Wisdom of Two Crowds: Current Practices of Misinformation Moderation on Reddit and How to Improve this Process-A Case Study of COVID-19

CSCW 2023

Lia Bozarth, Jane Im, Christopher Quarles, Ceren Budak

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

CENTER FOR AN INFORMED PUBLIC
UNIVERSITY of WASHINGTON

Dr. Lia Bozarth

Jane Im

Dr. Chris Quarles

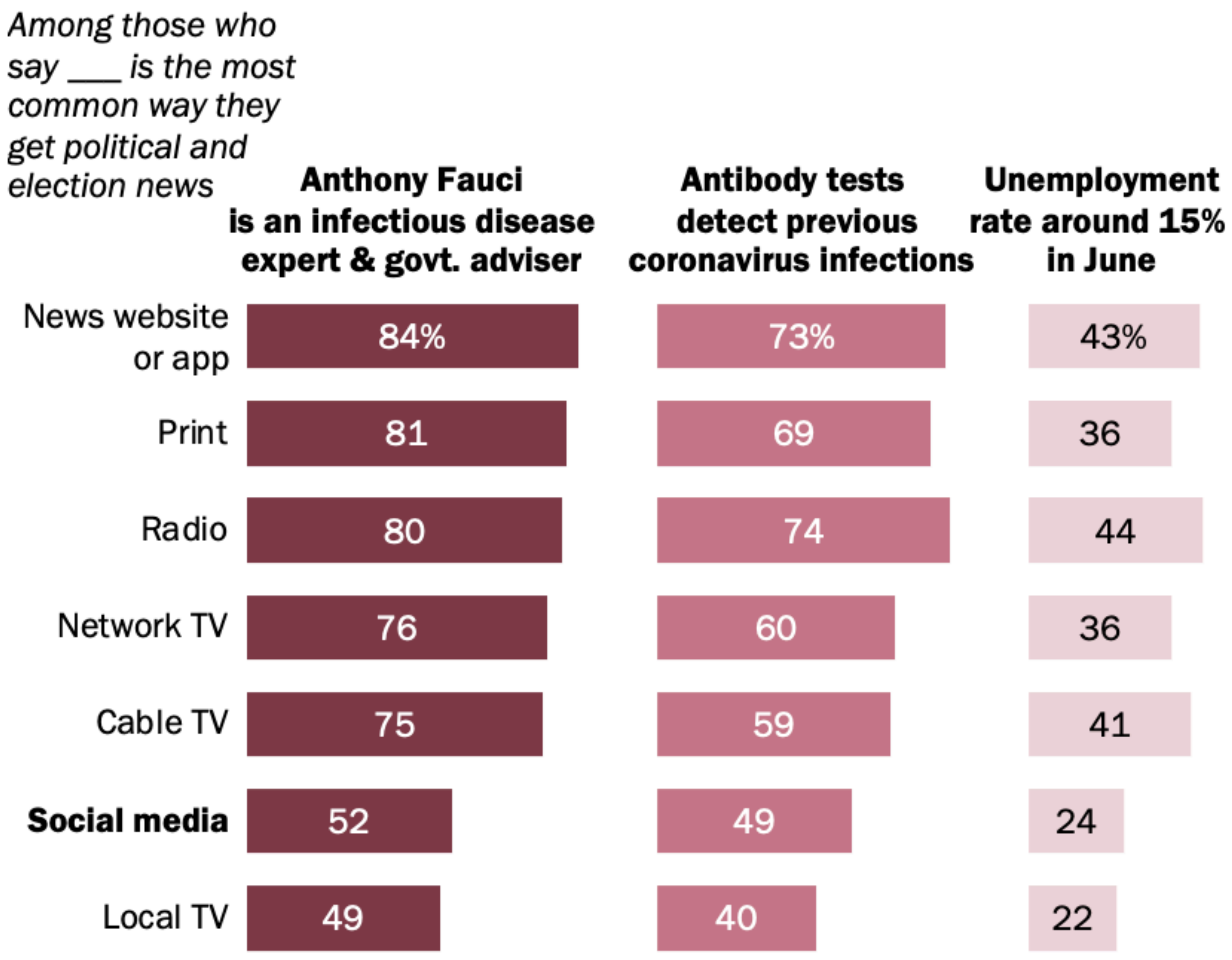
Dr. Ceren Budak

Dr. Lia Bozarth

Data Engineer, Center for an Informed Public
liafan@uw.edu

Background & Motivation

"Those who rely on social media for news are less likely to get the facts right about the coronavirus and politics and more likely to hear some unproven claims."
- Pew Research Center -



Source: Survey of U.S. adults conducted June 4-10, 2020.
"How Americans Navigated the News in 2020: A Tumultuous Year in Review"

PEW RESEARCH CENTER

Background & Motivation

Prior research has explored whether crowd signals available on social media can be aggregated to detect misinformation.

ordinary users on social media

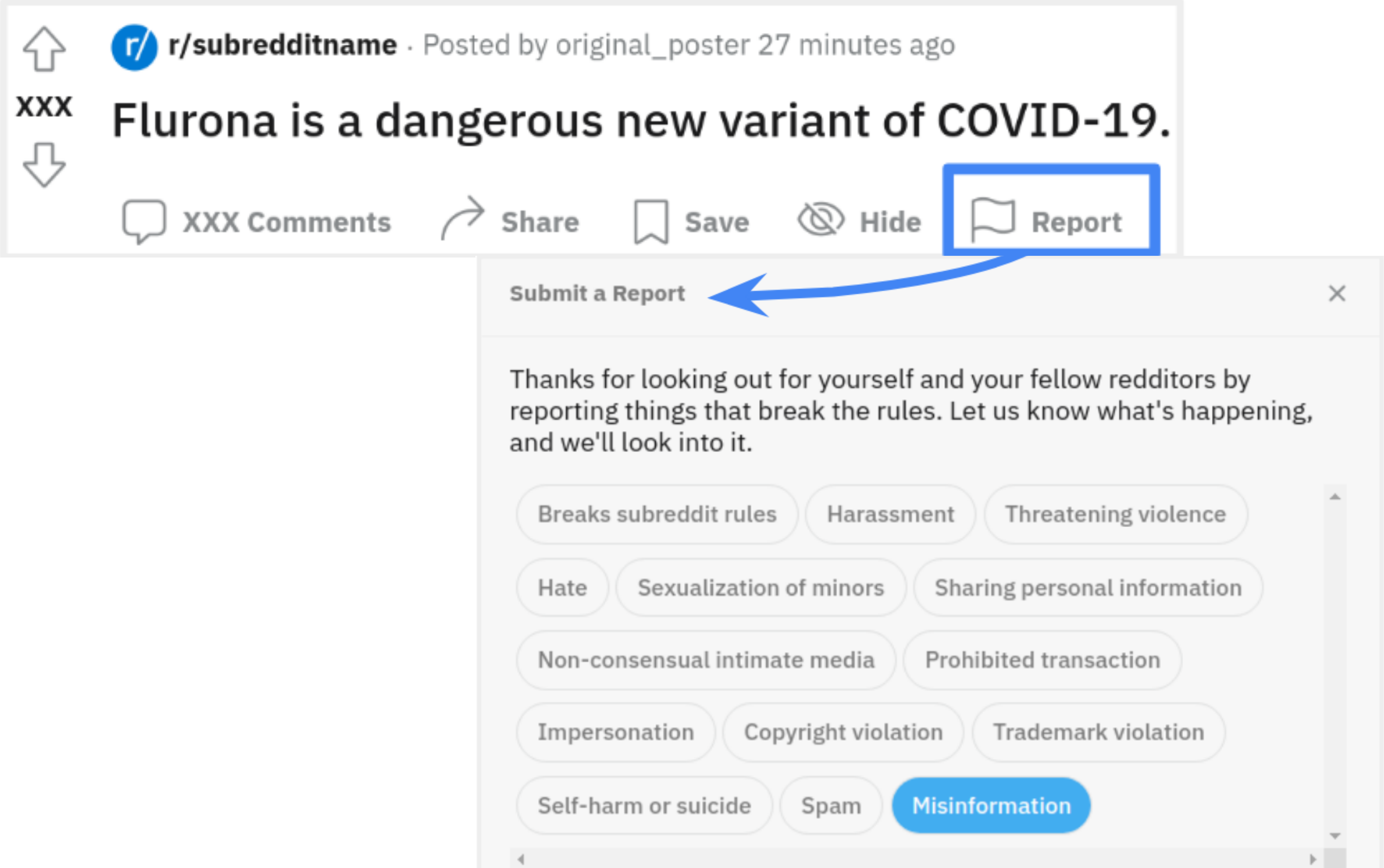
Background & Motivation

Platform users could give two kinds of signals regarding misinformation.

crowdsourced flagging	crowdsourced fact-checking

↓

users using platforms’ **report systems** to flag content to platform admins/moderators



r/subredditname · Posted by original_poster 27 minutes ago

XXX

Flurona is a dangerous new variant of COVID-19.

Submit a Report

Thanks for looking out for yourself and your fellow redditors by reporting things that break the rules. Let us know what's happening, and we'll look into it.

Background & Motivation

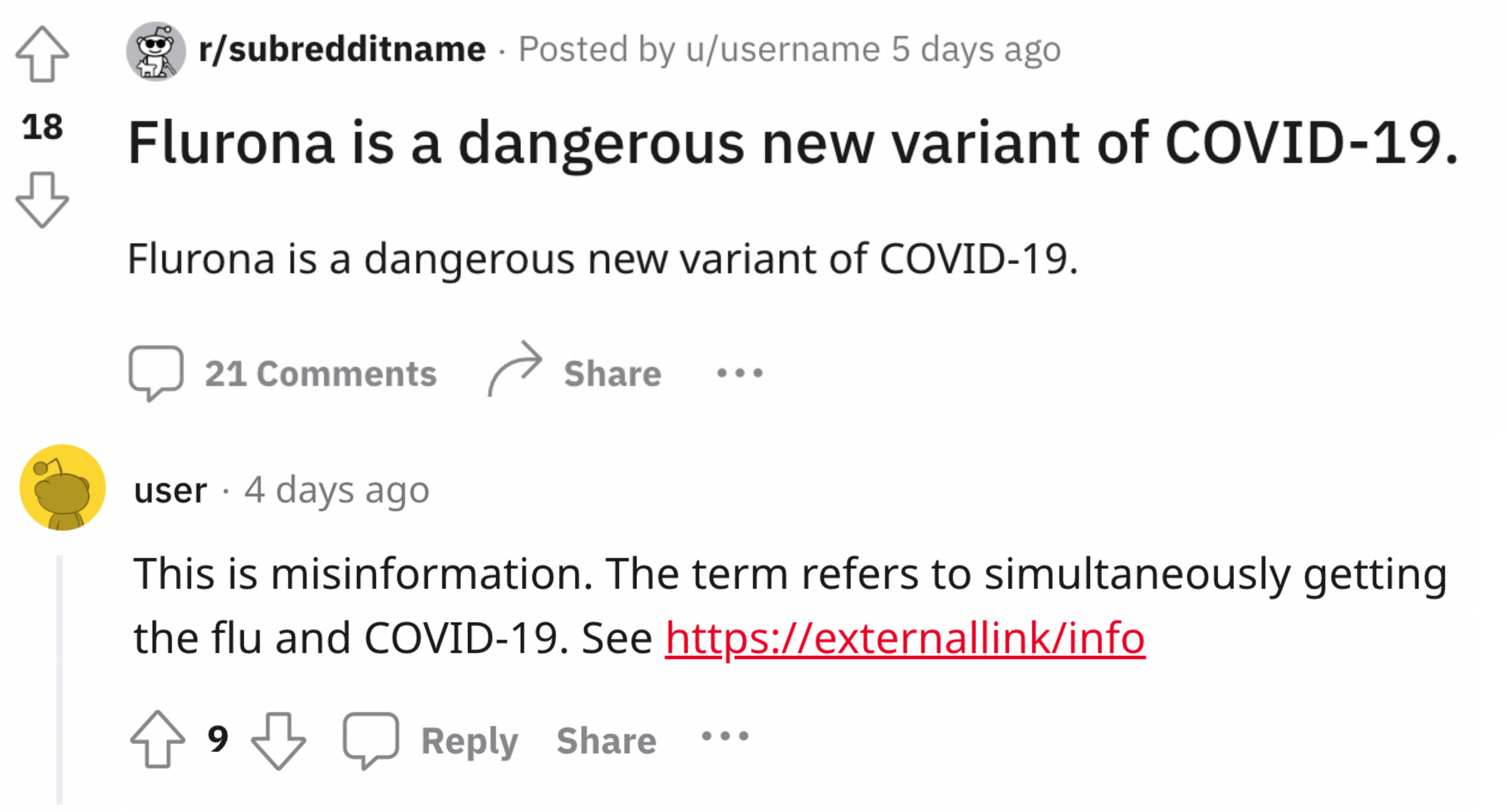
Platform users could give two kinds of signals regarding misinformation.

crowdsourced flagging

crowdsourced fact-checking

↓

users fact-checking other users’ posts or comments **via commenting**



r/subredditname · Posted by u/username 5 days ago

Flurona is a dangerous new variant of COVID-19.

