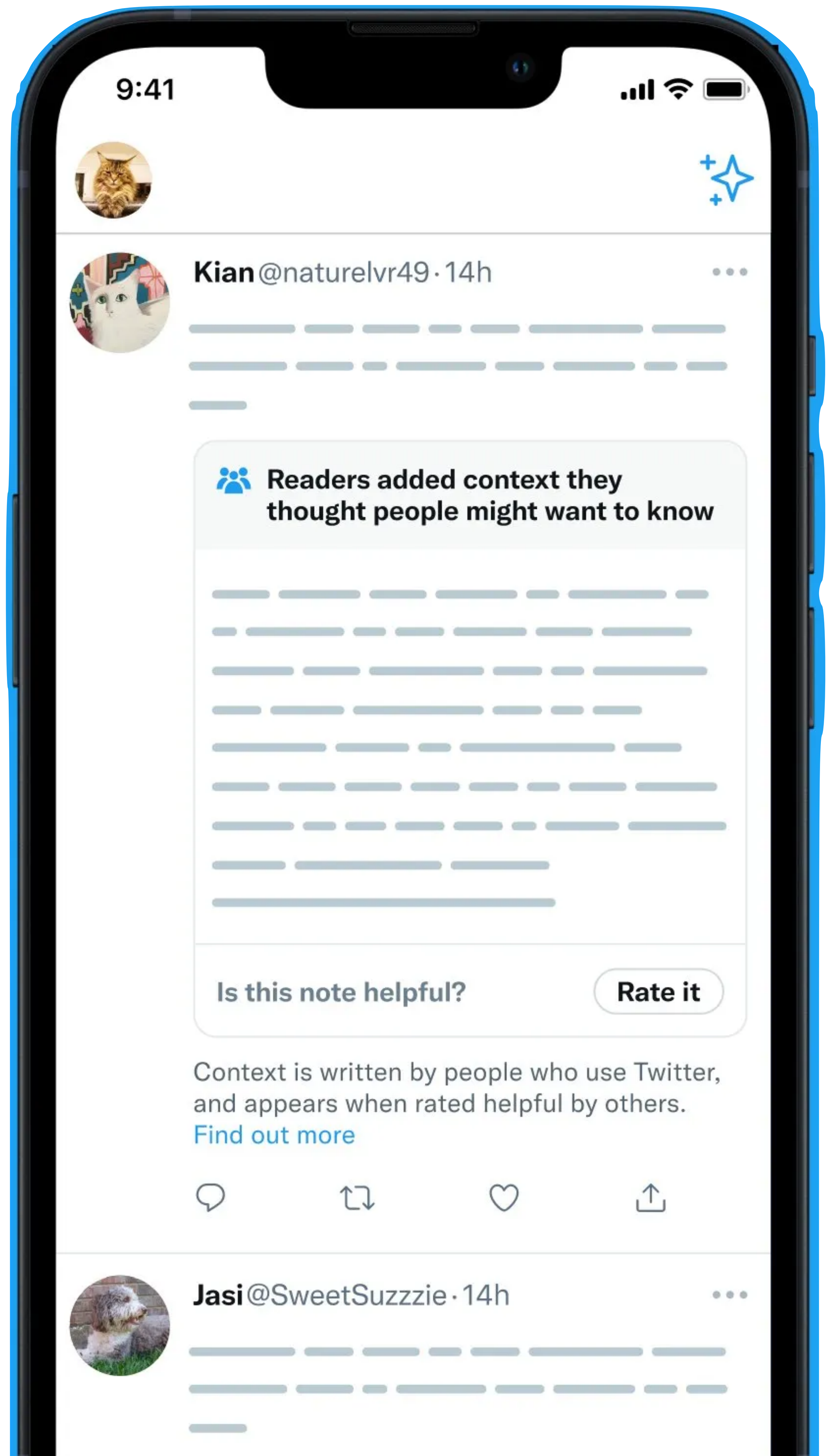
The *Supernote*

Enhancing Crowd-sourced factchecking using LLM-driven consensus

Soham De, Michiel Bakker, Martin Saveski

IC²S² 2024

Community Notes



9:41

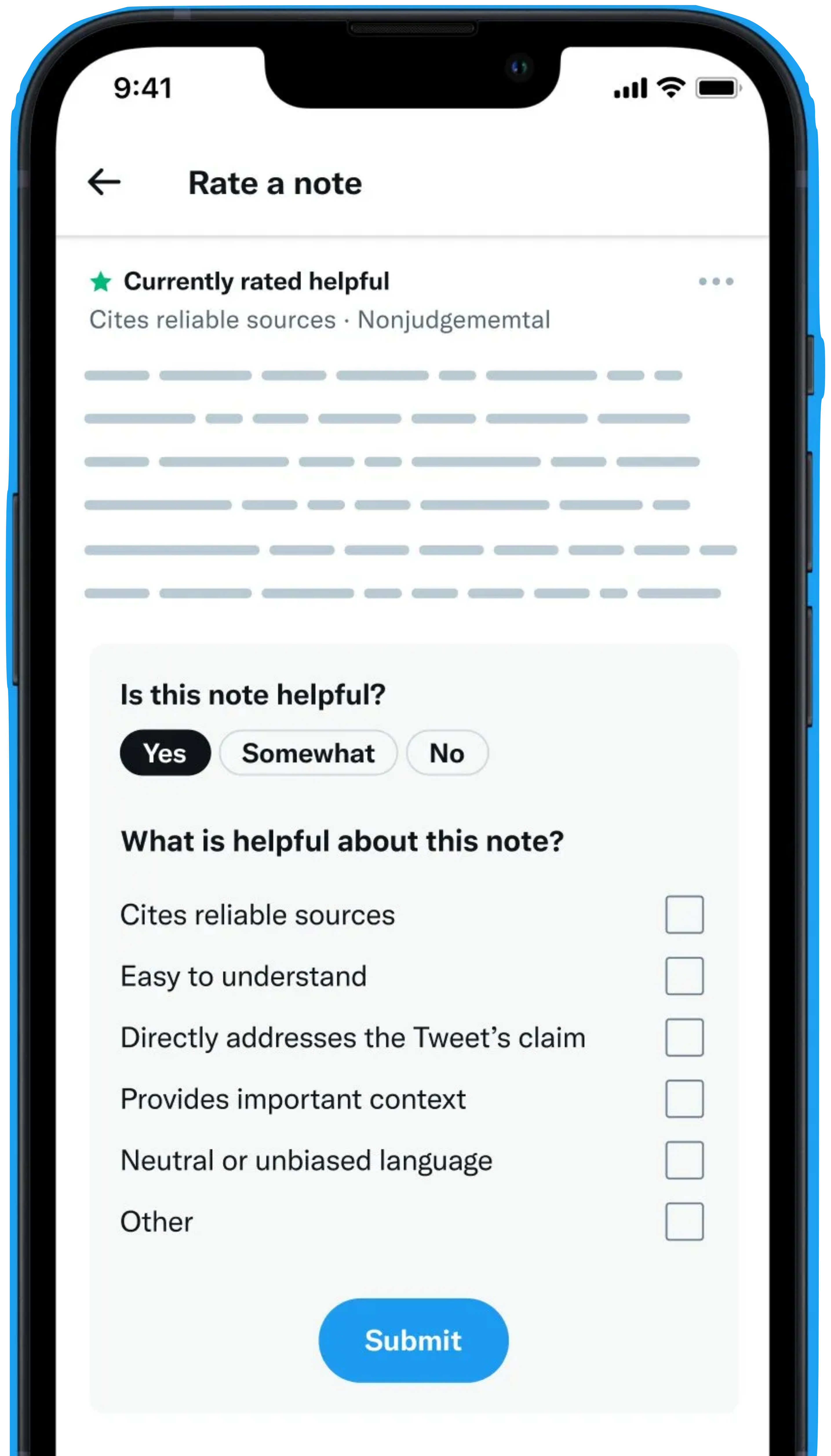
Kian @naturelvr49 · 14h

Readers added context they thought people might want to know

Is this note helpful? Rate it

Context is written by people who use Twitter, and appears when rated helpful by others. Find out more

Jasi @SweetSuzzzie · 14h



9:41

Rate a note

★ **Currently rated helpful**

Cites reliable sources · Nonjudgememtal

Is this note helpful?

[✓ Yes] Somewhat No

What is helpful about this note?

