

# Community Notes

**Jay Baxter, Sr. Staff ML Engineer @ X**

Presenting joint work with many others, including Keith Coleman,  
Sophie Hilgard, Brad Miller, Daniel Ortiz, and Jiansong Chao



Crypto Rover   
@rovercric

...

Trump is now leading Harris with 20% in the polls!

If he wins, the biggest [#Bitcoin](#) bull market ever starts!


## 2024 Election Forecast

Live and accurate forecasts by the world's largest prediction market.

Election in 0 DAYS 0 HRS 0 MIN 0 SEC

Presidency

Senate

 How it works  Share  Embed



Trump

60.7%  
+2.7%

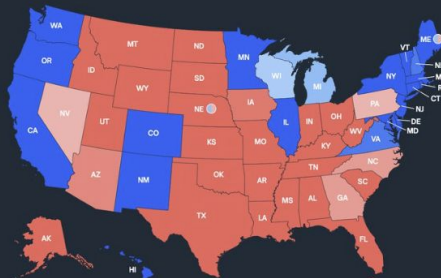


Polymarket  
[polymarket.com/elections](https://polymarket.com/elections)

39.3%  
+2.8%



Harris



Readers added context they thought people might want to know

This is a prediction market, not a poll.

[polymarket.com/elections](https://polymarket.com/elections)


Do you find this helpful?


Rate it

Context is written by people who use X, and appears when rated helpful by others. [Find out more.](#)





✕ Note details

 Note originally added to the image on this post, and is showing on 268 posts that include this image  
[See all posts with this note](#)

 1.5M+ views

Top tags selected by raters

-  Cites reliable sources
-  Easy to understand

Note Author

 Upstanding Creek Pigeon

Is your note about the post or the image?

About this specific post



About the image in this post, and should appear on all posts that include this image



9:41



← Posts with the same media



Kian@naturelvr49 · 14h



Readers added context

Is this note helpful?

Rate it



Kbpip@bip\_pop\_3 · 14h



# Notes are accurate & inform understanding

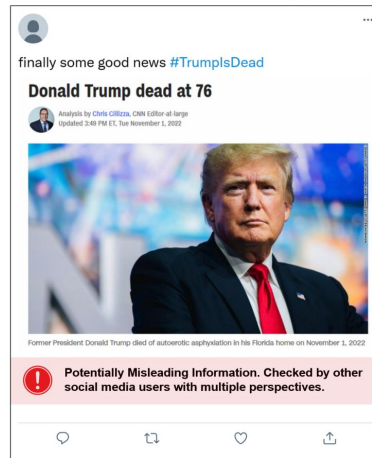
- **Accurate**
  - E.g. 97.5-99.5% accurate on COVID

1. Allen, Desai, Namazi et al  
<https://jamanetwork.com/journals/jama/fullarticle/2818054>

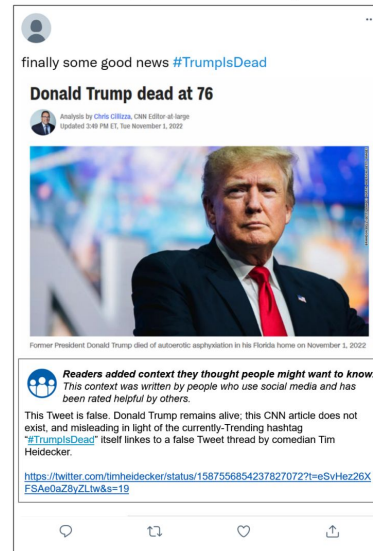
# Notes are accurate & inform understanding

- **Accurate**
  - E.g. 97.5-99.5% accurate on COVID
- **Inform Understanding**
  - Community Notes found more trustworthy than simple misinformation flags, bipartisanly (in survey)
  - Survey participants rate posts as less accurate when displayed with a note (vs. w/no note)