Flurona is a dangerous new variant of COVID-19.

user · 4 days ago

This is misinformation. The term refers to simultaneously getting the flu and COVID-19. See https://externallink/info

Background & Motivation

crowdsourced flagging	crowdsourced fact-checking
alerts platform admins/ moderators	does not necessarily alert platform admins/moderators (have to discover the thread)

Background & Motivation

While research has argued the promises of both signals, most prior work are **from the perspective of platforms**.

Lack of knowledge on how **moderators** in **self-governing online communities** leverage crowdsourced flagging and fact-checking.

Background & Motivation

While research has argued the promises of both signals, most prior work are **from the perspective of platforms**.

Lack of knowledge on how **moderators** in **self-governing online communities** leverage crowdsourced flagging and fact-checking.

Research Questions

How do moderators in self-governed online communities leverage crowd wisdom?

How do moderators use **crowdsourced flagging** (report feature) in their moderation?

Do moderators leverage **crowdsourced fact-checking** (users' comments) at all?

Research Questions

Are there any shared practices in moderators leveraging crowd wisdom, when there exist variances across communities?

And, if there are shared practices, how could they be improved?

Example of Self-Governed Online Communities: Reddit

Heavily relies on communities called “subreddits” to self-govern.

https://www.reddit.com/r/[subreddit_name]

Each subreddit has its own rules and moderators.

Join
Create Post
Hot
New
Top
Submit
About Community
Create Post

Keanu Reeves Being Awesome

r/KeanuBeingAwesome

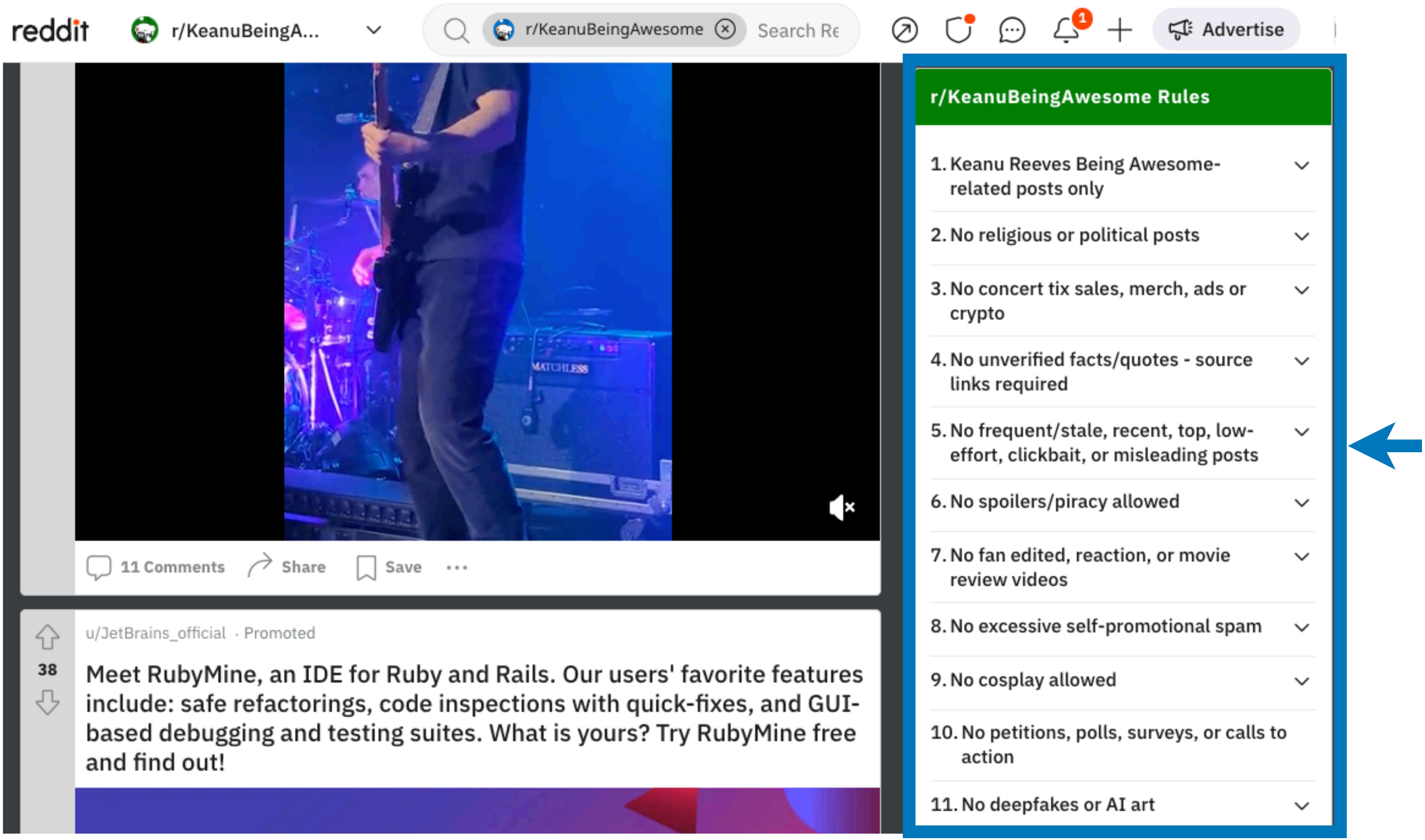
What mature themes are posted about or discussed in r/KeanuBeingAwesome ?

- Alcohol & tobacco
- Drug use
- Gambling
- Guns & weapons
- Military conflict & terrorism
- Nudity
- Profanity
- Sex & eroticism
- Shock & outrage
- Violence & gore
- None of the above
- Other

A subreddit for posts of Keanu Reeves Being Awesome. The Immortal was born on September 2, 1964, in Beirut, Lebanon. Of part-Hawaiian heritage on his father's side, /ki'ɑnu/ translates from Hawaiian to English as "cool breeze over the mountains."

Created Jan 11, 2018

362k	● 31	Top 1%
Members	Online	Ranked by Size

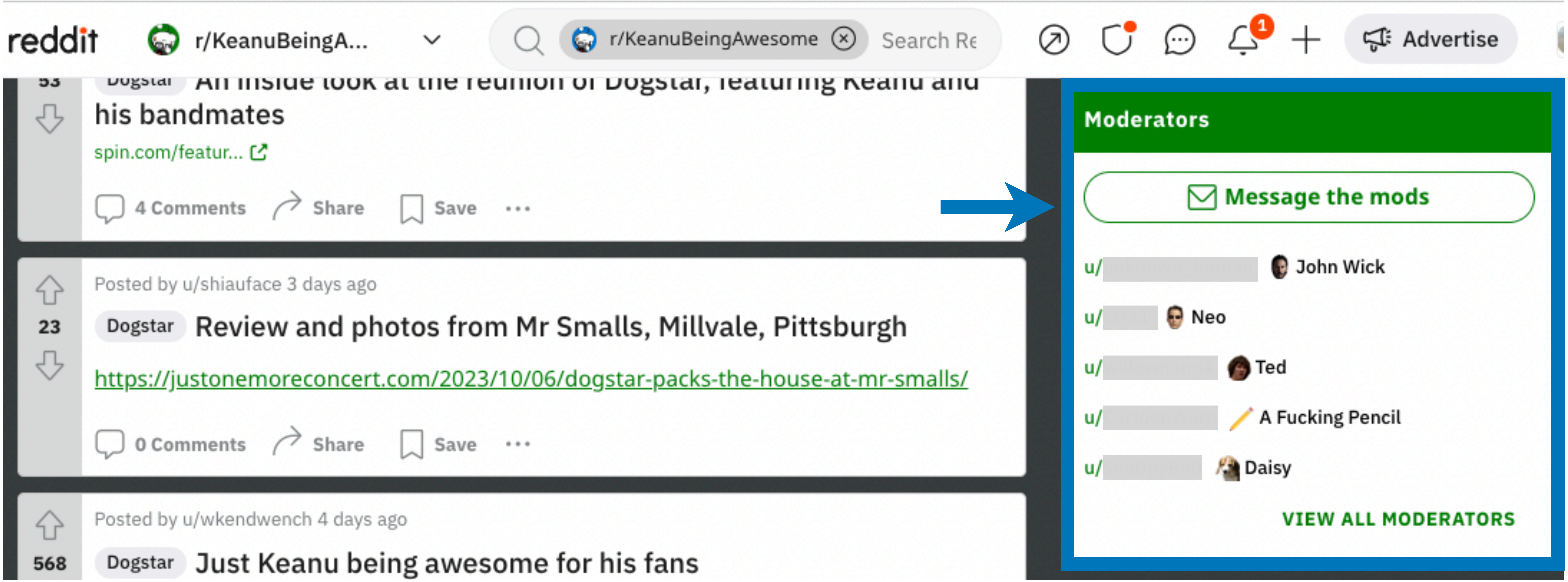


u/JetBrains_official · Promoted

Meet RubyMine, an IDE for Ruby and Rails. Our users' favorite features include: safe refactorings, code inspections with quick-fixes, and GUI-based debugging and testing suites. What is yours? Try RubyMine free and find out!

r/KeanuBeingAwesome Rules

1. Keanu Reeves Being Awesome-related posts only
2. No religious or political posts
3. No concert tix sales, merch, ads or crypto
4. No unverified facts/quotes - source links required
5. No frequent/stale, recent, top, low-effort, clickbait, or misleading posts
6. No spoilers/piracy allowed
7. No fan edited, reaction, or movie review videos
8. No excessive self-promotional spam
9. No cosplay allowed
10. No petitions, polls, surveys, or calls to action
11. No deepfakes or AI art



53 Dogstar An inside look at the reunion of Dogstar, featuring Keanu and his bandmates

spin.com/featur...

4 Comments Share Save ...

Posted by u/shiauface 3 days ago

23 Dogstar Review and photos from Mr Smalls, Millvale, Pittsburgh

https://justonemoreconcert.com/2023/10/06/dogstar-packs-the-house-at-mr-smalls/

0 Comments Share Save ...

Posted by u/wkendwench 4 days ago

568 Dogstar Just Keanu being awesome for his fans

Moderators

Message the mods

u/[redacted] John Wick

u/[redacted] Neo

u/[redacted] Ted

u/[redacted] A Fucking Pencil

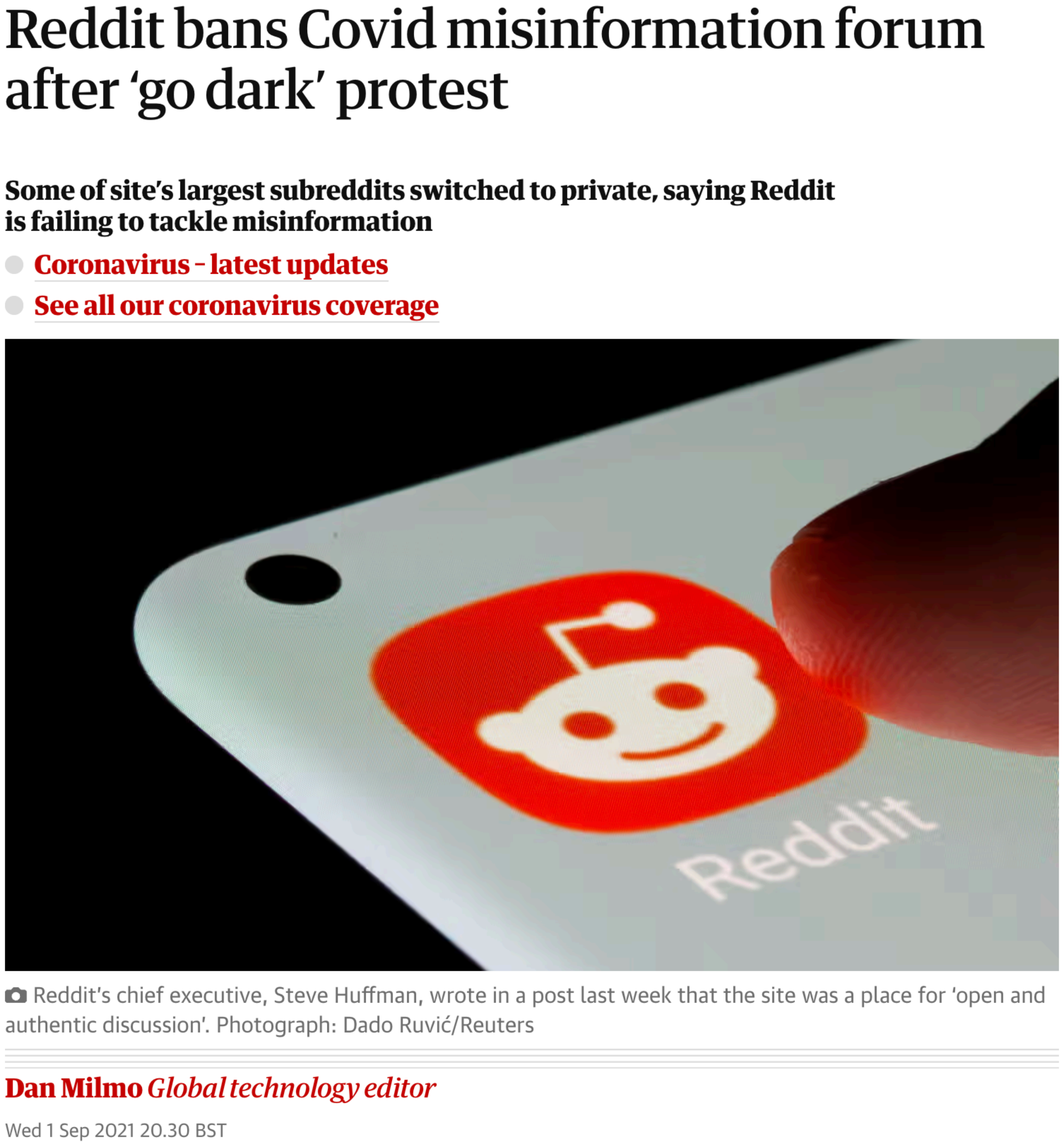
u/[redacted] Daisy

VIEW ALL MODERATORS

Context - COVID-19 misinformation on Reddit

Why?

