Finding | *** | 0.49 | SA > US |
| Case 1 | *** | 0.57 | SA > US |
| Case 2 | ** | 0.37 | SA > US |

Table 9: Summary of the main finding and validation checks under two alternative treatments of the "On the COVID-19 origins in Wuhan, China" label. The results are based on Mann-Whitney U Tests comparing misinformation bias scores between the US and SA for the top-10 videos in the SERPs. Each test reports the p-value ($p$), Mann-Whitney effect size ($r$), and mean rank difference. For example, SA > US indicates that bots in SA received more misinformative SERPs than bots in the US. These validation checks reinforce the robustness of our main findings, showing consistent results regardless of how lab leak theory videos were categorized. Note that: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

## Validation of Results

Although our classifier achieves comparable or exceeds the classification performance of models from prior studies (accuracy of 0.85), it has a 15% error rate, which may affect our findings. In this section, we aimed to confirm the reliability of our findings. To do so, we selected one misinformation topic. We manually annotated the top 10 search results, enabling us to compare the significance test results from fully manually annotated data with those from the classifier-predicted data. We chose the 5G Claims topic because, as shown in Figure 3, the topic exhibited statistical significance with the highest effect size ($r$=0.80), indicating that bots in SA received more misinformative search results than their counterparts in the US. For context, Figure 3 presents the results of the Mann-Whitney U Tests comparing the misinformation bias scores of top-10 search results between the US and SA across each topic.

For the 5G Claims topic, the classifier predicted and labeled 185 videos within the top-10 search results. The first author, as the expert, manually annotated the 185 videos. Subsequently, with complete manual annotations for the 5G Claims topic, we conducted the same analysis, employing the Mann-Whitney U Test to compare the misinformation bias scores for the top-10 search results related to the 5G Claims topic between the US and SA. We obtained a p-value of 6.1e-11 (i.e., p < 0.001) with an effect size $r$=0.84, which is consistent with our findings using classifier-predicted data. This additional analysis provides further evidence of the reliability of our findings despite a 15% error rate on our classifier.

## Robustness Analysis Under Alternative Treatments of Lab Leak Theory Class

Since the Lab Leak Theory are often featured heavily as misinformation, we conducted additional validation checks to evaluate the robustness of our main findings under two alternative treatments of the "On the COVID-19 origins in Wuhan, China" label. Originally, videos discussing COVID-

19 origins were labeled as "On the COVID-19 origins in Wuhan, China"(see §4) and later merged into a neutral "neither" class to reflect the lack of consensus about their status as misinformation.[21] Considering the top-10 videos in the SERPs, our main finding showed that bots in SA were exposed to significantly more misinformative SERPs than bots in the US ($p < 0.001$, effect size $r=0.49$). We replicate this main finding by conducting the same significance test while 1) excluding videos labeled under the "On the COVID-19 origins in Wuhan, China" class in the SERPs (Case 1), and 2) treating the class as misinformation (Case 2). Table 9 summarizes the results of the main finding and validation checks, demonstrating consistent results across all scenarios, regardless of how lab leak theory videos were categorized.

**Case 1: Eliminating the class.** We removed all videos labeled as "On the COVID-19 origins in Wuhan, China" from the dataset and repeated the significance test to compare the misinformation bias scores for the top-10 search results between the US and SA. Our results confirmed that bots in SA still encountered significantly more misinformative SERPs compared to bots in the US ($p < 0.001$, effect size $r=0.57$), consistent with our main findings.

**Case 2: Merging the class with misinformation.** We merged the "On the COVID-19 origins in Wuhan, China" class with the "supporting misinformation" class and conducted the same significance test. Again, SA bots were exposed to significantly more misinformative SERPs compared to bots in the US ($p < 0.01$, effect size $r=0.37$), consistent with our main findings.

These validation checks reinforce the robustness of our findings, demonstrating that bots in SA encountered significantly more misinformative SERPs than bots in the US regardless of how lab leak theory videos were categorized.