Condition (3) – *community flag*



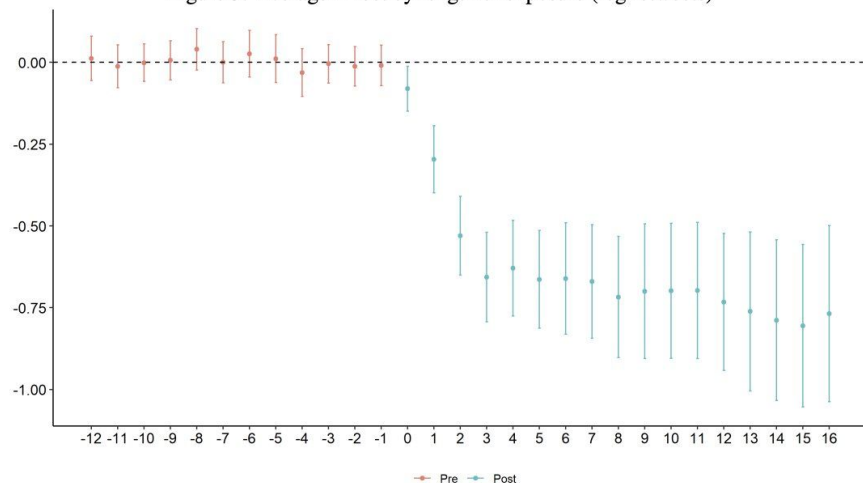
Condition (4) – *community note*



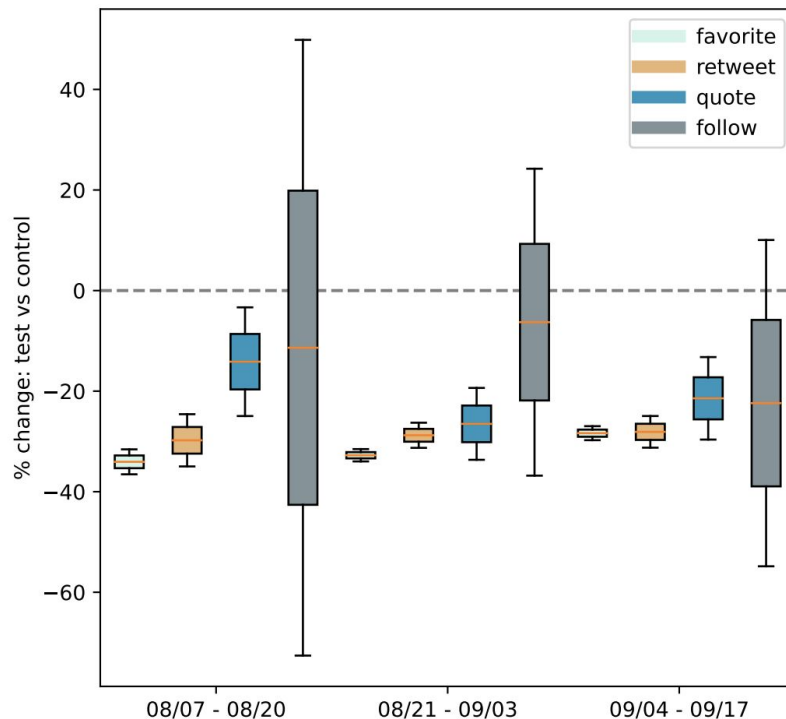
1. Allen, Desai, Namazi et al  
<https://jamanetwork.com/journals/jama/fullarticle/2818054>
2. Drolsbach, Solovev, Pröllochs 2024 <https://osf.io/preprints/osf/ydc42>

# Organic drop in Likes & Reposts on noted posts

Figure 5: Average Effect by length of exposure (log retweets)



Note : This Figure shows the average effect by length of exposure before (in red) and after (in blue) the treatment. The dependent variable is the number of retweets (in log)