Reddit moderators had publicly expressed that they had significant difficulty moderating COVID-19 misinformation.



Reddit bans Covid misinformation forum after 'go dark' protest

Some of site's largest subreddits switched to private, saying Reddit is failing to tackle misinformation

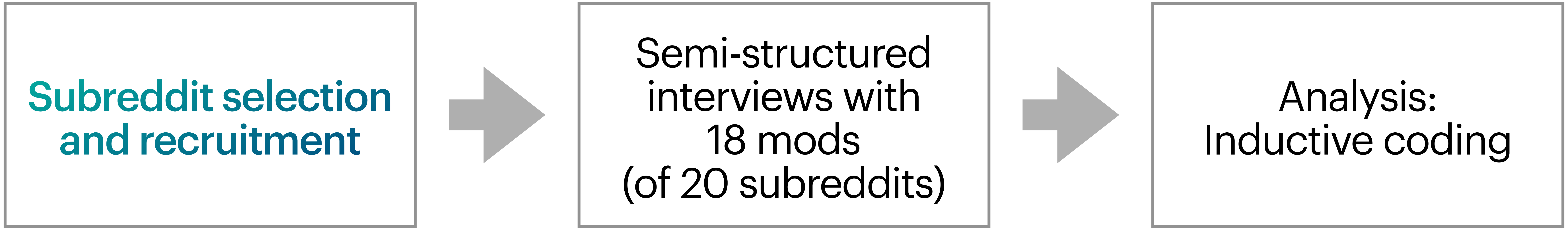
- Coronavirus - latest updates
- See all our coronavirus coverage

Reddit's chief executive, Steve Huffman, wrote in a post last week that the site was a place for 'open and authentic discussion'. Photograph: Dado Ruvić/Reuters

Dan Milmo *Global technology editor*

Wed 1 Sep 2021 20.30 BST

Study Process



Subreddit selection and recruitment

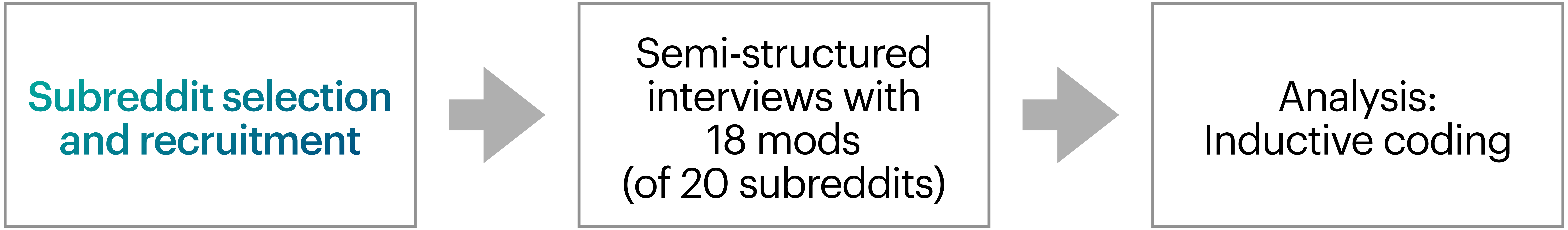
Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

Focus on subreddits of reasonable size where

1) COVID content has been actively moderated
2) COVID misinformation could have been an issue

Study Process



Subreddit selection and recruitment

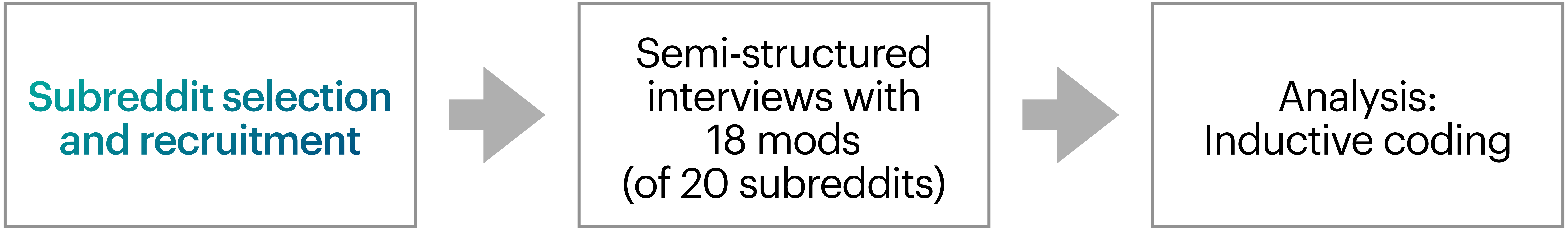
Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

selected a pool of 424 subreddits that

1) had ≥ 1000 subscribers (98th-percentile)
2) frequently hosted COVID-related posts
3) contained COVID-related posts that moderators took actions on

Study Process



Subreddit selection and recruitment

Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

selected a pool of 424 subreddits that

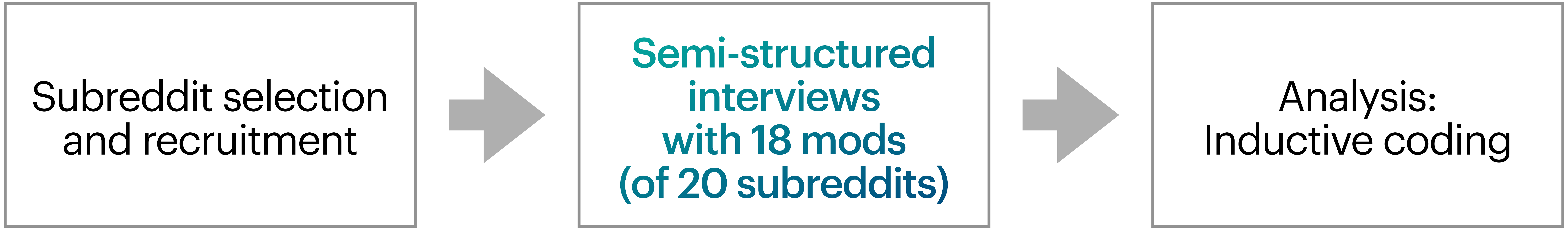
1) had ≥ 1000 subscribers (98th-percentile)
2) frequently hosted COVID-related posts
3) contained COVID-related posts that moderators took actions on

used COVID-related keywords from prior work

Chen, Emily, Kristina Lerman, and Emilio Ferrara. "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set." *JMIR public health and surveillance* 6, no. 2 (2020): e19273.

Kaur, Simranpreet, Pallavi Kaul, and Pooya Moradian Zadeh. "Monitoring the dynamics of emotions during COVID-19 using Twitter data." *Procedia Computer Science* 177 (2020): 423-430.

Study Process



Subreddit selection and recruitment

Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

whether subreddit is COVID-related or not:

COVID-specific (10/20)

non-COVID specific (10/20)

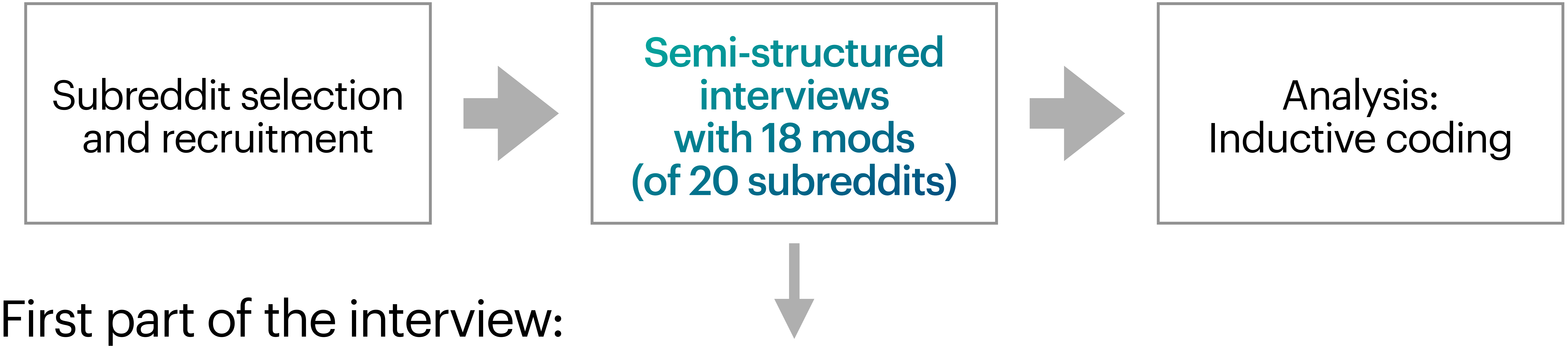
subreddit size:

small (4/20), medium (8/20), large (8/20)

subreddit ideological leaning:

liberal (10/20), not ideologically aligned (9/20), conservative (1/20)

Study Process



Subreddit selection and recruitment

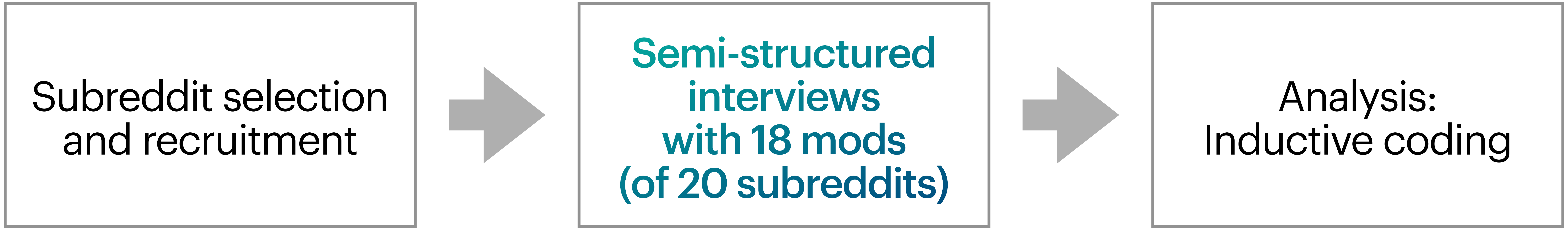
Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

First part of the interview:

1) how moderators conceptualize misinformation
2) moderation workflow and practices
3) role of crowd wisdom in the workflow

Study Process



Subreddit selection and recruitment

Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

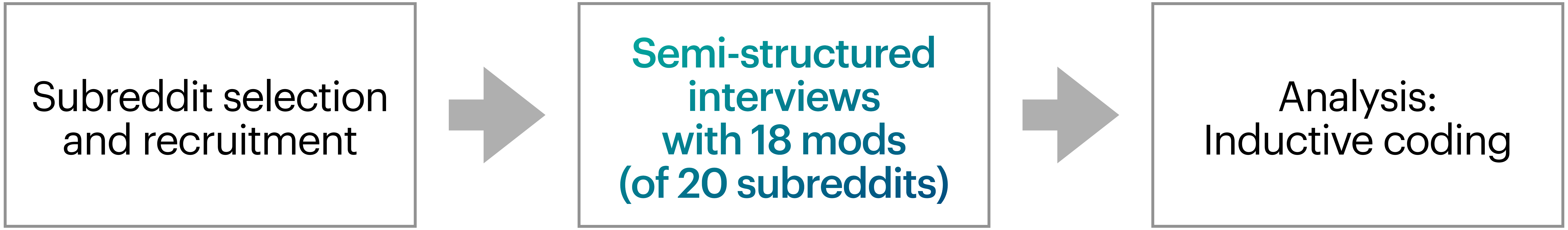
Second part of the interview:

Moderators' thoughts on novel moderation queue designs that **leverage crowdsourced fact-checking**.

crowdsourced fact-checking

similar posts + crowdsourced fact-checking

Study Process



Subreddit selection and recruitment

Semi-structured interviews with 18 mods (of 20 subreddits)

Analysis: Inductive coding

Second part of the interview:

Moderators’ thoughts on novel moderation queue designs that **leverage crowdsourced fact-checking**.