---

[21]As of September 2024, Abdelfatah and Arablouei (2024) maintained that the Lab Leak Theory is plausible.

| Scale | Annotation Label | Descriptions | Example Video (Title and ID) |
|---|---|---|---|
| -1 | Opposing COVID-19 Misinformation (Opposing) | The video opposes, debunks, satirizes OR provides counter-vailing contexts to COVID-19 misinformation OR disseminates/endorses public health authorities' policies OR explains the scientific contexts surrounding the public health policy and the COVID-19 pandemic OR provides counter-vailing contexts to the usage of misleading terms, such as "Kungflu" and "China Virus". | 1. Fox Allows Lunatic to Imply Kim Jong-Un Created Cronavirus (*Video ID: 8EGVzWRic1I*) 2. The Monday Times: Coronavirus Car, Ghosts, Tik Tok, 5G Conspiracy — The Daily Show With Trevor Noah (*Video ID: 2totKKH4770*) |
| 0 | Neutral COVID-19 Information (Neutral) | Videos that broadly cover anything related to the COVID-19 pandemic, but do not support nor oppose COVID-19 misinformation. For example, videos that report on the number of COVID-19 cases, discuss remote work hours during a pandemic, etc. | 1. Coronavirus outbreak: Hospitals in China swamped with patients (*Video ID: 9qAFanPeKtE*) 2. Africa Unite Against COVID 19 (*Video ID: 6EMEEn65JPg*) |
| 1 | Supporting COVID-19 Misinformation (Supporting) | The video supports OR provides evidence for COVID-19 misinformation OR suggests alternative treatment, prevention, and cure without scientific evidence OR exaggerates a claim to suggest a misinformative narrative OR promotes the usage of misleading terms without countervailing contexts. | 1. KT The Arch Degree Break's Down The Coronavirus! (MUST SEE) (*Video ID: ulncPbTgPlo*) 2. Origin of the CCP Virus: Fauci Emails Revealed; Ivermectin: A Vaccine Alternative?(*Video ID: edT1HYMjnzA*) |
| 2 | On the COVID-19 origins in Wuhan, China (COVID-19 origins) | Any videos that hypothesize OR cover the origins of COVID-19 in Wuhan, China as the main premise of the video (e.g. natural origins theory, lab leak theory) OR provide evidence for the lab leak theory OR trying to debunk/satirize the lab leak theory. Videos that speculate beyond the origin theories of the virus (e.g., COVID-19 was intentionally engineered as a bioweapon in the Wuhan Lab) are marked as supporting COVID-19 misinformation. | 1. Progressive Lab Leak Theory (*Video ID: 0iNZE6BEoXs*) 2. Why the US is taking a second look at the 'lab leak' theory about COVID-19 (*Video ID: ZRV_uYS2Buc*) |
| 3 | Irrelevant | Any video whose content is not related to the COVID-19 pandemic and vaccines. | 1. Dengue: Laboratory-acquired case reported in North Carolina, according to study (*Video ID: PxFJ3YSgG0Y*) 2. The Nuremberg Code (*Video ID: Eu_3Cbf7f8Y*) |
| 4 | YouTube video in a language other than English (Non-English) | The video's title, description, and/or contents in a language other than English. | 1. Can Coronavirus be Biological Weapon? Video Analysis by Major General S B Asthana (Retd) (*Video ID: WiHmN_DuQhQ*) 2. Shiva \| The Arrogant Kung Fu Fighter \| Episode 54 \| Download Voot Kids App (*Video ID: ISfm5pAWmxs*) |
| 5 | URL not accessible | Youtube Video URL is not accessible at the time of annotation. | - |

Table 10: Table containing our 7-point annotation labels. For each annotation label, we provide the scale value, description, and two example videos.