Cites reliable sources □

Easy to understand □

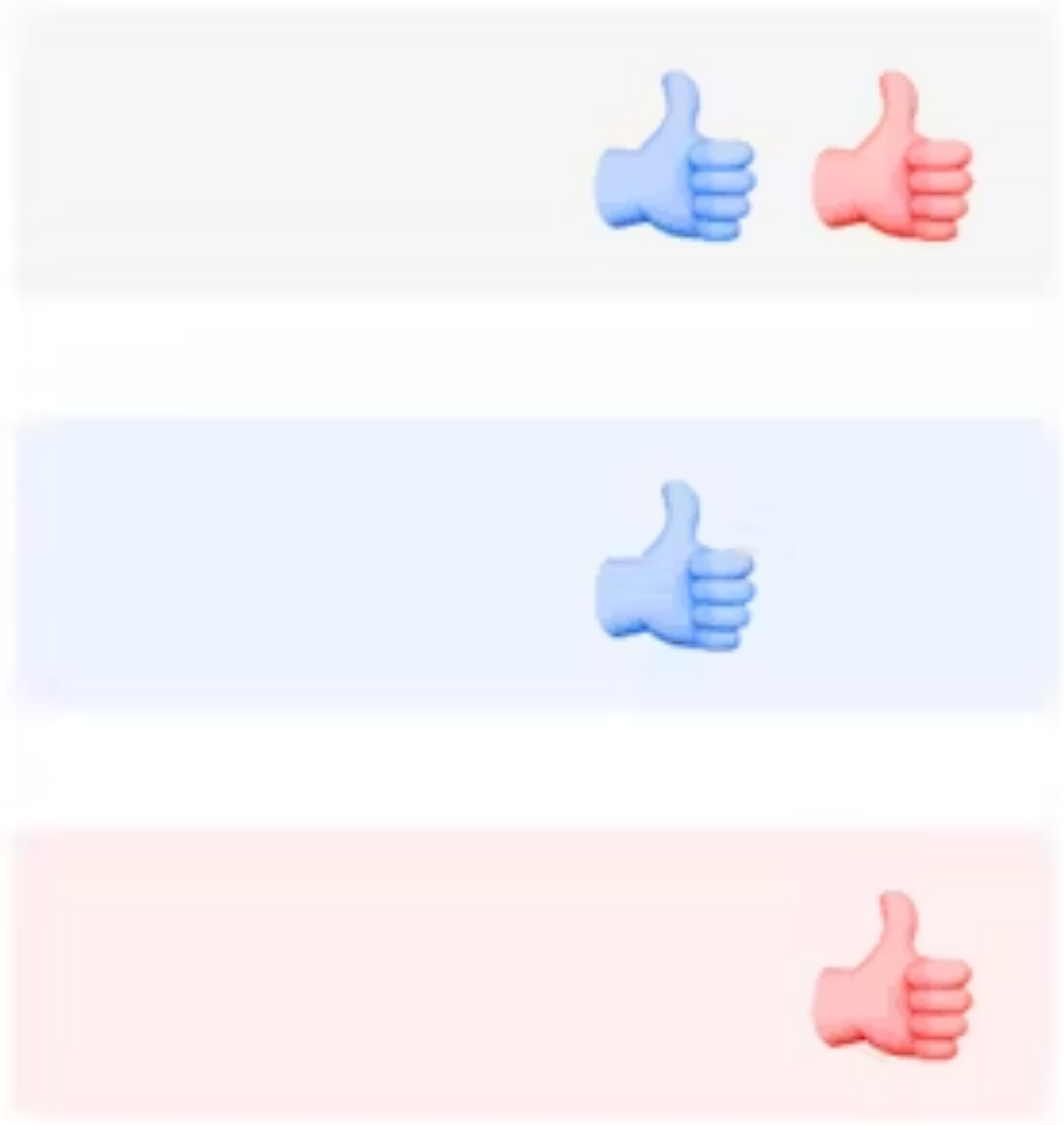
Directly addresses the Tweet’s claim □

Provides important context □

Neutral or unbiased language □

Other □

Submit



Poynter Institute on

Why Twitter’s Community Notes feature mostly fails to combat Misinformation

“It requires a cross-ideological agreement on truth, and … achieving that consensus is almost impossible”

Poynter Institute on
Why Twitter's Community Notes feature mostly fails to combat Misinformation

"It requires a cross-ideological agreement on truth, and … achieving that consensus is almost impossible"

Notes aren't rated helpful fast enough to reduce engagement with misinformation in the early stages of diffusion
(Chuai et al., 2023)