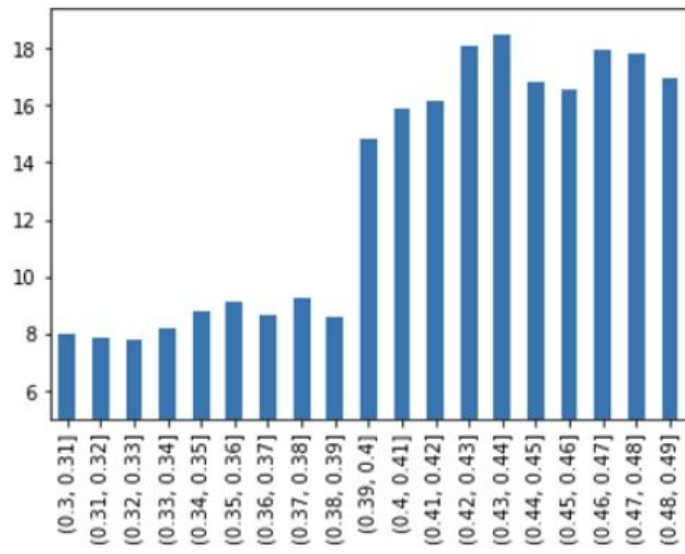


1. Renault, Amariles, Troussel 2024 <https://arxiv.org/abs/2404.02803>
2. Chuai, Pilarski, Lenzini, Pröllochs 2024 <https://osf.io/preprints/osf/3a4fe>
3. Wojcik et al 2022 <https://arxiv.org/abs/2210.15723>

# Impact on Post Deletion & Creation

- Noted posts are 80% more likely to be deleted by their authors
- Incentive change:
  - Users say they post differently knowing that they could be noted (and demonetized)

Figure 4: Tweet deletion



Note : This Figure shows the percentage of deleted tweets by Note Helpfulness Score



# Credible Neutrality

- **Contributors are regular users** (must have verified phone #, signed up >6mos ago...)
- **No external ground truth labels:** use bridging mechanism instead
- **Open source, transparent, reproducible, verifiable**
  - X never manually edits or changes the status of a note to take it down



# Transparency & verifiability

Open source algorithm & public data



## Download data

### Community Notes is built on data transparency

All Community Notes data are published here daily, so people have free access to analyze it, identify problems, and spot opportunities to make Community Notes better. We can't wait to learn with you.

Learn how to use and analyze Community Notes data [in our guide](#).

## Notes data

File no. 1 of 1



## Ratings data

File no. 1 of 8



twitter / communitynotes

Type  to search

<> Code

Issues 20

Pull requests 7

Discussions

Actions

Security

Insights

Settings

communitynotes

Public

Edit Pins

Unwatch 54

main

50 Branches

0 Tags

Go to file

Add file

<> Code

jbxter

Merge pull request #220 from twitter/jbxter/2024\_04\_26

adbc126 · 3 days ago

456 Commits

.github/workflows	Fix GitHub Workflows _redirects directory location (#75)	last year
documentation	Update ranking-notes.md	4 days ago
sourcecode	Freeze rater parameters in final scoring, and turn on stat...	3 days ago
.gitignore	Update gitignore	4 months ago
CODE_OF_CONDUCT.md	Replace Twitter strings (#128)	9 months ago
LICENSE	Update LICENSE	last year
README.md	Update README.md	3 days ago
birdwatch_paper_2022_10_27.pdf	Add Birdwatch paper PDF	2 years ago
requirements.txt	Freeze rater parameters in final scoring, and turn on stat...	3 days ago

# Bridging Algorithm (no external ground truth)



Readers added context they thought people might want to know

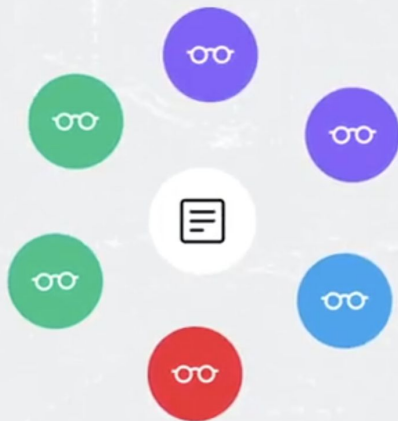
Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings.

[communitynotes.twitter.com/guide/en/about...](https://communitynotes.twitter.com/guide/en/about...)

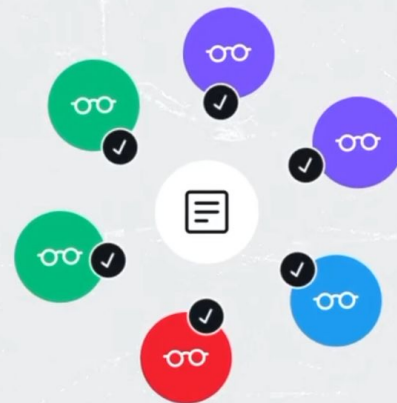
Do you find this helpful?