| ID | Model | Features | Vectorizer | Zero-/Few-Shot | Metadata | Temp. | Acc. | F1-M | F1-W |
|----|-------|----------|------------|----------------|----------|-------|------|------|------|
| | | | *SVM with Individual Features (FastText)* | | | | | | |
| 0 | SVM | Title | FastText | – | – | – | 0.64 | 0.64 | 0.64 |
| 1 | SVM | Description | FastText | – | – | – | 0.60 | 0.60 | 0.60 |
| 2 | SVM | Transcript | FastText | – | – | – | 0.58 | 0.58 | 0.58 |
| 3 | SVM | Tags | FastText | – | – | – | 0.62 | 0.62 | 0.62 |
| 4 | SVM | Comments | FastText | – | – | – | 0.49 | 0.47 | 0.47 |
| | | | *SVM with Feature Combinations (FastText)* | | | | | | |
| 5 | SVM | Title, Description | FastText | – | – | – | 0.68 | 0.67 | 0.68 |
| 6 | SVM | Title, Transcript | FastText | – | – | – | 0.71 | 0.71 | 0.71 |
| 7 | SVM | Title, Tags | FastText | – | – | – | 0.71 | 0.71 | 0.71 |
| 8 | SVM | Title, Comments | FastText | – | – | – | 0.66 | 0.66 | 0.66 |
| 9 | SVM | Title, Description, Transcript | FastText | – | – | – | 0.69 | 0.69 | 0.69 |
| 10 | SVM | Title, Description, Tags | FastText | – | – | – | 0.68 | 0.68 | 0.68 |
| 11 | SVM | Title, Description, Comments | FastText | – | – | – | 0.66 | 0.66 | 0.66 |
| 12 | SVM | Title, Description, Transcript, Tags | FastText | – | – | – | 0.69 | 0.69 | 0.69 |
| 13 | SVM | Title, Description, Transcript, Comments | FastText | – | – | – | 0.67 | 0.67 | 0.67 |
| | | | *SVM with All Features* (Various Vectorizers)* | | | | | | |
| 14 | SVM | All Features* | Count(unigram) | – | – | – | 0.64 | 0.64 | 0.64 |
| 15 | SVM | All Features* | Count(unigram, bigram) | – | – | – | 0.68 | 0.68 | 0.68 |
| 16 | SVM | All Features* | Count(unigram, bigram, trigram) | – | – | – | 0.68 | 0.69 | 0.69 |
| 17 | SVM | All Features* | TFIDF(unigram) | – | – | – | 0.78 | 0.78 | 0.78 |
| 18 | SVM | All Features* | TFIDF(unigram, bigram) | – | – | – | 0.78 | 0.78 | 0.78 |
| 19 | SVM | All Features* | TFIDF(unigram, bigram, trigram) | – | – | – | 0.77 | 0.76 | 0.77 |
| 20 | SVM | All Features* | Word2Vec | – | – | – | 0.76 | 0.76 | 0.76 |
| 21 | SVM | All Features* | FastText | – | – | – | 0.71 | 0.71 | 0.72 |
| | | | *SVM with Feature Combinations (TFIDF)* | | | | | | |
| 22 | SVM | Title, Description | TFIDF(unigram) | – | – | – | 0.70 | 0.70 | 0.70 |
| 23 | SVM | Title, Description | TFIDF(unigram, bigram) | – | – | – | 0.72 | 0.72 | 0.72 |
| 24 | SVM | Title, Transcript | TFIDF(unigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| 25 | SVM | Title, Transcript | TFIDF(unigram, bigram) | – | – | – | 0.75 | 0.75 | 0.75 |
| 26 | SVM | Title, Tags | TFIDF(unigram) | – | – | – | 0.70 | 0.70 | 0.70 |
| 27 | SVM | Title, Tags | TFIDF(unigram, bigram) | – | – | – | 0.71 | 0.71 | 0.71 |
| 28 | SVM | Title, Description, Transcript | TFIDF(unigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| 29 | SVM | Title, Description, Transcript | TFIDF(unigram, bigram) | – | – | – | 0.76 | 0.76 | 0.76 |
| 30 | SVM | Title, Description, Tags | TFIDF(unigram) | – | – | – | 0.72 | 0.72 | 0.72 |
| 31 | SVM | Title, Description, Tags | TFIDF(unigram, bigram) | – | – | – | 0.75 | 0.75 | 0.76 |