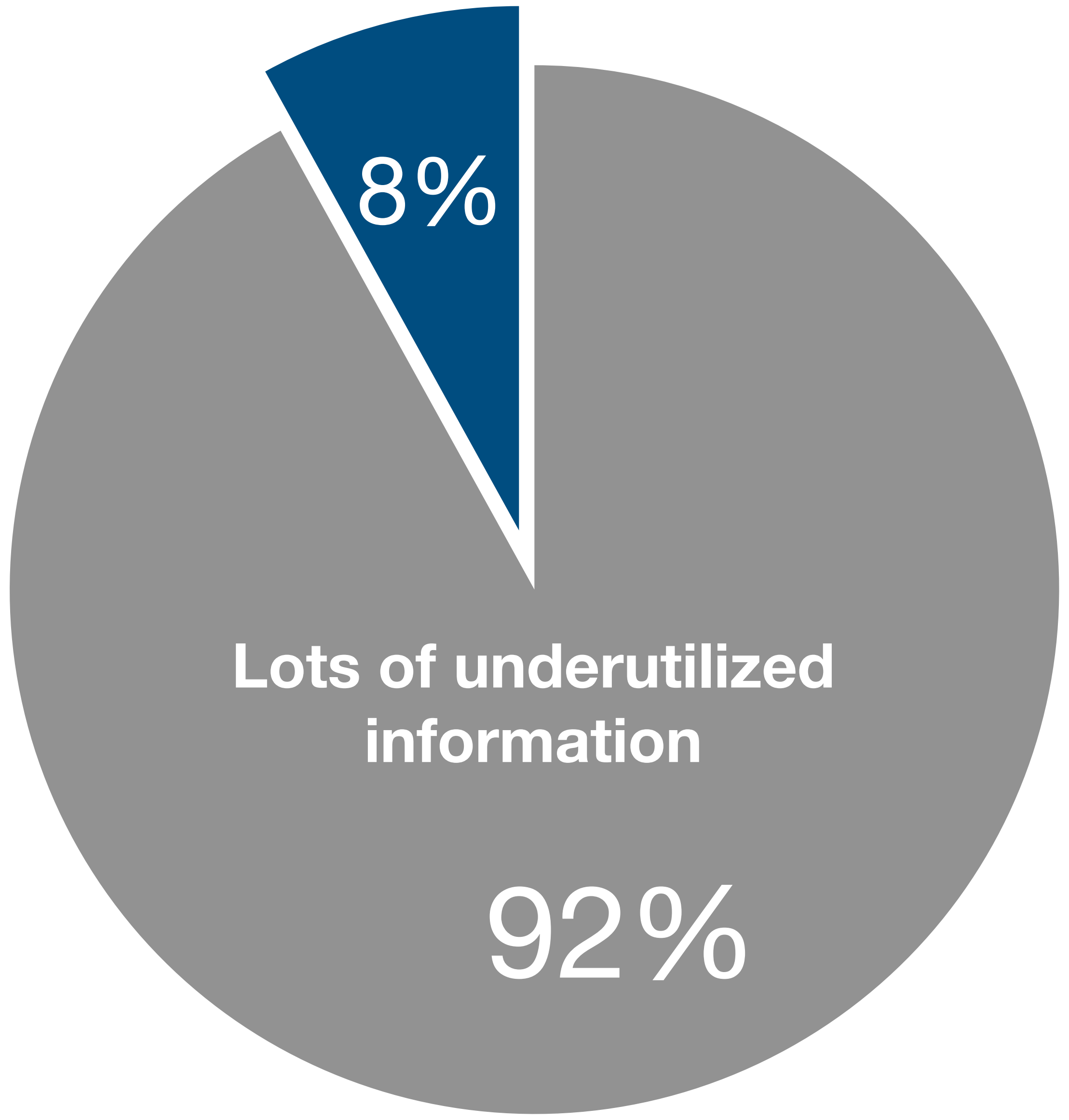
Poynter Institute on

Why Twitter’s Community Notes feature mostly fails to combat Misinformation

“It requires a cross-ideological agreement on truth, and … achieving that consensus is almost impossible”

Notes aren’t rated helpful fast enough to reduce engagement with misinformation in the early stages of diffusion

(Chuai et al., 2023)



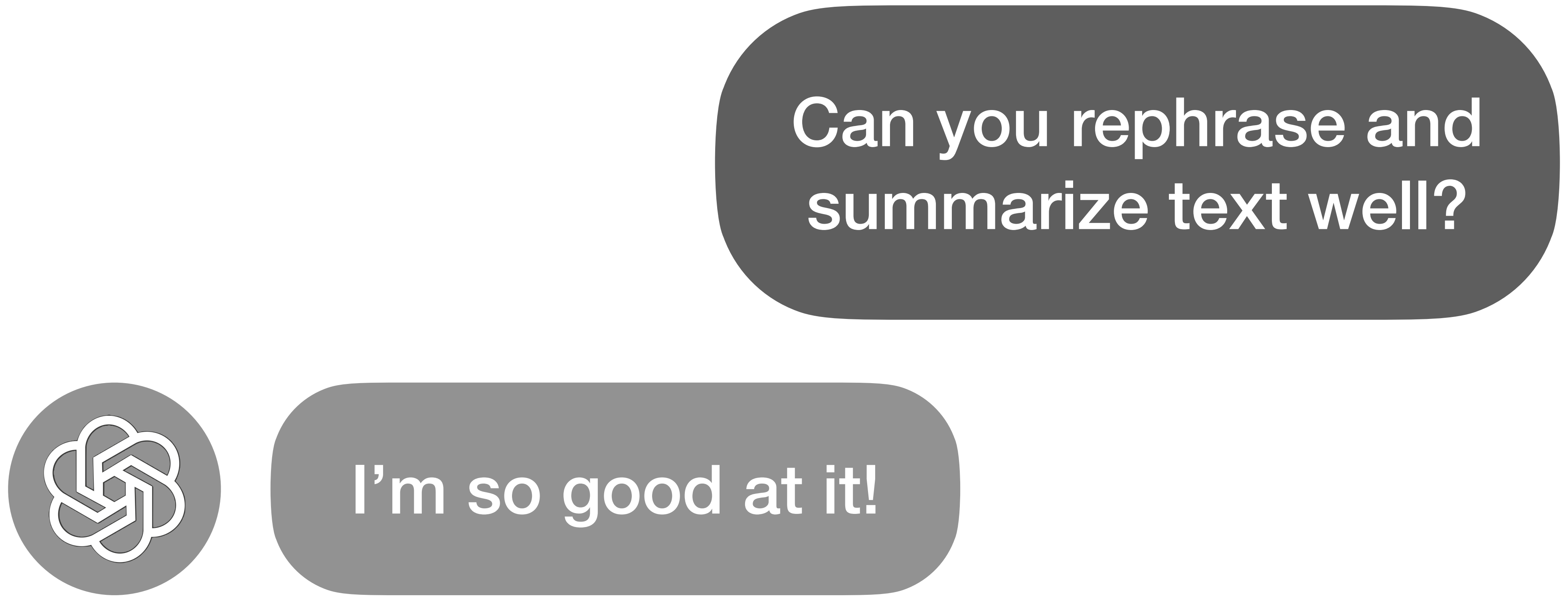
as a consequence,

More than 90% of notes written are not shown on X.