**Rate it**

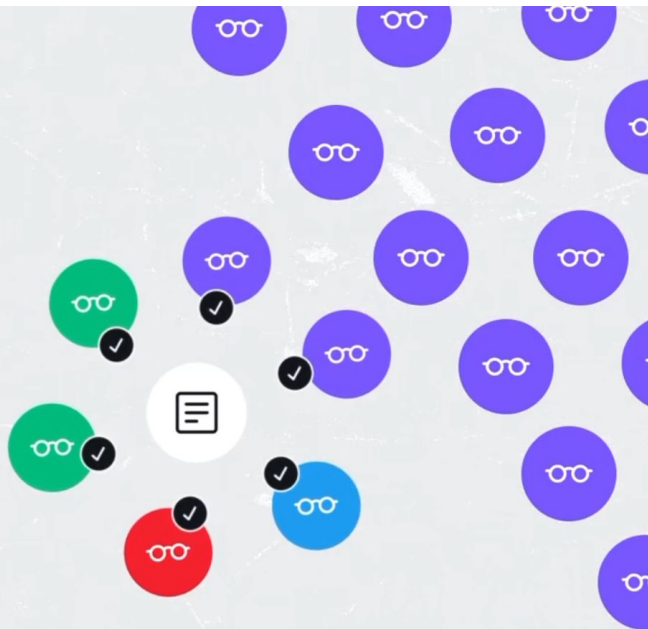
Context is written by people who use X, and appears when rated helpful by others. [Find out more.](#)



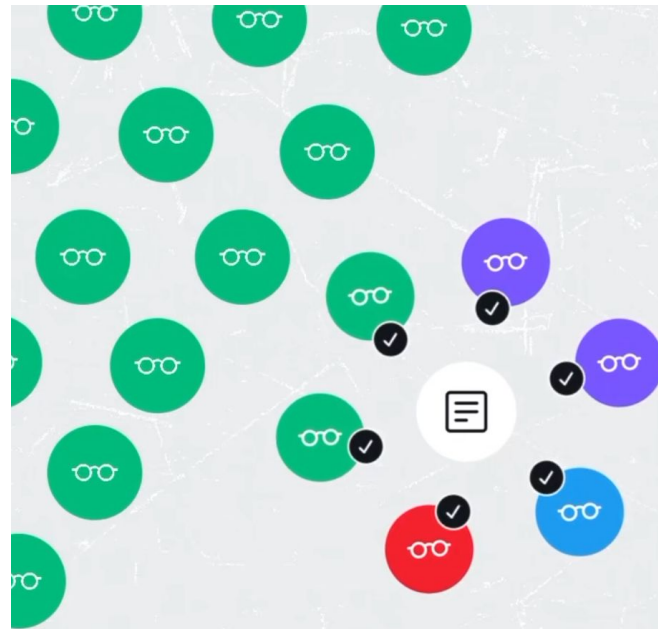
**Community Notes are  
rated by contributors of  
multiple perspectives**