| 32 | SVM | Title, Description, Transcript, Tags | TFIDF(unigram) | – | – | – | 0.75 | 0.75 | 0.75 |
| 33 | SVM | Title, Description, Transcript, Tags | TFIDF(unigram, bigram) | – | – | – | 0.78 | 0.78 | 0.78 |
| 34 | SVM | Title, Description, Transcript, Comments | TFIDF(unigram) | – | – | – | 0.75 | 0.75 | 0.75 |
| 35 | SVM | Title, Description, Transcript, Comments | TFIDF(unigram, bigram) | – | – | – | 0.76 | 0.76 | 0.76 |
| | | | *XGBoost with Feature Combination (TFIDF)* | | | | | | |
| 36 | XGB | Title, Description | TFIDF(unigram) | – | – | – | 0.68 | 0.68 | 0.68 |
| 37 | XGB | Title, Description | TFIDF(unigram, bigram) | – | – | – | 0.67 | 0.67 | 0.67 |
| 38 | XGB | Title, Transcript | TFIDF(unigram) | – | – | – | 0.71 | 0.71 | 0.71 |
| 39 | XGB | Title, Transcript | TFIDF(unigram, bigram) | – | – | – | 0.69 | 0.69 | 0.69 |
| 41 | XGB | Title, Tags | TFIDF(unigram) | – | – | – | 0.69 | 0.69 | 0.69 |
| 42 | XGB | Title, Tags | TFIDF(unigram, bigram) | – | – | – | 0.69 | 0.69 | 0.69 |
| 43 | XGB | Title, Description, Transcript | TFIDF(unigram) | – | – | – | 0.70 | 0.69 | 0.69 |
| 44 | XGB | Title, Description, Transcript | TFIDF(unigram, bigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| 45 | XGB | Title, Description, Tags | TFIDF(unigram) | – | – | – | 0.74 | 0.74 | 0.74 |
| 46 | XGB | Title, Description, Tags | TFIDF(unigram, bigram) | – | – | – | 0.75 | 0.75 | 0.76 |
| 47 | XGB | Title, Description, Transcript, Tags | TFIDF(unigram) | – | – | – | 0.71 | 0.71 | 0.71 |
| 48 | XGB | Title, Description, Transcript, Tags | TFIDF(unigram, bigram) | – | – | – | 0.72 | 0.72 | 0.72 |
| 49 | XGB | Title, Description, Transcript, Comments | TFIDF(unigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| 50 | XGB | Title, Description, Transcript, Comments | TFIDF(unigram, bigram) | – | – | – | 0.72 | 0.72 | 0.72 |
| 51 | XGB | All Features* | TFIDF(unigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| 52 | XGB | All Features* | TFIDF(unigram, bigram) | – | – | – | 0.73 | 0.73 | 0.73 |
| | | | *DeBERTa Models* | | | | | | |
| 53 | DeBERTa (base) | Title, Description, Transcript, Tags | – | – | – | – | 0.81 | 0.81 | 0.81 |
| 54 | DeBERTa (large) | Title, Description, Transcript, Tags | – | – | – | – | **0.85** | **0.85** | **0.85** |
| | | | *GPT-4 Turbo Models* | | | | | | |
| 55 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Zero-Shot | Complete | 0 | 0.76 | 0.76 | 0.76 |
| 56 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Zero-Shot | Truncated | 0 | 0.78 | 0.78 | 0.78 |
| 57 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Zero-Shot | Truncated | 0.2 | 0.79 | 0.79 | 0.79 |
| 58 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Zero-Shot | Truncated | 0.7 | 0.77 | 0.77 | 0.77 |
| 59 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Few-Shot | Truncated | 0 | 0.78 | 0.78 | 0.78 |
| 60 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Few-Shot | Truncated | 0.2 | 0.78 | 0.78 | 0.78 |
| 61 | GPT-4 Turbo | Title, Description, Transcript, Tags | – | Few-Shot | Truncated | 0.7 | 0.79 | 0.79 | 0.79 |