Poynter Institute on

Why Twitter's Community Notes feature mostly fails to combat Misinformation

"It requires a cross-ideological agreement on truth, and … achieving that consensus is almost impossible"



Can you rephrase and summarize text well?

I'm so good at it!

LLMs can construct very good consensus statements to find agreement

(Bakker et al, 2022)

Goals

1. **Improve fact-checking note-quality**
   - Use information in existing notes
     (useful information may be spread across several notes)
   - Ensure adherence to principles of good fact-checking[1]

[1] Fact-checking principles from X: https://communitynotes.x.com/guide/en/contributing/examples

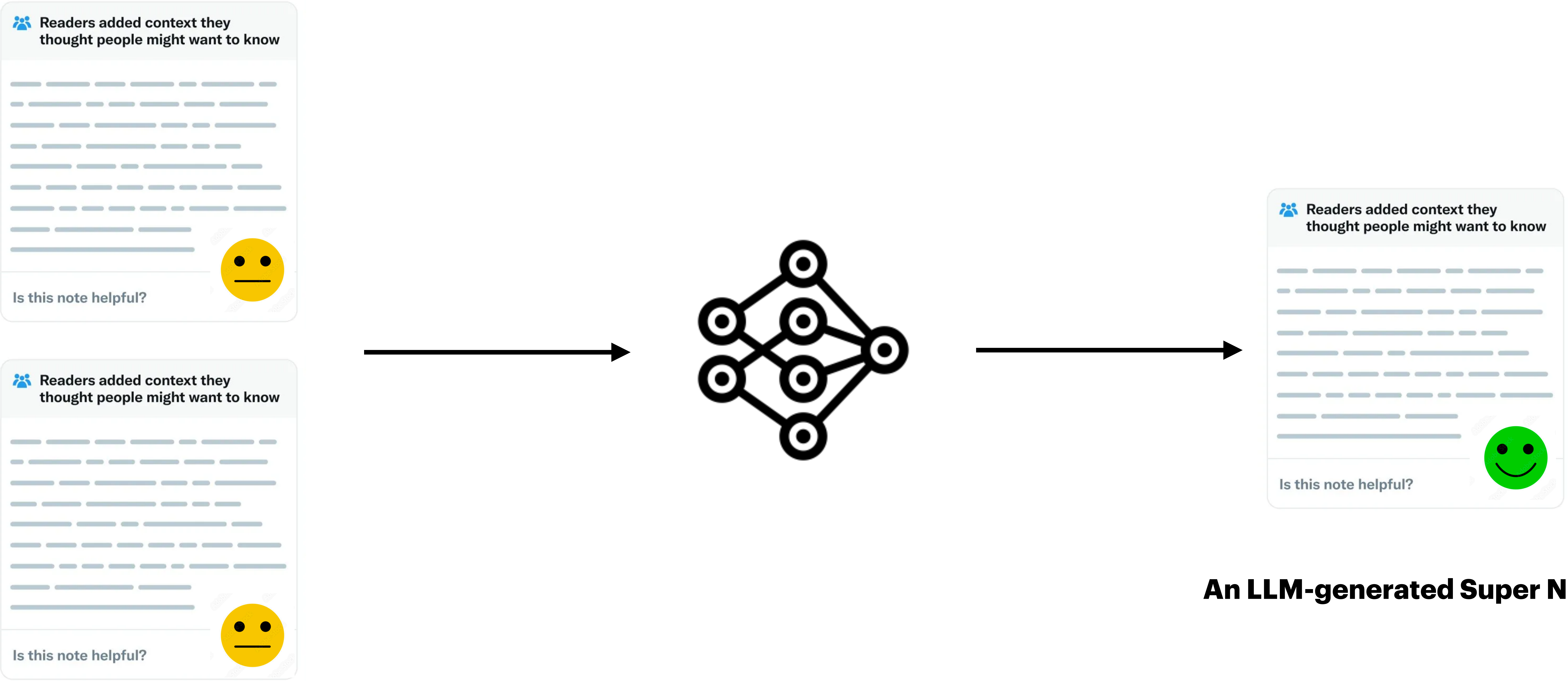
Goals

1. **Improve fact-checking note-quality**
    - Use information in existing notes
      (useful information may be spread across several notes)
    - Ensure adherence to principles of good fact-checking

2. **Scale up crowd-sourced fact-checking**
    - Use LLMs to summarize existing fact-checks
    - Produce notes more likely to draw cross-ideological agreement

Proposed Solution

A Supernote



Readers added context they thought people might want to know

Is this note helpful?