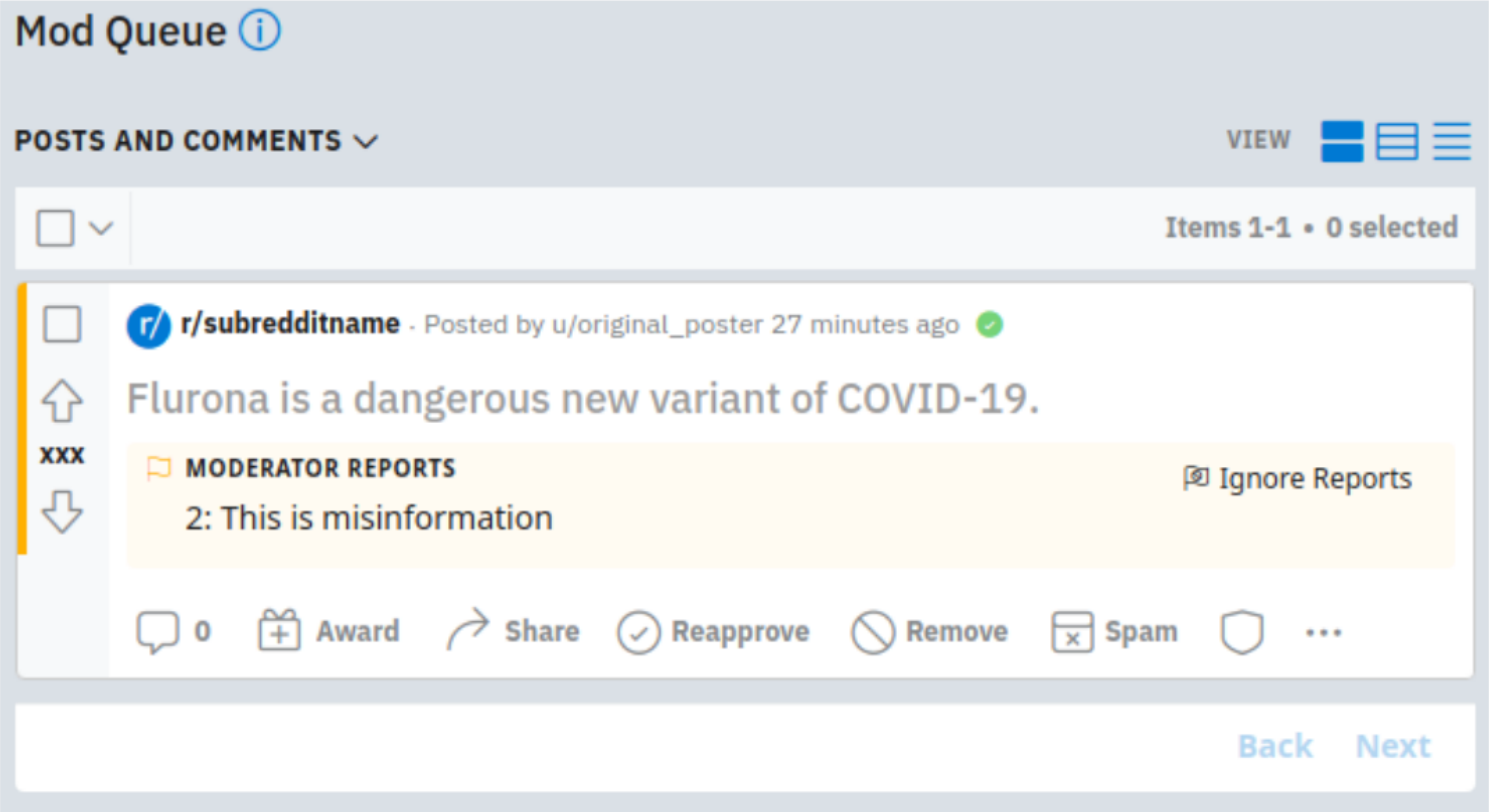
crowdsourced fact-checking

similar posts + crowdsourced fact-checking

versus

labels from professional fact-checkers

default mod queue: crowdsourced flagging (users' reports) (which we forked off for our alternative designs)



Mod Queue

POSTS AND COMMENTS

VIEW

□ Items 1-1 • 0 selected

□ r/subredditname · Posted by u/original_poster 27 minutes ago

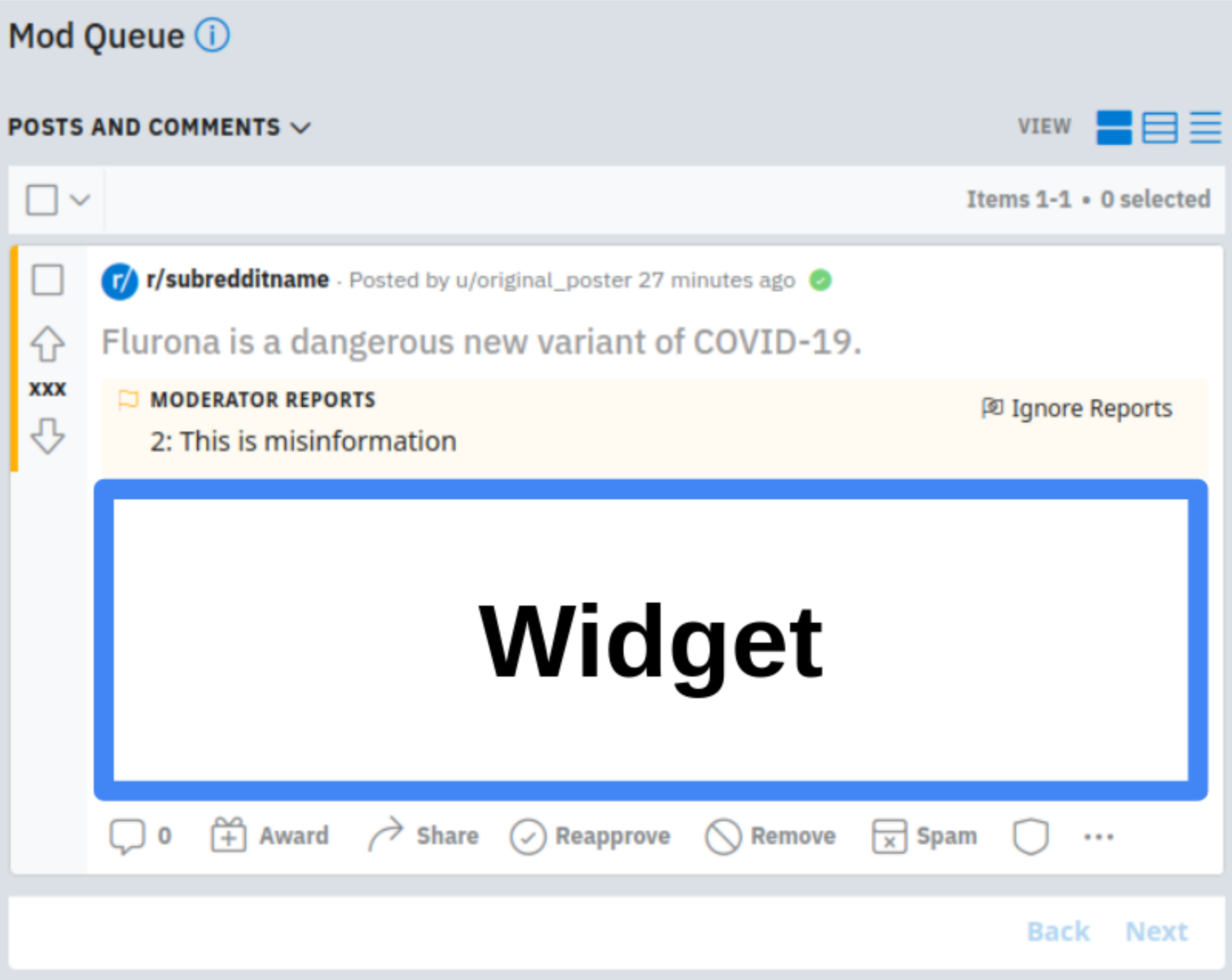
xxx

Flurona is a dangerous new variant of COVID-19.

MODERATOR REPORTS

2: This is misinformation

Ignore Reports

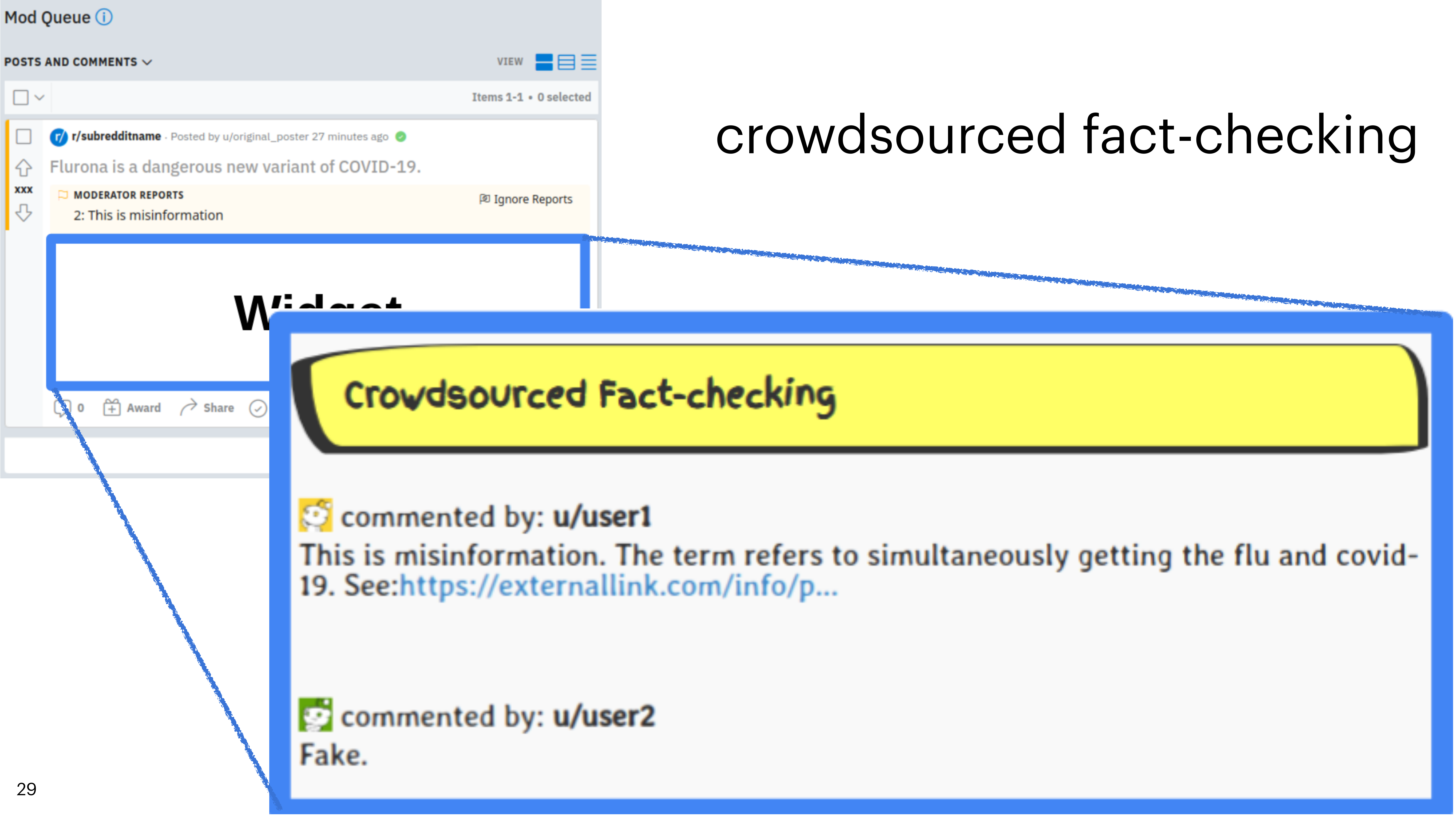


Flurona is a dangerous new variant of COVID-19.

2: This is misinformation

design probes shown during interviews

crowdsourced fact-checking



Mod Queue

POSTS AND COMMENTS

VIEW

Items 1-1 • 0 selected

r/subredditname · Posted by u/original_poster 27 minutes ago

Flurona is a dangerous new variant of COVID-19.