**And only notes that have  
a broad appeal are  
shown on Tweets**

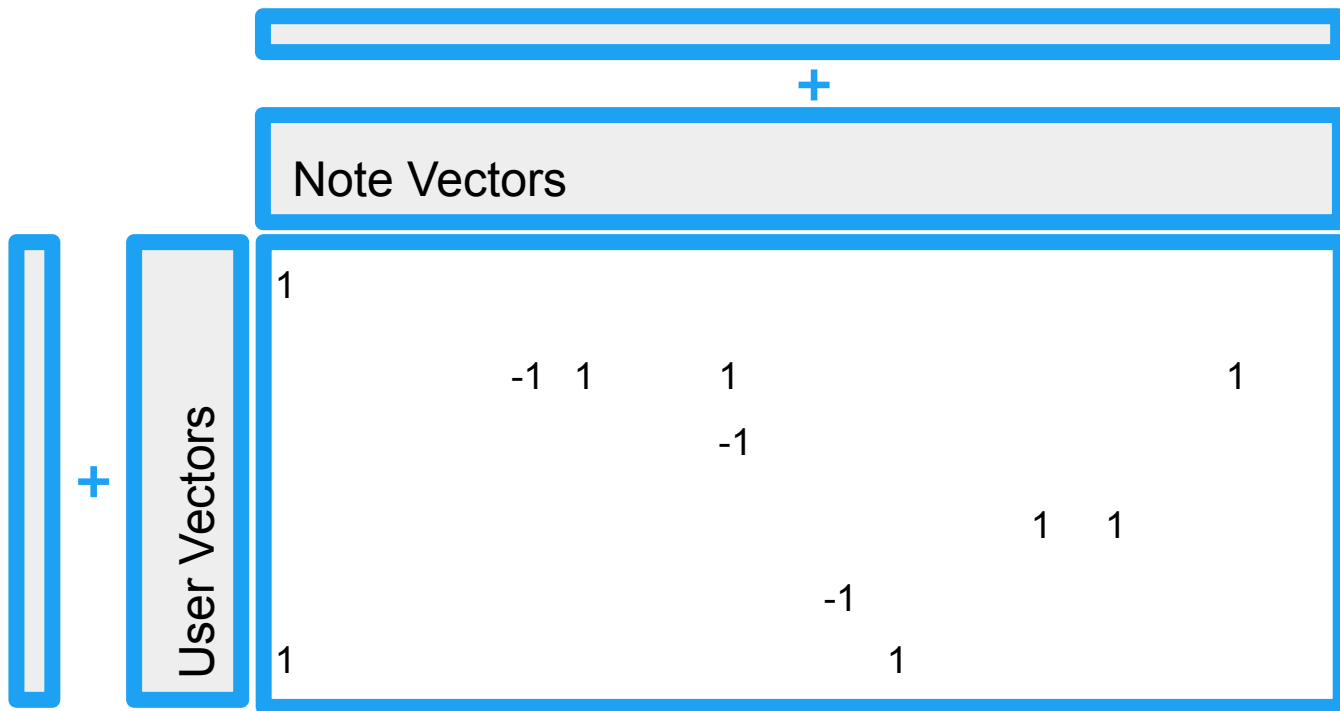


**This means that pile-ons  
aren't effective**



**And one group alone can't  
determine what notes  
get shown**

# Factorize the note rating matrix



1: rated helpful  
-1: rated unhelpful

# Finding broadly helpful notes: matrix factorization

Recommender systems  
typically rank by this

$$\hat{r}_{user,note} = \vec{v}_{user} \cdot \vec{v}_{note} + i_{user} + i_{note} + i_{glob}$$

Predicted Rating  
(from user, of note)