Readers added context they thought people might want to know

Is this note helpful?

Readers added context they thought people might want to know

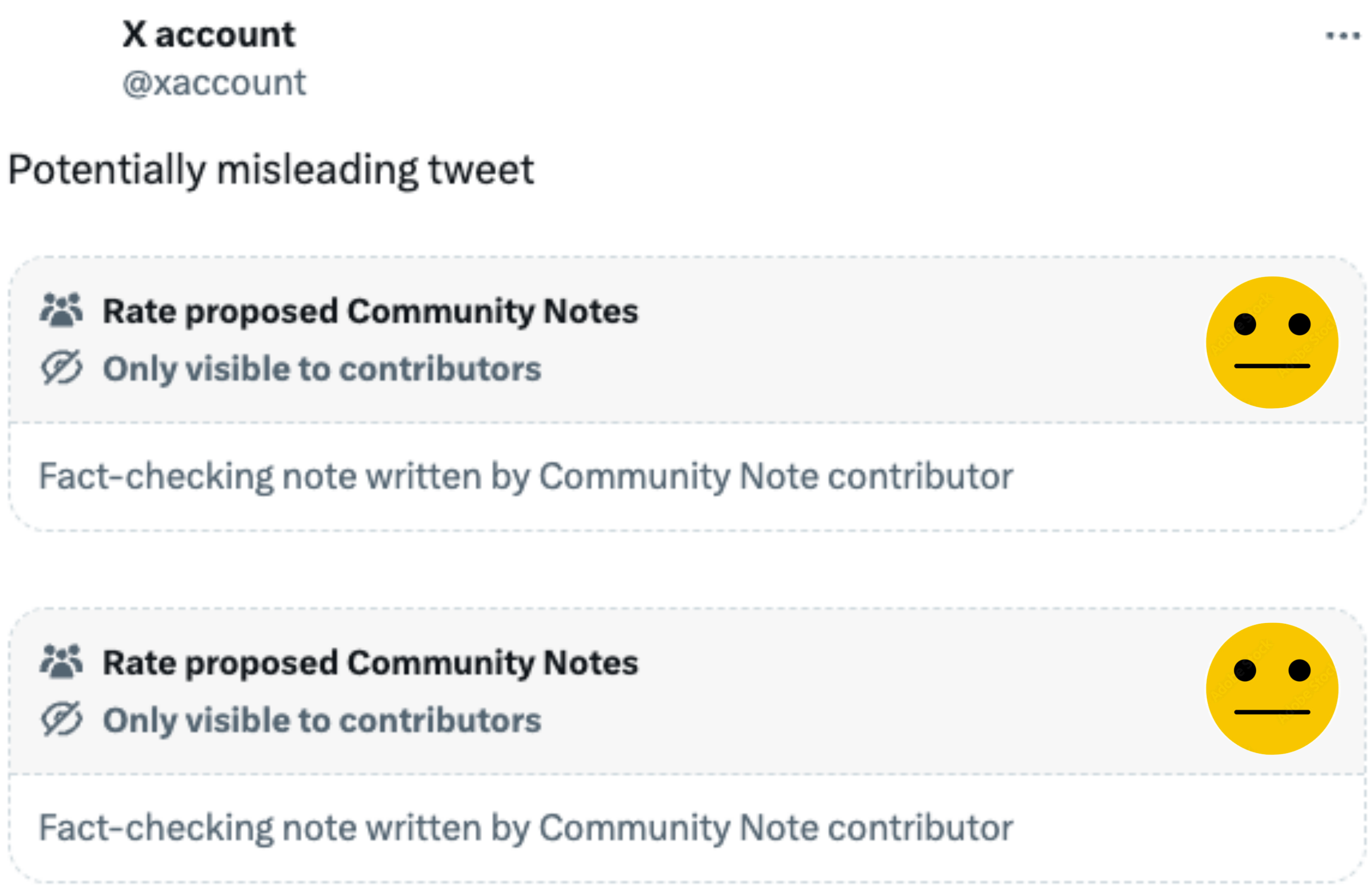
Is this note helpful?

