COVID misinfo is the biggest challenge for Twitter's Birdwatch program, data shows

As of October 6th, Twitter's Birdwatch community moderation program has been expanded to all US users.

It's a big step for Birdwatch, which was officially launched in beta in January 2021, and marks a step up for the platform's efforts to reduce the spread of misinformation on the platform. But as the scheme expands, data reviewed by *The Verge* suggests that the most common topics being fact-checked are already covered by Twitter's misinformation policies, raising new questions as to the overall impact of the program.

At its core, the promise of Birdwatch is to "decentralize" the process of fact-checking misinformation, putting power into the hands of the community of users rather than a tech company. But fact-checking encompasses a huge range of topics, from trivial and easily debunked rumors to complex claims that may hinge on fundamental uncertainties in the scientific process.

"It can speak to the internet's random curiosities that pop up"

In public statements, Twitter executives involved in the program have focused on the easier decisions. In a call with reporters last month, Keith Coleman, vice president of product at Twitter, suggested that the strength of Birdwatch was in addressing statements that were not covered by Twitter's misinformation policies or weren't serious enough to be assigned in-house fact-checking resources. "It can speak to the internet's random curiosities that pop up," *Gizmodo* quotes Coleman as saying. "Like, is there a giant void in space? Or, is this bat actually the size of a human?"

a:hover]:shadow-highlight-franklin [&>a]:shadow-underline-black dark:[&>a:hover]:shadow-highlight-franklin dark:[&>a]:shadow-underline-white md:text-26″>METHODOLOGY
We downloaded Birdwatch data up to September 20th. This dataset contained 37,741 notes in total, of which 32,731 were unique.

We used Python's natural language tool kit library to parse the notes and extract the most common significant words appearing in them.

To do this, we discarded conjunction words like "and," "but," "there," "which," and "about" and excluded words that were frequently used in the process of constructing a fact check, such as "tweet," "source," "claims," "evidence," and "article." We also ignored words inside URLs — which Twitter includes as part of the note text — and reduced plurals to their singular form (so "cars" would be counted as "car").

The processed data gives us a good overview of topics that are commonly addressed or have context added to them using the Birdwatch system.

➡ To explore the full data yourself, you can browse our interactive database of Birdwatch notes.

But cases from the beta phase of the program show that many Birdwatch users are attempting to tackle more serious misinformation issues on the platform and overlapping significantly with Twitter's existing policies. Birdwatch data released by Twitter shows that COVID-related topics are by far the most common subject addressed in Birdwatch notes. What's more, many of the accounts that posted the tweets that were annotated have since been suspended, suggesting that Twitter's internal review process is catching content violations and taking action.

As part of its broader open-source efforts, Twitter maintains a regularly updated dataset of all Birdwatch notes is freely

available to download from the project blog. *The Verge* analyzed this data, looking through a dataset that spanned from January 22nd, 2021, to September 20th, 2022. Using computational tools to collate and summarize the data, we can gain an insight into the major topics of Birdwatch notes that would be hard to gain from manual review.

Data shows that Birdwatch users have spent a lot of time reviewing tweets related to COVID, vaccination, and the government's response to the pandemic. The word frequency list shows us that "COVID" is the most common subject term, with the related term "vaccine" ranking at number three on the list.

Of these notes, the type of claims being commonly fact-checked evolves over time as public understanding of the pandemic changes. Tweets from 2021 address false narratives claiming that Dr. Anthony Fauci somehow had a personal role in creating the novel coronavirus or shedding doubt on the safety and effectiveness of vaccines as they became available.

Other Birdwatch notes from this time address unproven or dangerous treatments for COVID, like ivermectin and hydroxychloroquine.



While some of the more outlandish COVID myths are easy to fact-check — like the idea that the virus was a hoax, is mostly harmless, or gets spread by 5G towers — other claims about transmission, severity, and mortality can be harder to definitively correct.

For example, as vaccines rolled out in January 2021, one Birdwatch user tried to add context to an argument over one vaccine brand's effectiveness at preventing hospitalization vs. preventing any infection whatsoever. New Jersey Governor Phil Murphy tweeted that trial data for the Johnson & Johnson vaccine showed "COMPLETE protection against hospitalization and death" and provoked an angry response from a statistician who linked to trial data showing only "66% efficacy" from the vaccine.

"The [tweet] author is confusing the reported efficacy of preventing hospitalization and death, with the reported overall efficacy of preventing infection," a Birdwatch note added helpfully, referencing *Bloomberg* coverage that clearly distinguished between the metrics.

More questionably, another Birdwatch user attempted to fact-check a claim widely reported by mainstream news outlets, using a blog post on a prepping website as a citation. Where news outlets followed the CDC's lead in reporting that the

omicron variant made up 73 percent of new infections as of December 2021, a blog post on *ThePrepared.com* argued that the claim may have stemmed from an error in the CDC's statistical modeling. The blog post was tightly argued, but without confirmation from a more reliable and vetted source, it's hard to know whether the annotation helped the situation or simply muddied the waters.

Birdwatch users rated tweets like these as being some of the most problematic to deal with. (By filling in a survey when creating a note, users can rate tweets on four binary values qualifying how misleading, believable, harmful, and difficult to fact-check the claims are). It's clear that accurate, accessible communication of scientific findings is a difficult task, but public health outcomes depend on surfacing accurate health advice and preventing bad advice from proliferating. Experts agree that platforms need strong, clear and coordinated standards for addressing misinformation about the pandemic, and it seems unlikely that community-driven moderation meets this bar.

Though COVID is a main topic of Birdwatch notes, it's far from the only one.

In the word frequency list, "earthquake" and "prediction" rank highly due to a large number of identically worded notes that were attached to tweets from accounts that falsely claim to be able to predict earthquakes around the world.

There's no evidence that earthquakes can reliably be predicted, but inaccurate earthquake predictions keep going viral online. With 48K followers at time of writing, the @Quakeprediction Twitter account is one of the worst offenders, posting a steady stream of predictions of elevated earthquake risk in California. One Birdwatch user seems to have taken it on themselves to attach a warning note to more than 1,300 tweets from this and other earthquake prediction accounts, each time linking to a debunk from the US Geological

Survey explaining that scientists have never predicted an earthquake.

It’s unclear why the user focused on earthquakes, but the end result is a human reviewer ironically behaving more like automated fact-checking software: looking for a pattern in tweets and responding with an identical action each time.