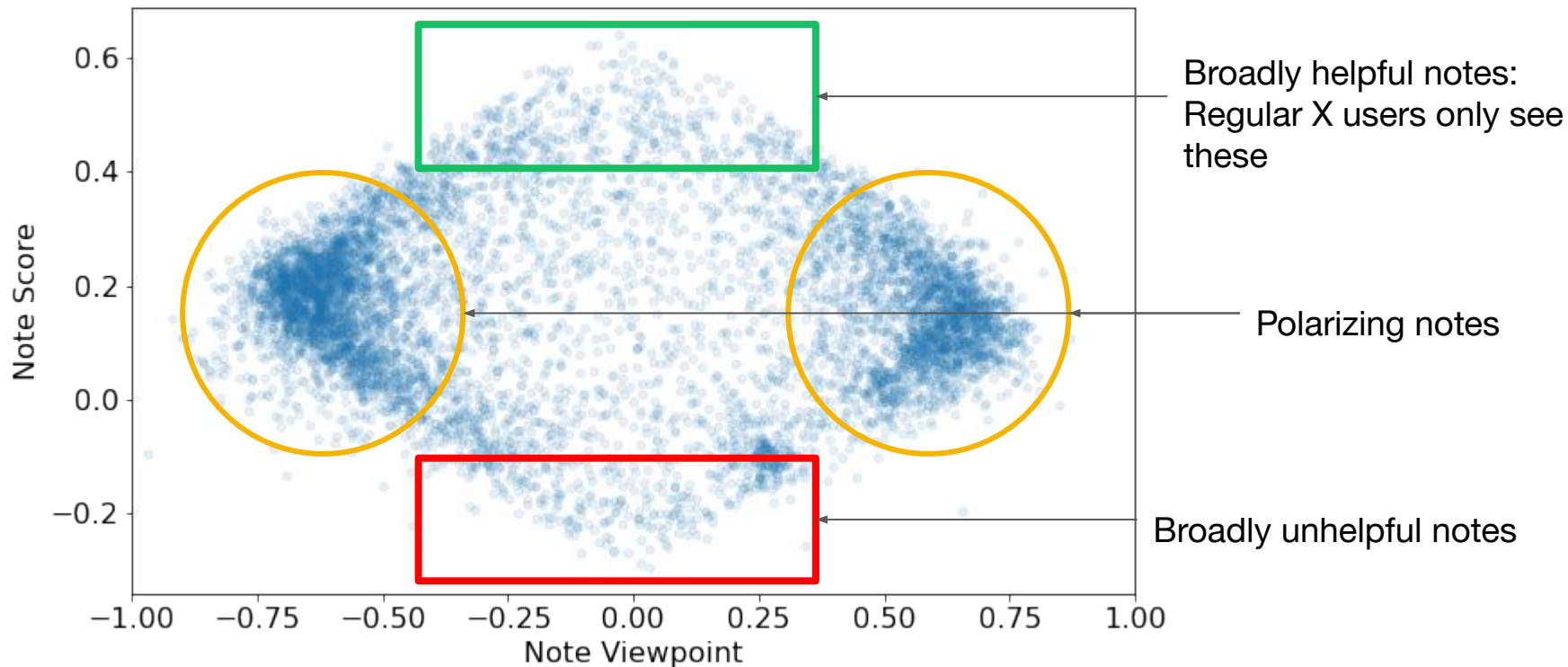
User and Note  
Viewpoint Similarity

Intercept terms  
(How high baseline ratings are)

Community Notes uses the note intercept  
as a non-personalized note score.

(Add extra regularization to force ratings  
to be explained by viewpoints whenever  
possible)

# The best (and worst) notes receive broad consensus





# Rating Tags: Beyond Helpfulness

Run same bridging matrix  
factorization on tags

Bridging-based tag consensus can  
override helpful consensus

## What was unhelpful about it?

Sources not included or unreliable

☐

Sources do not support note

☐

Incorrect information

☐

Opinion or speculation

☐

Typos or unclear language

☐

Misses key points or irrelevant

☐

Argumentative or biased language

☐

Note not needed on this post

☐

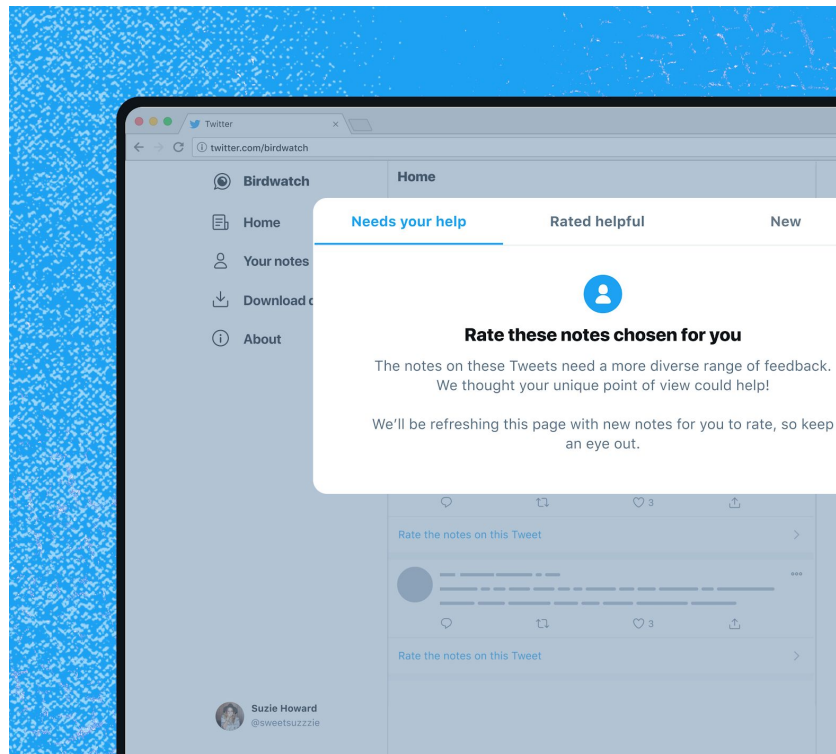
Other

☐

Submit

# Needs Your Help (Tab + Notifications)

- Address selection bias
  - Make sure enough raters with each viewpoint have rated a note
- Mitigate coordinated manipulation
  - If distribution of organic ratings is very different than Needs Your Help ratings => don't trust organic ratings