MODERATOR REPORTS

2: This is misinformation

Ignore Reports

Widget

Crowdsourced Fact-checking

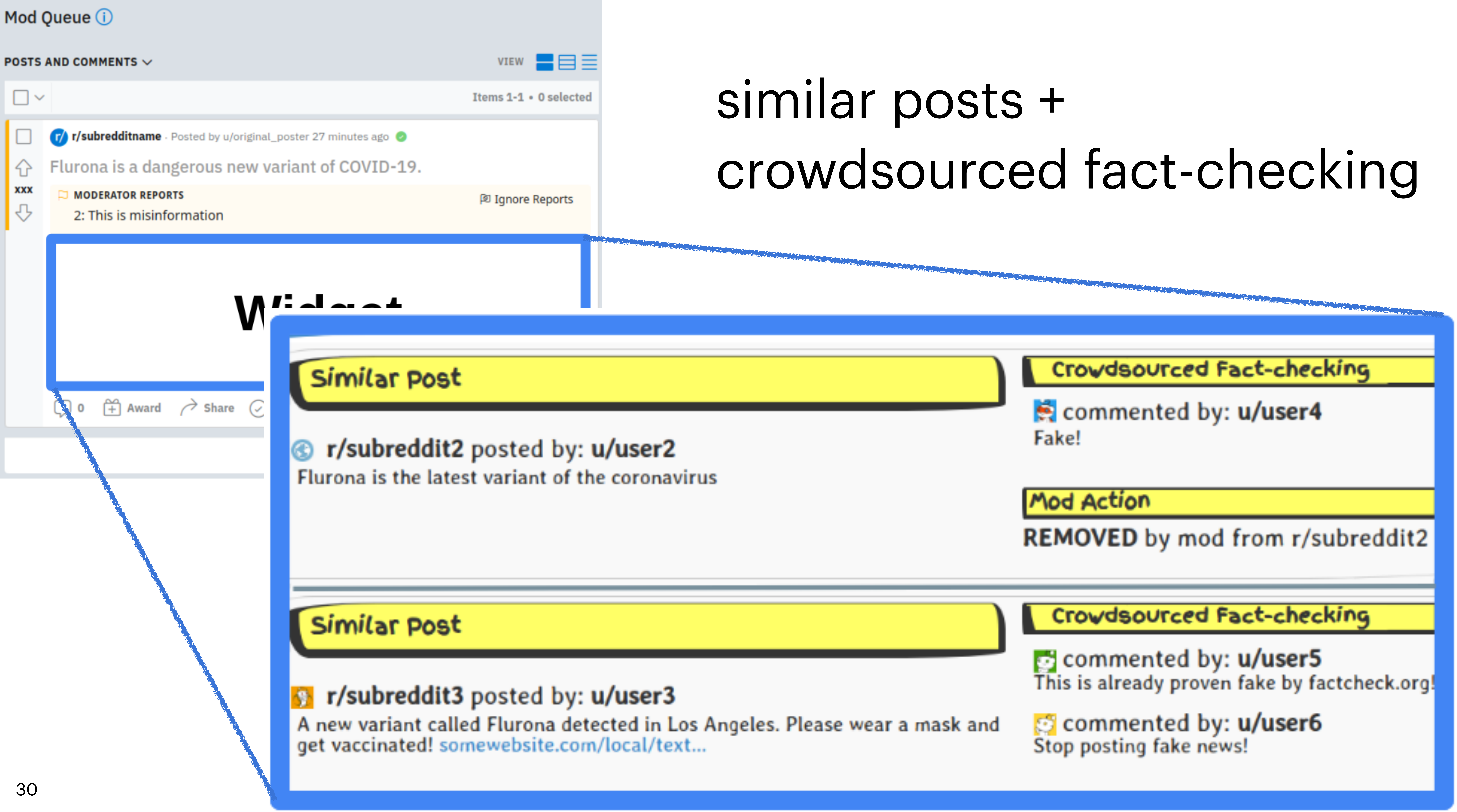
commented by: **u/user1**

This is misinformation. The term refers to simultaneously getting the flu and covid-19. See:https://externallink.com/info/p...

commented by: **u/user2**

Fake.

similar posts + crowdsourced fact-checking



Mod Queue ⓘ

POSTS AND COMMENTS

VIEW

□ Items 1-1 • 0 selected

□ r/subredditname · Posted by u/original_poster 27 minutes ago

Flurona is a dangerous new variant of COVID-19.

xxx

MODERATOR REPORTS

2: This is misinformation

Ignore Reports

Widget

0 Award Share

Similar Post

r/subreddit2 posted by: u/user2

Flurona is the latest variant of the coronavirus

Crowdsourced Fact-checking

commented by: u/user4

Fake!

Mod Action

REMOVED by mod from r/subreddit2

Similar Post

r/subreddit3 posted by: u/user3

A new variant called Flurona detected in Los Angeles. Please wear a mask and get vaccinated! somewebsite.com/local/text...

Crowdsourced Fact-checking

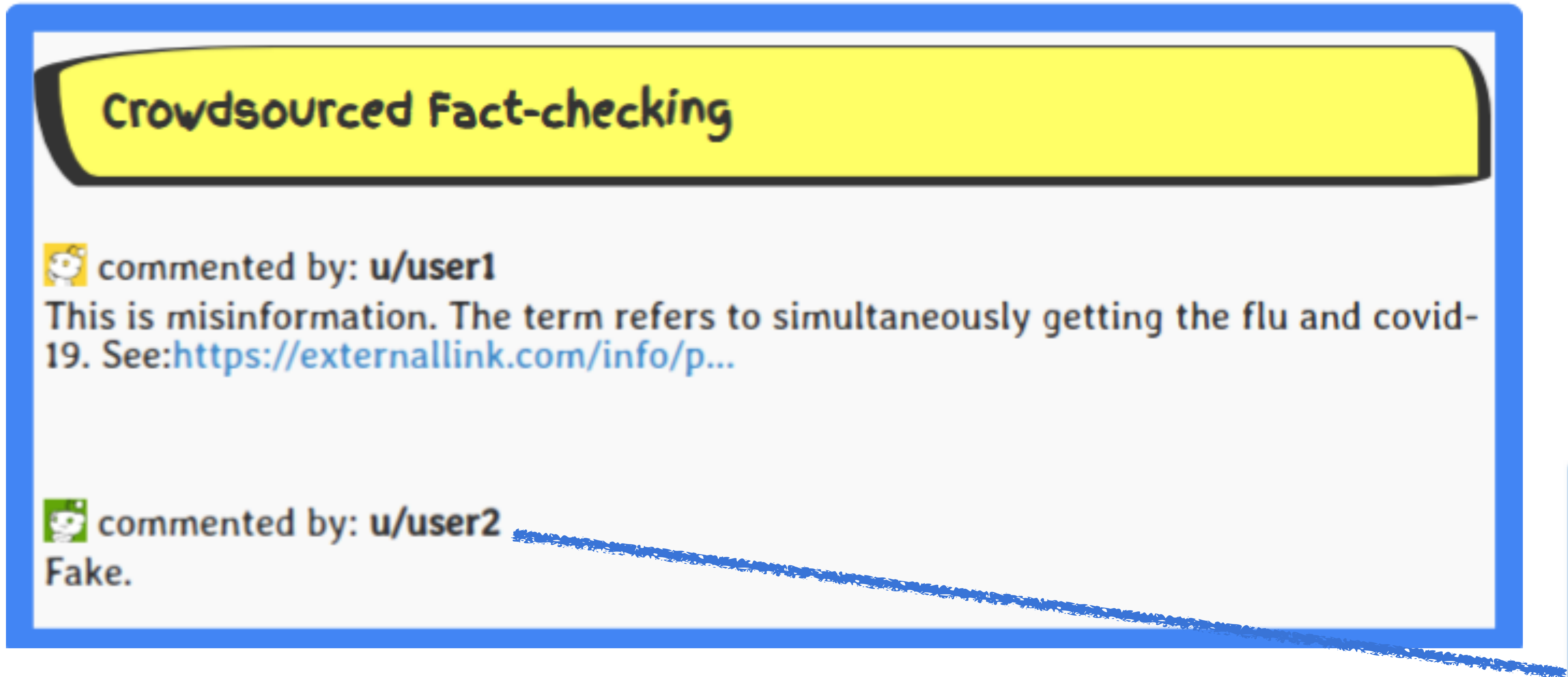
commented by: u/user5

This is already proven fake by factcheck.org!

commented by: u/user6

Stop posting fake news!

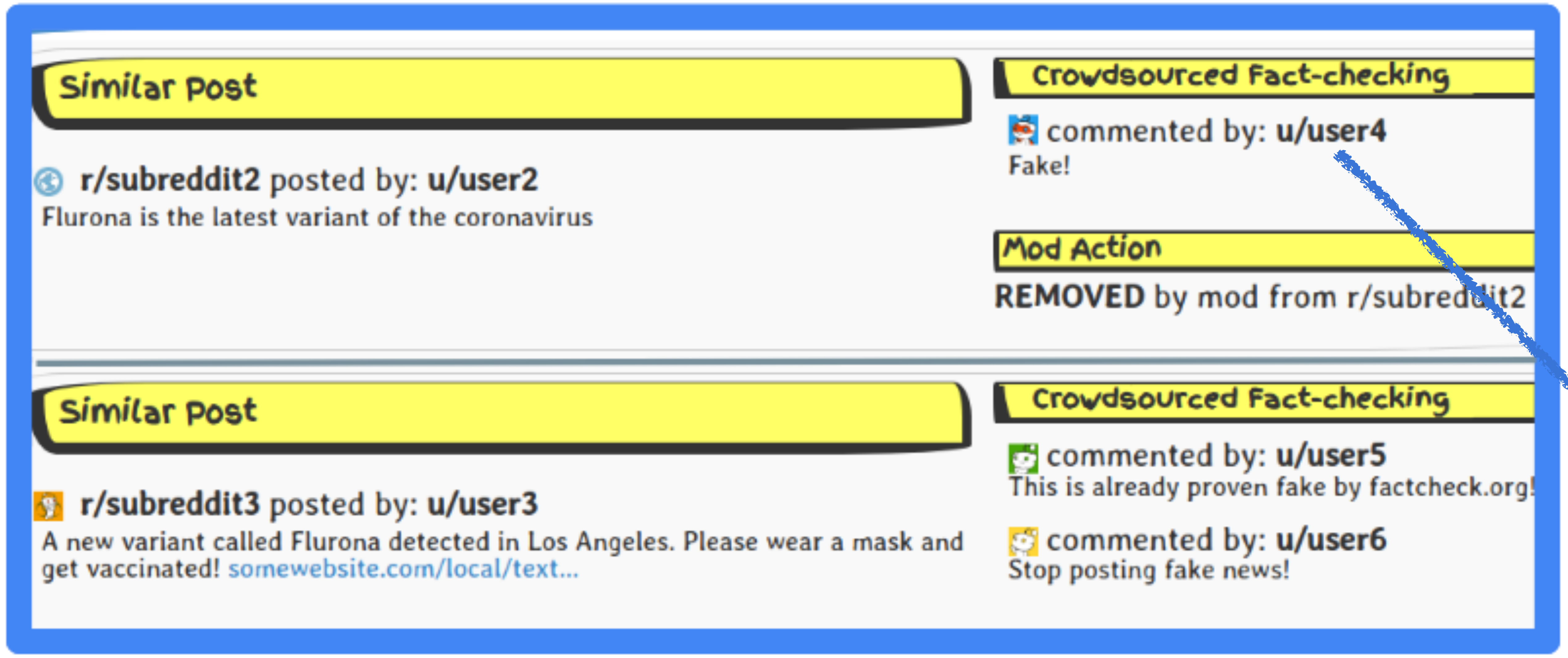
hovering over username shows user attributes



Crowdsourced Fact-checking

commented by: **u/user1**
This is misinformation. The term refers to simultaneously getting the flu and covid-19. See:https://externallink.com/info/p...

commented by: **u/user2**
Fake.



Similar Post

r/subreddit2 posted by: **u/user2**
Flurona is the latest variant of the coronavirus

Crowdsourced Fact-checking

commented by: **u/user4**
Fake!