Take notes that **aren’t rated Helpful yet**

An LLM-generated Super Note

Proposed Solution

In practice



X account
@xaccount

Potentially misleading tweet

Rate proposed Community Notes
Only visible to contributors

Fact-checking note written by Community Note contributor

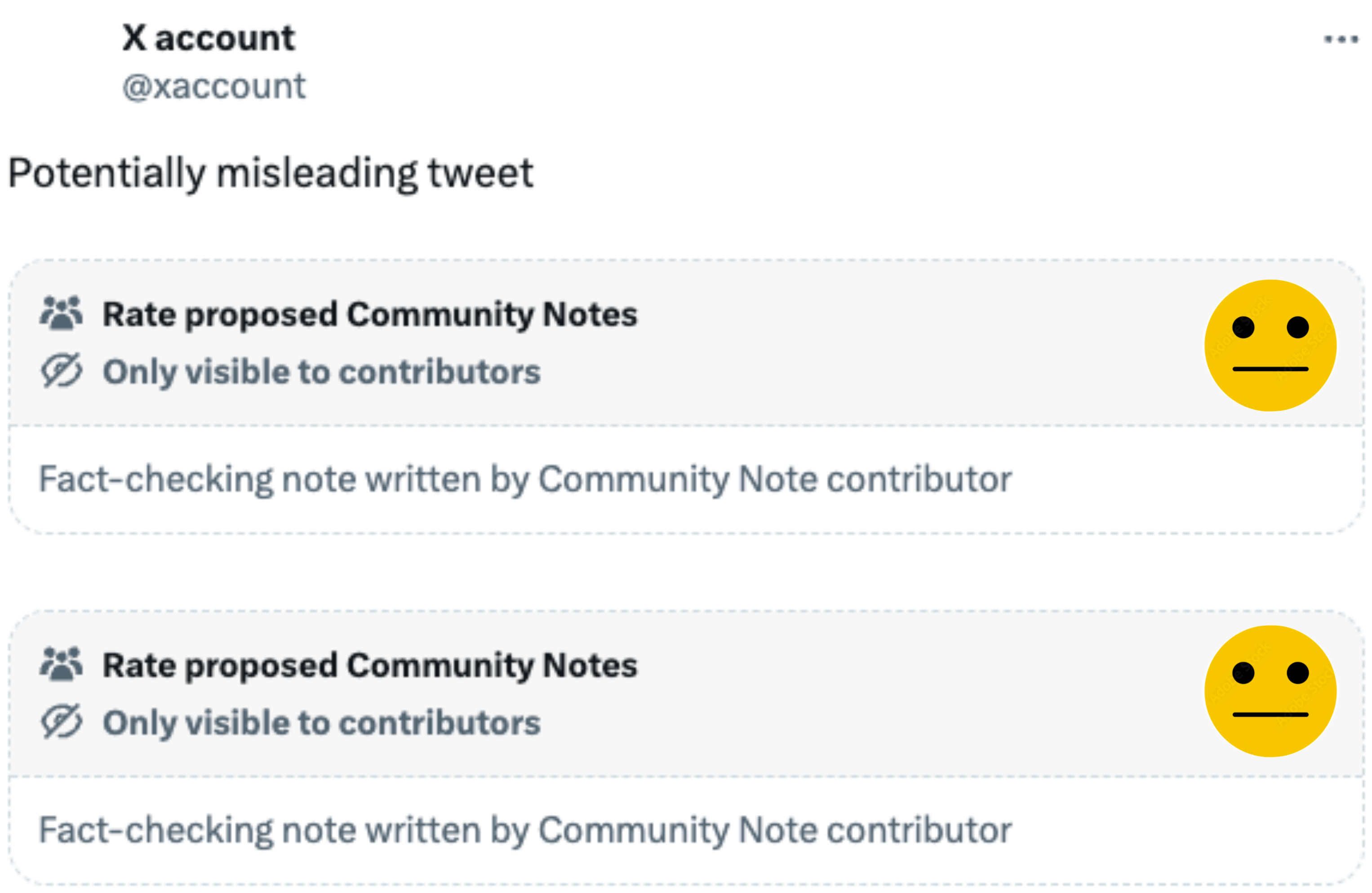
Rate proposed Community Notes
Only visible to contributors

Fact-checking note written by Community Note contributor

This tweet has a few notes on it, but **none of them are currently rated helpful**

Proposed Solution

In practice



X account
@xaccount

Potentially misleading tweet

Rate proposed Community Notes 😐
Only visible to contributors

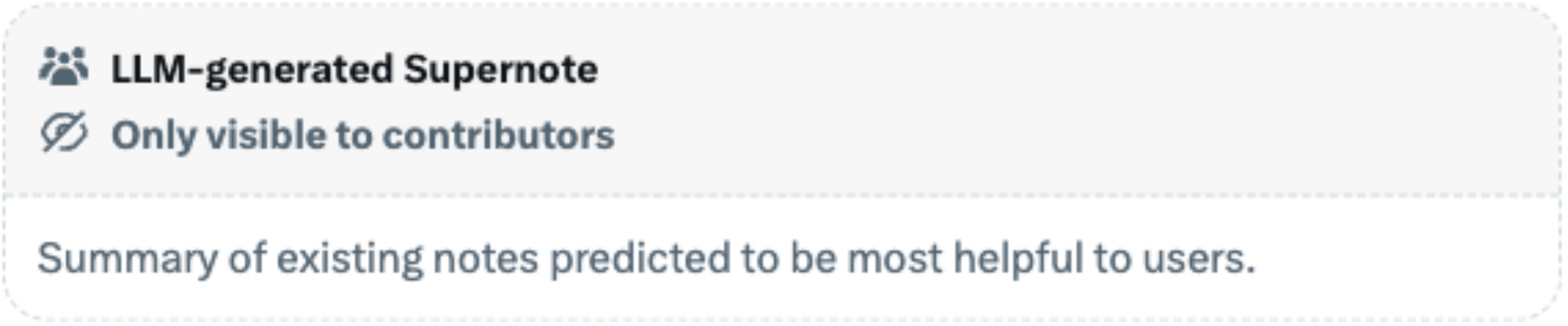
Fact-checking note written by Community Note contributor

Rate proposed Community Notes 😐
Only visible to contributors

Fact-checking note written by Community Note contributor

This tweet has a few notes on it, but **none of them are currently rated helpful**

We attempt to generate a **Supernote** - the summary of existing notes predicted to be most helpful to users.