Table 11: Performances on our held-out test set (273 videos) across Support Vector Machines (SVM), XGBoost (XGB), DeBERTa-v3, and GPT-4 Turbo (Version 1106) on various features, vectorizers, and parameters. We trained and evaluated 62 models. For GPT-4, we experimented with providing complete vs. truncated video metadata (first 500 tokens of each video metadata) in the prompts; these are indicated by the values "Truncated" and "Complete" in the column "Metadata." The highest performance value is bolded; DeBERTa-v3-large (Index 54) performed the best. Note that Temp: Temperature, Acc.: Accuracy, F1-M: Macro F1-score, F1-W: Weighted F1-score, and All Features*: Title, Description, Transcript, Tags, Comments.

***Zero-shot prompt***

You are a public health expert with comprehensive knowledge of the COVID-19 pandemic and the misinformation surrounding it. You are aware of the various health misinformation in videos on YouTube. You are tasked with carefully analyzing the provided video metadata to discern whether the YouTube video falls into one of three labels: opposing the COVID-19 misinformation with scientific information and public health guidance, supporting the COVID-19 misinformation, or presenting neutral information.

Using the provided label descriptions and examples, please evaluate the YOUTUBE VIDEO METADATA and assign a label. Below, we provided bulleted descriptions and a few examples for each label, defining what videos would fall into the label category:

OPPOSES COVID-19 MISINFORMATION: [OPPOSING DESCRIPTION]

SUPPORTS COVID-19 MISINFORMATION: [SUPPORTING DESCRIPTION]

NEUTRAL INFORMATION:[NEUTRAL DESCRIPTION]

Now, given what you learned from the label descriptions and examples above, please evaluate and assign a label to the YOUTUBE VIDEO METADATA and provide justification on your label with direct excerpts(s) from the YOUTUBE VIDEO METADATA. FORMAT your response as a JSON object in the following structure [(LABEL, EXCERPTS, JUSTIFICATION)].

YOUTUBE VIDEO METADATA starts here *****:
Video Title: [TITLE]
Video Description: [DESCRIPTION]
Video Transcript: [TRANSCRIPT]
Video Tags: [TAGS]

Table 12: Zero-shot prompt to determine whether a YouTube video opposes COVID-19 misinformation, supports COVID-19 misinformation, or is neutral information. For each label, we employed the same descriptions provided to AMT workers on their annotation task. Note that "Neutral Information" here refers to the merged category, consisting of "Neutral COVID-19 Information," "COVID-19 origins," and "Irrelvant" labels. Please see subsection "Consolidating from 5-classes to 3-classes" in §5 for more information.

*Few-shot prompt*

You are a public health expert with comprehensive knowledge of the COVID-19 pandemic and the misinformation surrounding it. You are aware of the various health misinformation in videos on YouTube. You are tasked with carefully analyzing the provided video metadata to discern whether the YouTube video falls into one of three labels: opposing the COVID-19 misinformation with scientific information and public health guidance, supporting the COVID-19 misinformation, or presenting neutral information.

Using the provided label descriptions and examples, please evaluate the YOUTUBE VIDEO METADATA and assign a label. Below, we provided bulleted descriptions and a few examples for each label, defining what videos would fall into the label category:

OPPOSES COVID-19 MISINFORMATION: [OPPOSING DESCRIPTION]

SUPPORTS COVID-19 MISINFORMATION: [SUPPORTING DESCRIPTION]

NEUTRAL INFORMATION:[NEUTRAL DESCRIPTION]

We provide five examples of the task, each featuring video metadata, label, and reasoning.
EXAMPLE 1 starts here ****:
VIDEO_TITLE: [EXAMPLE1_VIDEO_TITLE]
VIDEO_DESCRIPTION: [EXAMPLE1_VIDEO_DESCRIPTION]
VIDEO_TRANSCRIPT: [EXAMPLE1_TRANSCRIPT]
VIDEO_TAGS: [EXAMPLE1_TAGS]
LABEL: [EXAMPLE1_LABEL]
REASONING: [EXAMPLE1_REASONING]
...
...
...
EXAMPLE 5 starts here ****:
VIDEO_TITLE: [EXAMPLE5_VIDEO_TITLE]
VIDEO_DESCRIPTION: [EXAMPLE5_VIDEO_DESCRIPTION]
VIDEO_TRANSCRIPT: [EXAMPLE5_TRANSCRIPT]
VIDEO_TAGS: [EXAMPLE5_TAGS]
LABEL: [EXAMPLE5_LABEL]
REASONING: [EXAMPLE5_REASONING]

Now, given what you learned from the label descriptions and examples of the task, please evaluate and assign a label to the YOUTUBE VIDEO METADATA and provide justification on your label with direct quote(s) from the YOUTUBE VIDEO METADATA. FORMAT your response as a JSON object in the following structure [(LABEL, QUOTE, JUSTIFICATION)].