Mod Action

REMOVED by mod from r/subreddit2

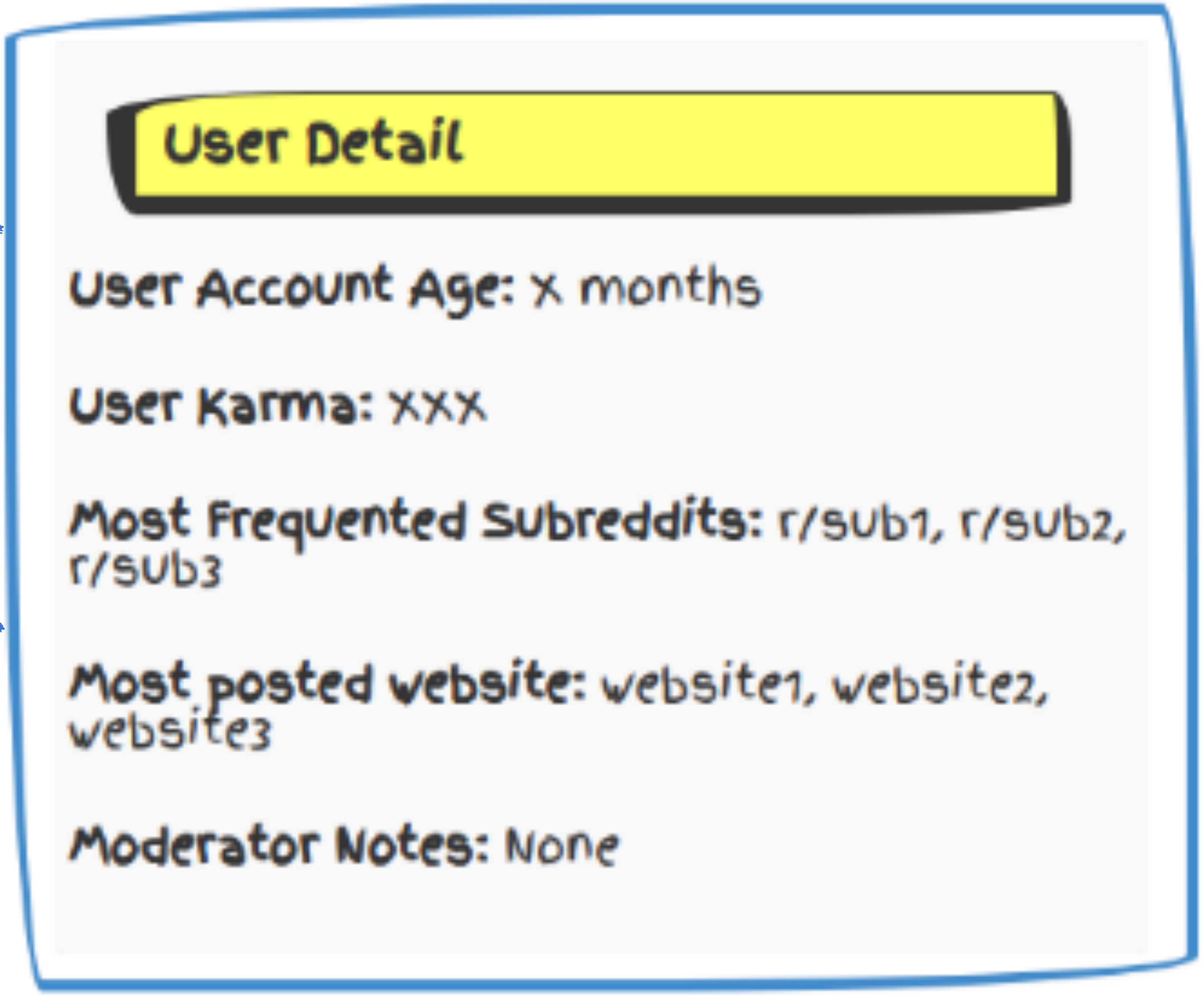
Similar Post

r/subreddit3 posted by: **u/user3**
A new variant called Flurona detected in Los Angeles. Please wear a mask and get vaccinated! somewebsite.com/local/text...

Crowdsourced Fact-checking

commented by: **u/user5**
This is already proven fake by factcheck.org!

commented by: **u/user6**
Stop posting fake news!



User Detail

User Account Age: x months

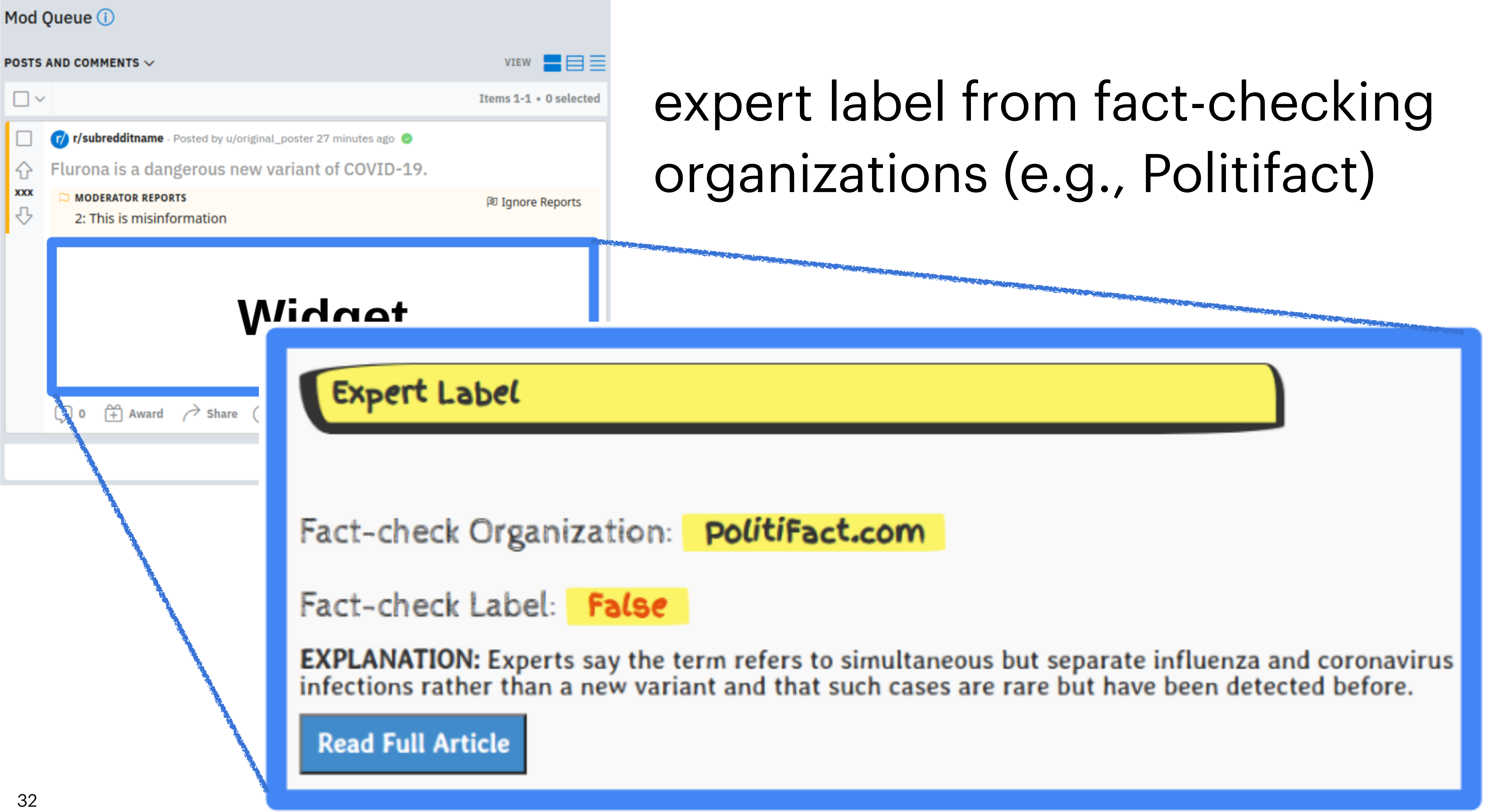
User Karma: xxx

Most Frequented Subreddits: r/sub1, r/sub2, r/sub3

Most posted website: website1, website2, website3

Moderator Notes: None

expert label from fact-checking organizations (e.g., Politifact)



r/subredditname · Posted by u/original_poster 27 minutes ago

Flurona is a dangerous new variant of COVID-19.

MODERATOR REPORTS

2: This is misinformation

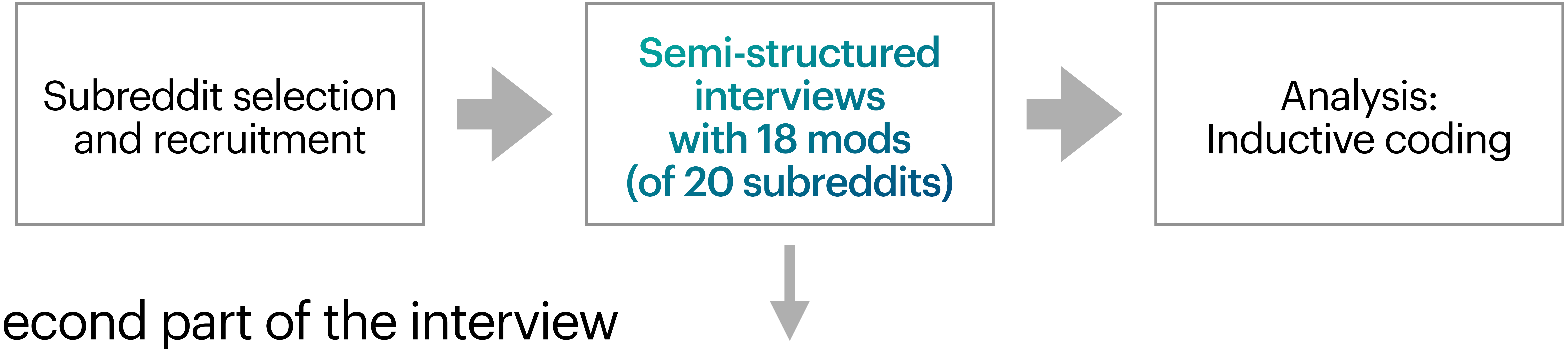
Expert Label

Fact-check Organization: PolitiFact.com

Fact-check Label: False

EXPLANATION: Experts say the term refers to simultaneous but separate influenza and coronavirus infections rather than a new variant and that such cases are rare but have been detected before.

Study Process



Subreddit selection and recruitment

Semi-structured interviews with 18 mods (of 20 subreddits)

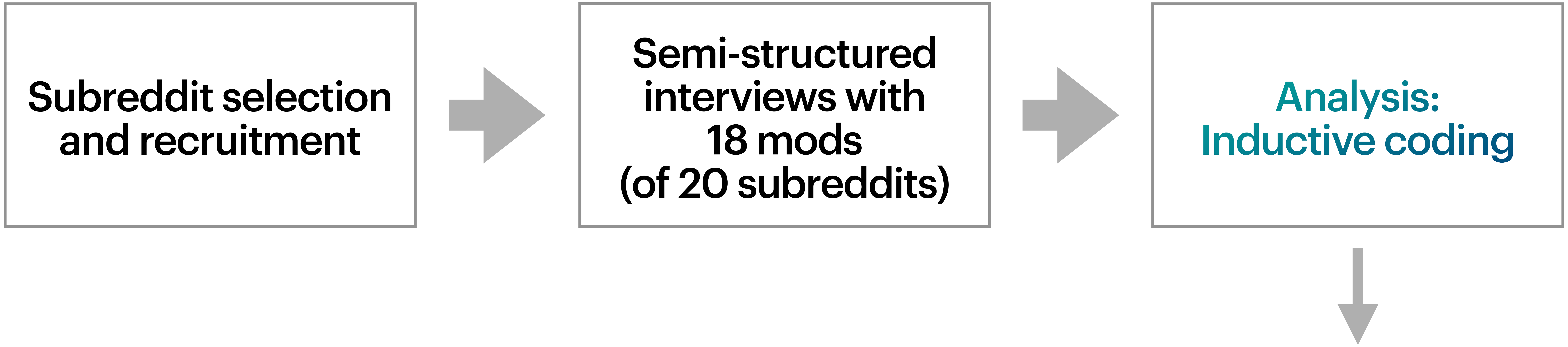
Analysis: Inductive coding

Second part of the interview

For each widget design:

1) how useful the mods find the widget
2) potential drawbacks of the widget

Study Process



Subreddit selection and recruitment

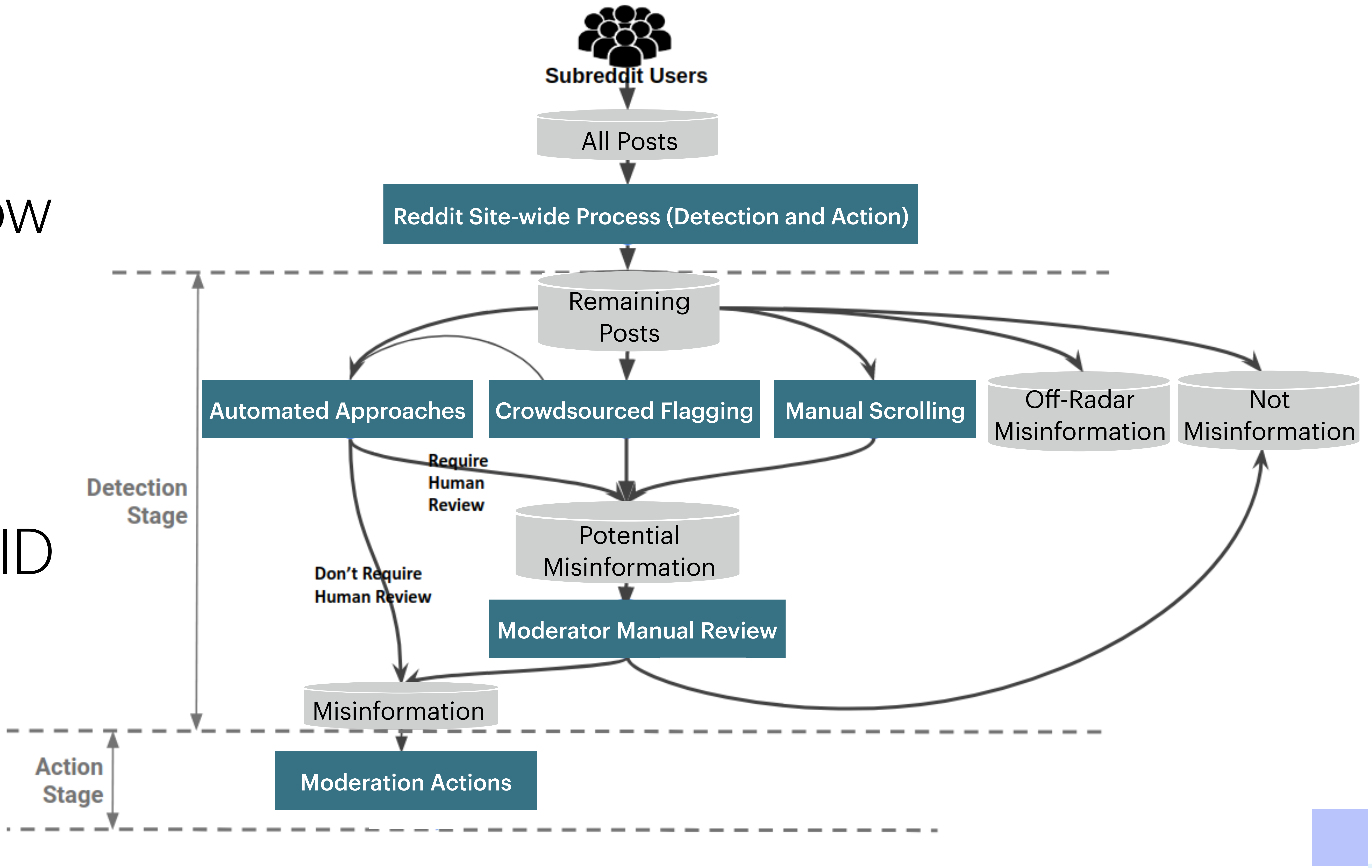
Semi-structured interviews with 18 mods (of 20 subreddits)