# Reputation (& similarities to prediction markets)

×

115

Rating Impact

Last update: 51 minutes ago

+38

**Ratings that helped a note earn the status of Helpful**

Nice work! These ratings identified Helpful notes that get shown as context on posts and help keep people informed.

+85

**Ratings that helped a note reach the status of Not Helpful**

These ratings improve Community Notes by giving feedback to note authors, and allowing contributors to focus on the most promising notes.

-8

**Ratings of Not Helpful on notes that ended up with a status of Helpful**

Don't worry, everyone gets some of these! These ratings are common and can lead to status changes if enough people agree that a "Helpful" note isn't sufficiently helpful.

-0 **×2**

**Ratings of Helpful on notes that ended up with a status of Not Helpful**

These ratings are counted twice because they often indicate support for notes that others deemed low-quality.

- You can think of each note as a “prediction market” that resolves to its final status (helpful or not helpful)
- Rater reputation goes up if your rating matches the outcome, down if it doesn't
  - Ratings are secret until 48 hours later

# Reputation

×

**115**  
**Rating Impact**  
Last update: 51 minutes ago

**+38**  
**Ratings that helped a note earn the status of Helpful**  
Nice work! These ratings identified Helpful notes that get shown as context on posts and help keep people informed.

**+85**  
**Ratings that helped a note reach the status of Not Helpful**  
These ratings improve Community Notes by giving feedback to note authors, and allowing contributors to focus on the most promising notes.

**-8**  
**Ratings of Not Helpful on notes that ended up with a status of Helpful**  
Don't worry, everyone gets some of these! These ratings are common and can lead to status changes if enough people agree that a "Helpful" note isn't sufficiently helpful.

**-0** **×2**  
**Ratings of Helpful on notes that ended up with a status of Not Helpful**  
These ratings are counted twice because they often indicate support for notes that others deemed low-quality.

×

**6**  
**Writing impact**  
Last update: 51 minutes ago

**+6**  
**Your notes that earned the status of Helpful**  
Well done! These notes are now showing to everyone who sees the post, adding context and helping keep people informed.

**-0**  
**Your notes that reached the status of Not Helpful**  
These notes have been rated Not Helpful by enough contributors, including those who sometimes disagree in their past ratings. You can see these notes and the feedback they've received on your profile.  
Note statuses can change as more people rate them.

**41**  
**Notes that need more ratings**  
Notes that don't yet have a status of Helpful or Not Helpful.

i **Did you know?**  
Note statuses aren't reached by majority rule. To identify notes that are helpful to a wide range of people, note statuses require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings.  
[Learn More](#)



1



# Will there be at least one "Currently Rated Helpful" Community Note on Biden's X.com post "inflation was 0% last month"?

#X (Twitter)



Jay Baxter

Play

10



1.1k

\$0.00 earned

resolved Dec 4

Resolved **YES**



1D

1W


1M

ALL




# Community Notes + Prediction Markets?

🔄 You reposted

 **Josh Stark** ✓  
@0xstark

biggest devcon ever

 **Lazar** 🇹🇼 ✓ @0xlazar · Oct 25

Exciting stats about @EFDevcon I just learned about during today's sync with other volunteer leads:

- 12,000+ people are expected to show up
- 900+ volunteers signed up...

[Show more](#)

94% chance that, according to consensus of Community Notes users, the 2024 Bangkok Devcon will be the most-attended Devcon ever.

Buy YES Buy NO

<https://vitalik.eth.limo/general/2024/11/09/infofinance.html>

# Things I am personally interested in

- Mechanisms, e.g. prediction market mashups, multidimensional bridging
- More notes, faster, more places
  - E.g. matching, LLMs
- Bridging algorithms beyond notes

# Thank you!

- We are hiring an ML engineer!
  - Also linked on my bio on X
- DM me on X: @\_jaybaxter\_