YOUTUBE VIDEO METADATA starts here *****:
Video Title: [TITLE]
Video Description: [DESCRIPTION]
Video Transcript: [TRANSCRIPT]
Video Tags: [TAGS]

Table 13: Few-shot prompt to determine whether a YouTube video opposes COVID-19 misinformation, supports COVID-19 misinformation, or is neutral information. For each label, we employed the same descriptions and examples provided to AMT workers on their annotation task. Note that "Neutral Information" here refers to the merged category, consisting of "Neutral COVID-19 Information," "COVID-19 origins," and "Irrelvant" labels. Please see subsection "Consolidating from 5-classes to 3-classes" in §5 for more information.

**For each label, we provided bulleted descriptions and a few examples -- each of the bulleted descriptions define what videos would fall into the label category. As you go through the video metadata in the aforementioned priority order, please use the label descriptions and the examples in order to annotate the YouTube Video URL. Please read the descriptions and the examples below carefully**

- **Opposes misinformation and problematic content on COVID-19**
  - Videos that oppose, debunk, or offer countervailing contexts to the misinformation and problematic content on the COVID-19 pandemic and vaccines. For example, a video arguing that COVID-19 is not bioengineered, that Bill Gates did not plant microchips in the vaccine, that vaccines do not violate the Nuremberg Code.
  - Videos that explicitly provide, endorse, disseminate, or aligns with public health authorities/organizations' (e.g. the World Health Organization, the Center for Disease Control, etc) COVID-19 policies and recommendations, including but not limited to explaining how to properly wash your hands, properly wear masks, social distance, and getting vaccinated.
  - Videos that explicitly explain the correct scientific context behind the COVID-19 virus (i.e. symptoms, spread of the virus, how the virus interacts within the body, etc) and public health policies (i.e. vaccines, masking, social distancing, etc). For example, videos that help the audience understand how vaccines work and how vaccines were developed, videos that scientifically explain the COVID-19 symptoms, etc. Note that some videos may claim to use "science" to spread misinformation.
  - Satire videos that make fun of the misinformation or problematic content around the COVID-19 pandemic and vaccines.

**Example 1:** Fox Allows Lunatic to Imply Kim Jong-Un Created Cronavirus
**Label:** Opposes misinformation and problematic content on COVID-19
**Reason for assigning label:** Video description and transcript debunks the misinformation that North Korea made coronavirus as a bioweapon, and argues that people need to listen to pandemic specialists, doctors, and scientists instead of "lunatic Christian preachers"



Figure 9: For each annotation label, we provided bulleted descriptions and a few examples, including the reasoning behind assigning the label. These annotation label descriptions were provided in both the Qualification Test and the actual annotation task.

## Qualification Test Overview

You will be graded on 5 questions in total including the eligibility question. You qualify if you fulfill our eligibility criteria and answer all four labeling questions mentioned below correctly. Please read the instructions carefully before attemptimg the questions.

Figure 10: Figure illustrating our Qualification Test instructions. Our test included 5 questions, of which four were labeling questions and one was an eligibility question that required the addition of the authors' university. A full score of 100 was required to qualify for the test.

**Task Instruction**

In this task, you will be presented with a YouTube Video URL that you have to annotate based on the given possible labels defined below. To annotate the Youtube Video, please go through the video metadata in the following priority order: 1) video title and description, 2) content of the video/video transcript (use the overall premise of the video). If you come across a video that discusses events or individuals that require more contextual information, we encourage you to perform an online search to better understand the context and determine the annotation. Please carefully go through the figures depicting YouTube video metadata, list of possible labels for each video, and the descriptions and examples of each label below.

A figure depicting YouTube's video titles and descriptions:



Figure 11: Task instruction in the Qualification Test explaining how to annotate YouTube videos for our task. The same instructions were provided in the actual annotation task.

## Question 1

Please open the YouTube video URL below and assign it a label that best suits the video.

YouTube URL 1

○ Opposes misinformation and problematic content on COVID-19

○ Supports misinformation and problematic content on COVID-19

○ Neutral COVID-19 information

○ On the COVID-19 origins in Wuhan, China

○ Irrelevant

○ URL not accessible

○ Youtube video in a language other than English

Figure 12: Example of the Qualification Test question.

Figure 13: Interface of our YouTube video annotation task on Amazon Mechanical Turk.