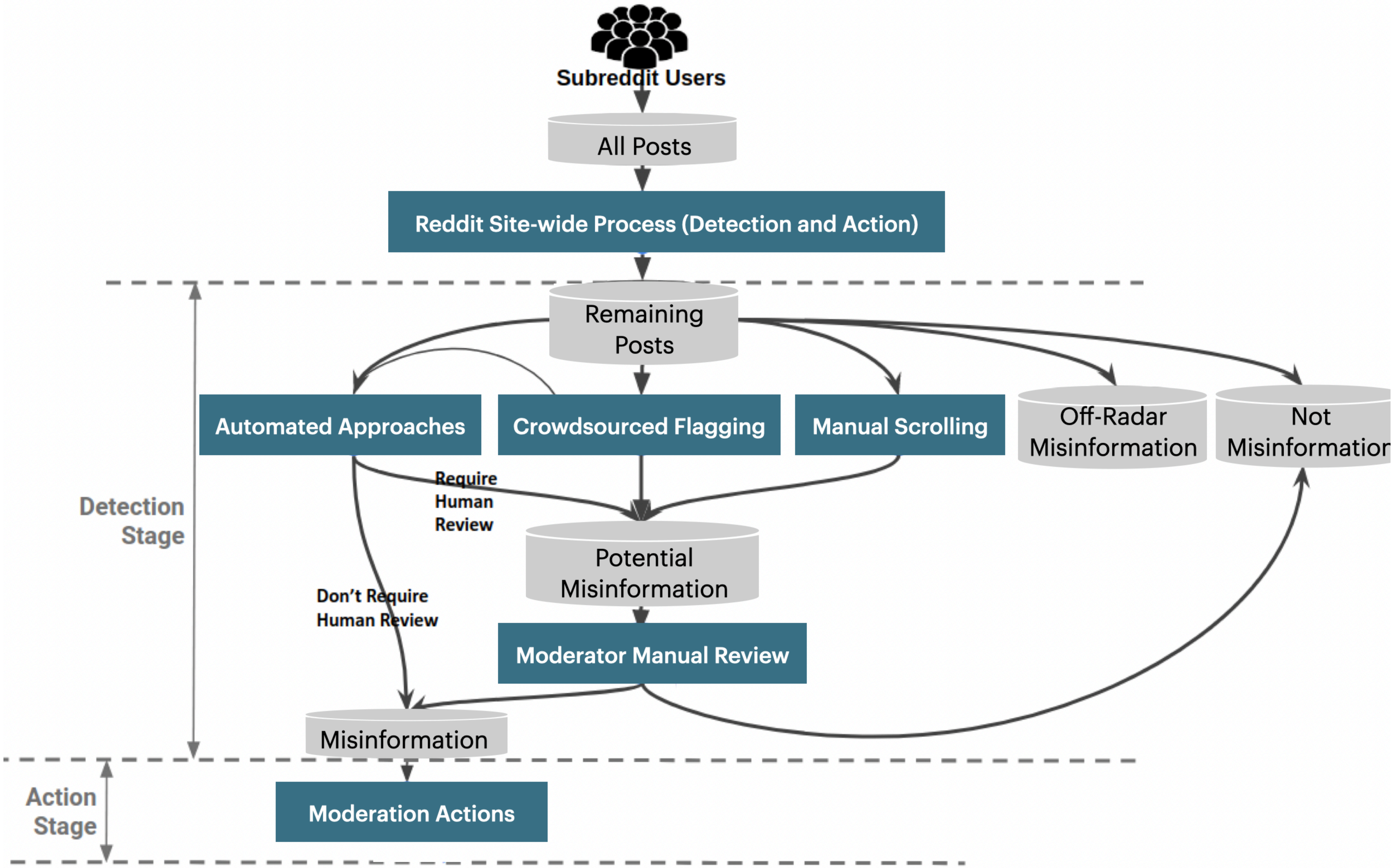
Analysis: Inductive coding

The first and second authors conducted line-by-line open coding.

Findings — Generalizable Moderation Workflow Model

A general workflow model that encapsulates processes for moderating COVID misinformation.




Subreddit Users
All Posts
Reddit Site-wide Process (Detection and Action)
Remaining Posts
Automated Approaches
Crowdsourced Flagging
Manual Scrolling
Off-Radar Misinformation
Not Misinformatior
Require Human Review
Don't Require Human Review
Potential Misinformation
Moderator Manual Review
Misinformation
Moderation Actions
Detection Stage
Action Stage

Findings — Generalizable Moderation Workflow Model

content facticity

user intent

perceived harm

Findings — Generalizable Moderation Workflow Model

content facticity

whether the content is about obvious, blatantly false claims (e.g., misinformation about vaccines being dangerous)

"Vaccines were found to be safe, specifically Pfizer, Moderna, and Johnson & Johnson. They were found to be overall safe. To be completely candid, we just don't have time for that anymore..." - [P18]

Findings — Generalizable Moderation Workflow Model

user intent

whether the user acted in good faith
(account age, karma, activity in different subreddits, activity gaps, most frequently posted websites, username, user ideology, etc.)

“Typically, we’ll look at the user’s history. And more often than not, it is somebody bouncing subreddit to subreddit and we’ll just delete their comment, ban them and just move on...” - [P10]

Findings — Generalizable Moderation Workflow Model

perceived harm

prioritize misinformation that they perceive would lead to harm at the individual level or at the societal level

"But it is dangerous that you think that [alternative cures] will cure Coronavirus. Because I really felt like this could lead to someone getting hurt. And so, that's something I was really particular about."
- [P17]

Findings — Wisdom of Two Crowds

wisdom of users wisdom of mods

Almost all participants are heavily reliant on reports from crowds of users to **identify potential misinformation.**

Most widely used method — compared to automated approaches or manual scrolling.

Findings — Wisdom of Two Crowds

wisdom of users wisdom of mods

“So there were the users flagging. It is the number one way it [potential misinformation] comes to our attention.

Number two is by scanning the background of new users coming into the community, [using Saferbot] to flag people that may be liable to post misinformation.” - [P1]

Findings — Wisdom of Two Crowds

wisdom of users wisdom of mods

Barrier — **Prevalent misuse** of reporting feature

“We started getting a lot of trolling and spamming of moderator reports by people who didn’t like the moderation. That’s probably the single biggest problem that I’ve had in the whole process, …” - [P9]

Findings — Wisdom of Two Crowds

wisdom of users wisdom of mods

Moderators want **more information about users** to combat users' misuse of reporting.

"Let the moderators see who's doing the reporting. Because I strongly suspect that a lot of those types of reports, the frivolous reporting, is the same subset of people. It's the same small group of people doing it." - [P5]

Findings — Wisdom of Two Crowds

wisdom of users

wisdom of mods

When encountering **difficult, ambiguous cases,** participants received support from moderation teams and other subreddits' moderators, often using Discord or Slack.

Findings — Wisdom of Two Crowds

wisdom of users

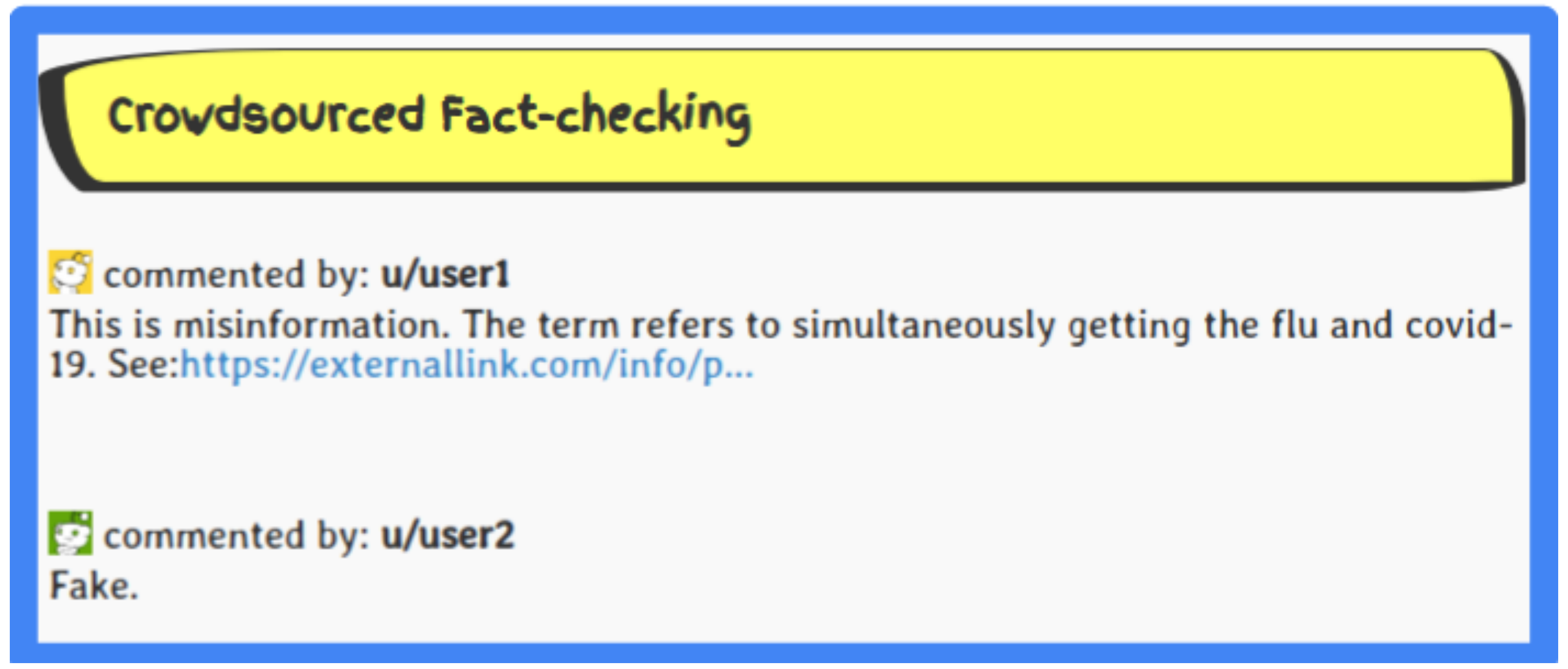
wisdom of mods

“What we would do is, we would hop into Discord with them [moderators of other subreddits]. They gave us access to their general Discord that they use for coordinating and planning.

And then, if they had questions on whether something that someone was talking about was actually true or not, they could just bring it to us.” - [P3]

Findings: Mods’ Thoughts on Modqueue Designs

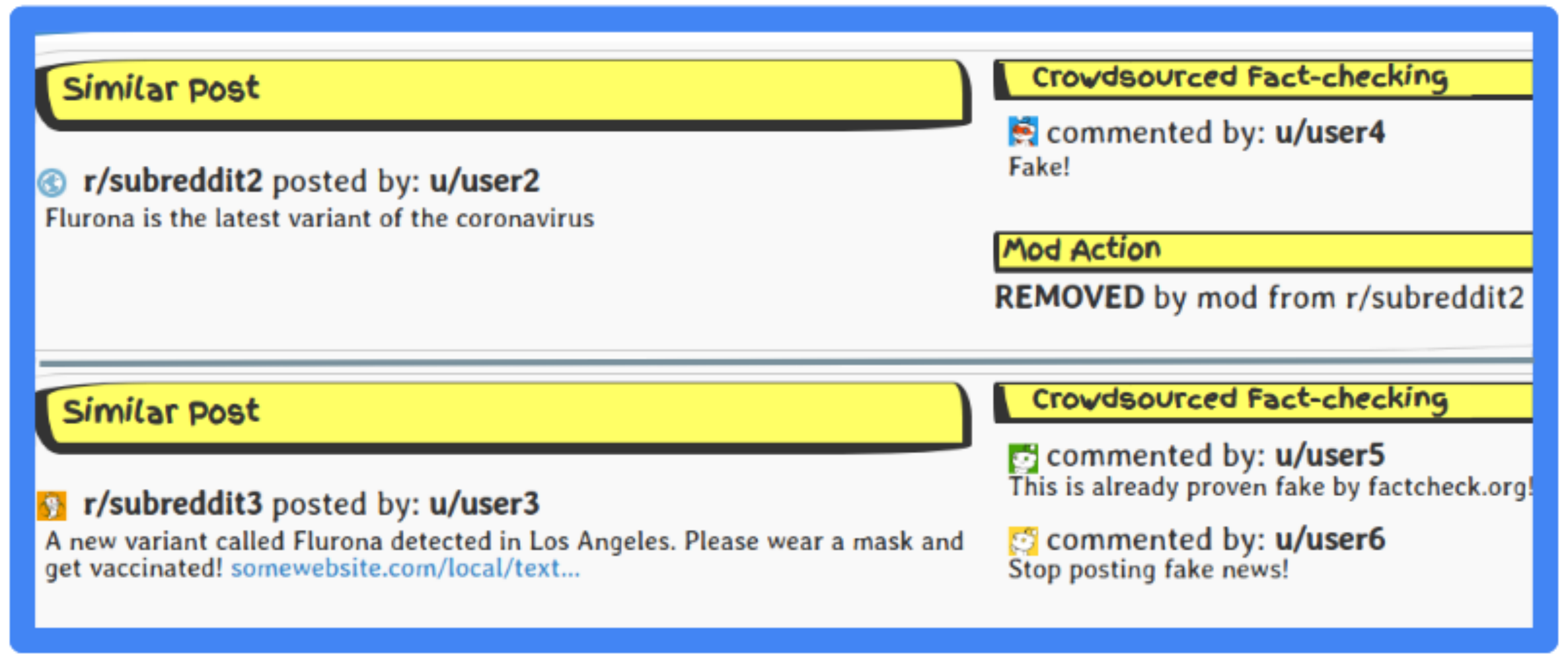
crowd fact-checking



Crowdsourced Fact-checking

commented by: **u/user1**
This is misinformation. The term refers to simultaneously getting the flu and covid-19. See:https://externallink.com/info/p...

commented by: **u/user2**
Fake.



Similar Post

r/subreddit2 posted by: **u/user2**
Flurona is the latest variant of the coronavirus

Crowdsourced Fact-checking

commented by: **u/user4**
Fake!

Mod Action

REMOVED by mod from r/subreddit2

Similar Post

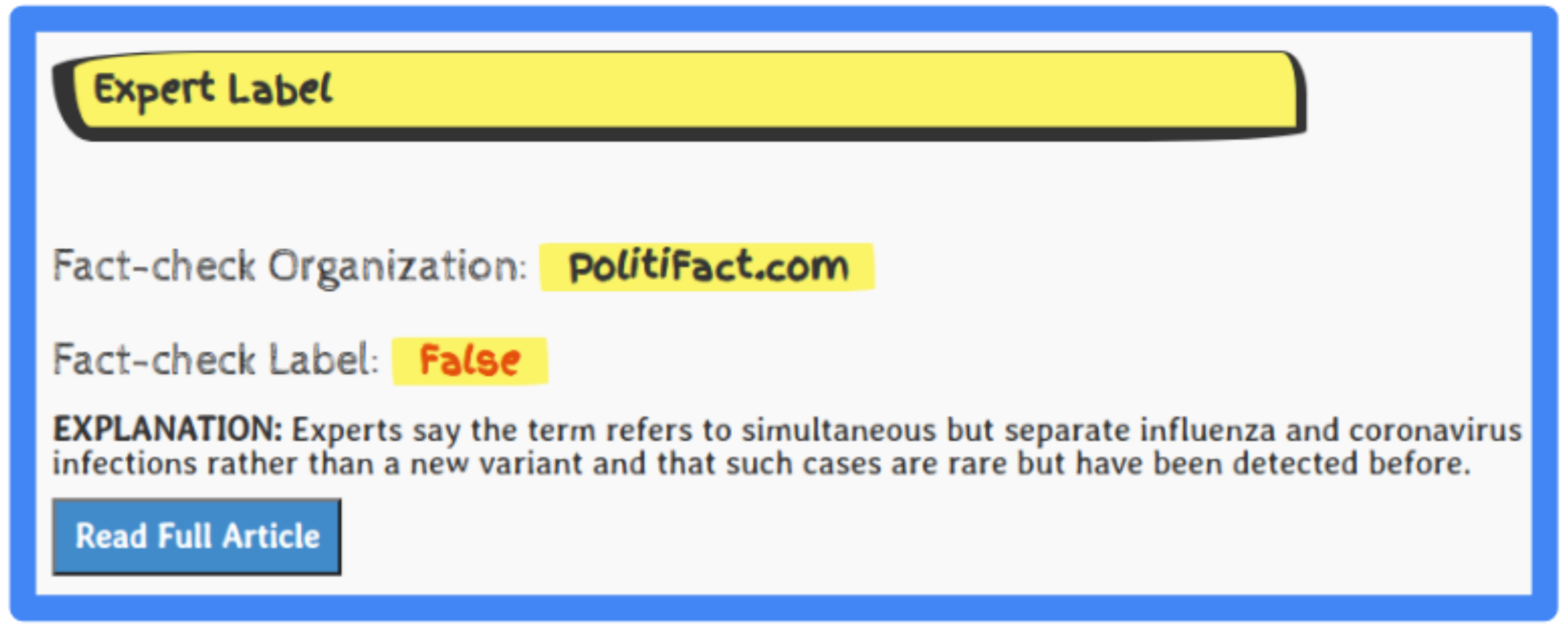
r/subreddit3 posted by: **u/user3**
A new variant called Flurona detected in Los Angeles. Please wear a mask and get vaccinated! somewebsite.com/local/text...

Crowdsourced Fact-checking

commented by: **u/user5**
This is already proven fake by factcheck.org!

commented by: **u/user6**
Stop posting fake news!

expert labels



Expert Label

Fact-check Organization: PolitiFact.com

Fact-check Label: False

EXPLANATION: Experts say the term refers to simultaneous but separate influenza and coronavirus infections rather than a new variant and that such cases are rare but have been detected before.

Read Full Article

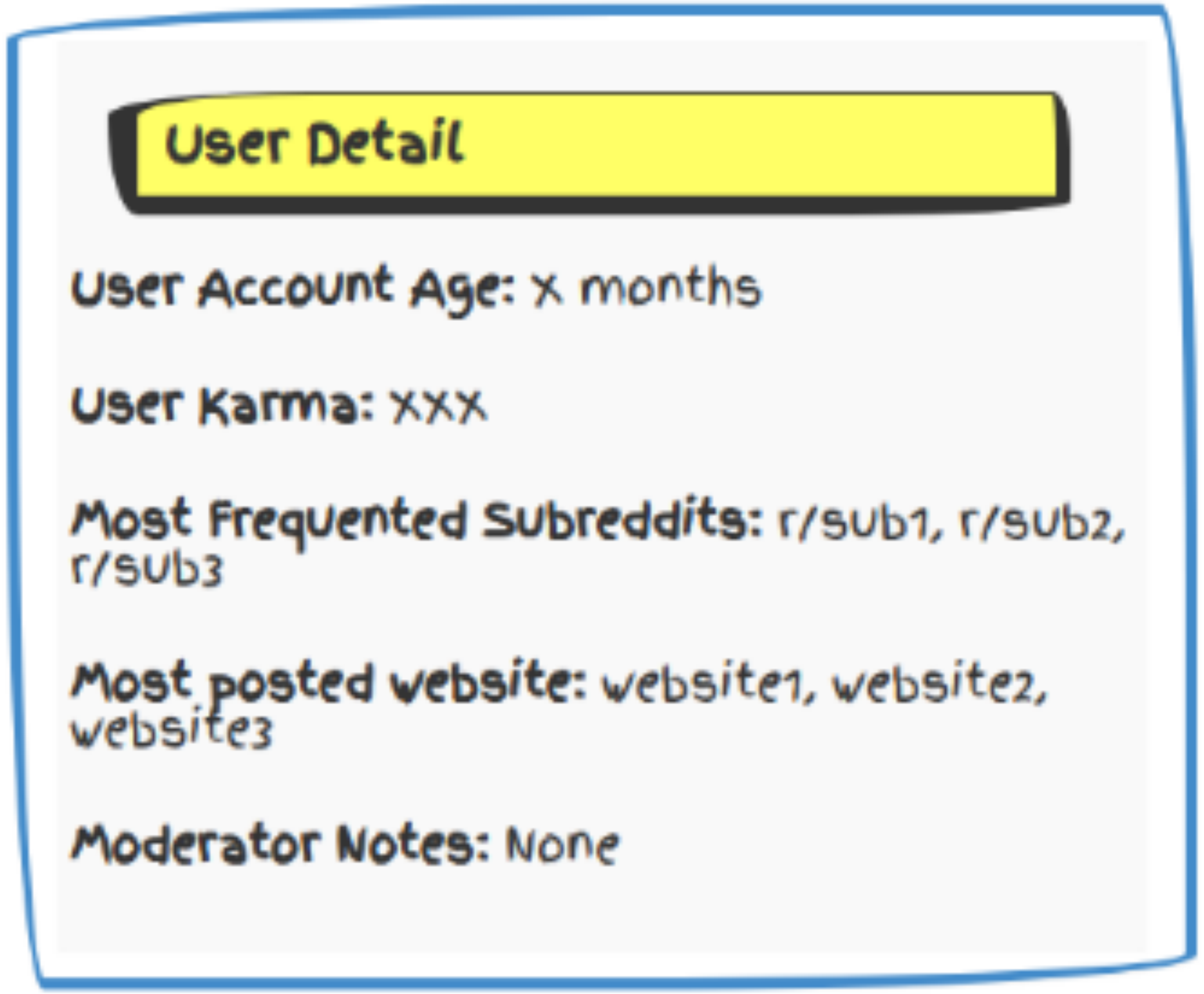
The vast majority of participants viewed all designs favorably.

Findings: Mods’ Thoughts on Modqueue Designs

crowd fact-checking

expert labels

Almost all participants appreciated user popup that provide more account information.



User Detail

User Account Age: x months

User Karma: xxx

Most Frequented Subreddits: r/sub1, r/sub2, r/sub3

Most posted website: website1, website2, website3

Moderator Notes: None

Findings: Mods’ Thoughts on Modqueue Designs

crowd fact-checking expert labels

Nearly half **preferred crowd signals** over labels from professional fact-checkers.

Why?

Because crowd signals can assist participants with **evaluating user intent**.

Findings: Mods’ Thoughts on Modqueue Designs

crowd fact-checking

expert labels

“Yeah. That [user reputation] would absolutely add weight to making a decision probably easier. Google does this with their reviewers. If I see, like Wikipedia, somebody who’s done lots of good articles that I would know, I would rely on their report quicker. Or, if I see somebody has a new account, then I’m gonna look at it a little closer.” - [P10]

Findings — Mods' Feedback on Alternative Modqueue Designs

crowd fact-checking

expert labels

A quarter of the participants **distrust professional fact-checkers**, raising important concerns about misinformation moderation.

Why?

Because they viewed fact-checking organizations as being too politicized or lacking medical expertise.

Takeaways

1. There exists a **general workflow model** that encompasses processes mods use for handling misinformation. It's centered around **content facticity**, **user intent**, and **perceived harm**.
2. Platforms should **make it easier for mods to leverage both wisdom of crowds of users and mods**, given that mods find them helpful.
3. Platforms should provide mods with more **signals about user characteristics** to leverage crowd wisdom effectively.

Takeaways

1. There exists a **general workflow model** that encompasses processes mods use for handling misinformation. It's centered around **content facticity**, **user intent**, and **perceived harm**.
2. Platforms should **make it easier for mods to leverage both wisdom of crowds of users and mods**, given that mods find them helpful.
3. Platforms should provide mods with more **signals about user characteristics** to leverage crowd wisdom effectively.

Takeaways

1. There exists a **general workflow model** that encompasses processes mods use for handling misinformation. It's centered around **content facticity**, **user intent**, and **perceived harm**.
2. Platforms should **make it easier for mods to leverage both wisdom of crowds of users and mods**, given that mods find them helpful..
3. Platforms should provide mods with more **signals about user characteristics** to leverage crowd wisdom effectively.

Thank you!

Lia's contact:
liafan@uw.edu

Jane's contact:
imjane@umich.edu

bit.ly/moderation-crowd-wisdom

Takeaways

1. There exists a **general workflow model** that encompasses processes mods use for handling misinformation. It's centered around **content facticity**, **user intent**, and **perceived harm**.
2. Platform designers should make it easier for mods to leverage both wisdom of **crowds of users and mods**.
3. Platforms should provide mods with more **signals about user characteristics** to leverage crowd wisdom effectively.