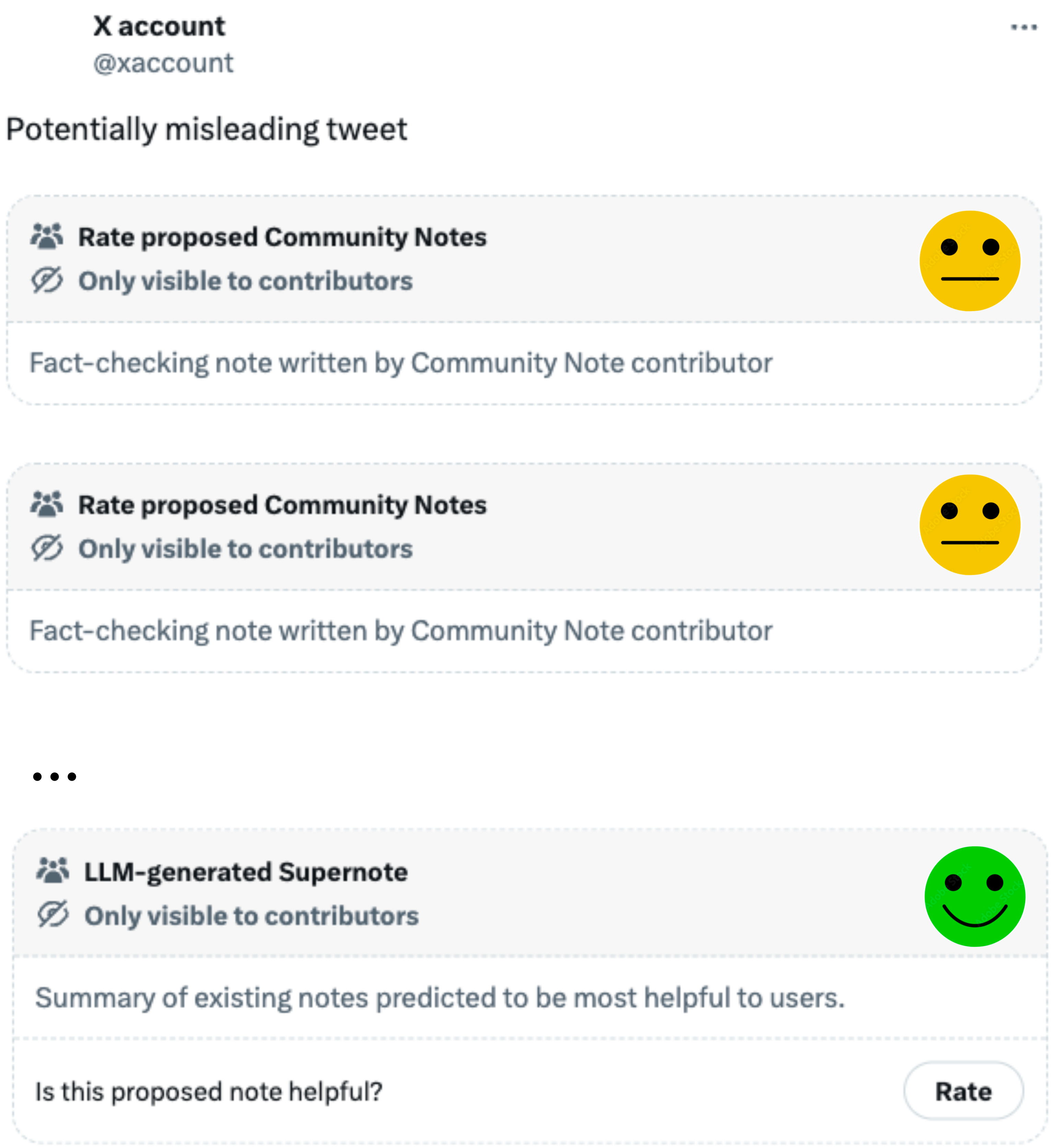


LLM-generated Supernote
Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

Proposed Solution

In practice



X account
@xaccount

Potentially misleading tweet

Rate proposed Community Notes 😐
Only visible to contributors

Fact-checking note written by Community Note contributor

Rate proposed Community Notes 😐
Only visible to contributors

Fact-checking note written by Community Note contributor

...

LLM-generated Supernote 🙂
Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

Is this proposed note helpful? **Rate**

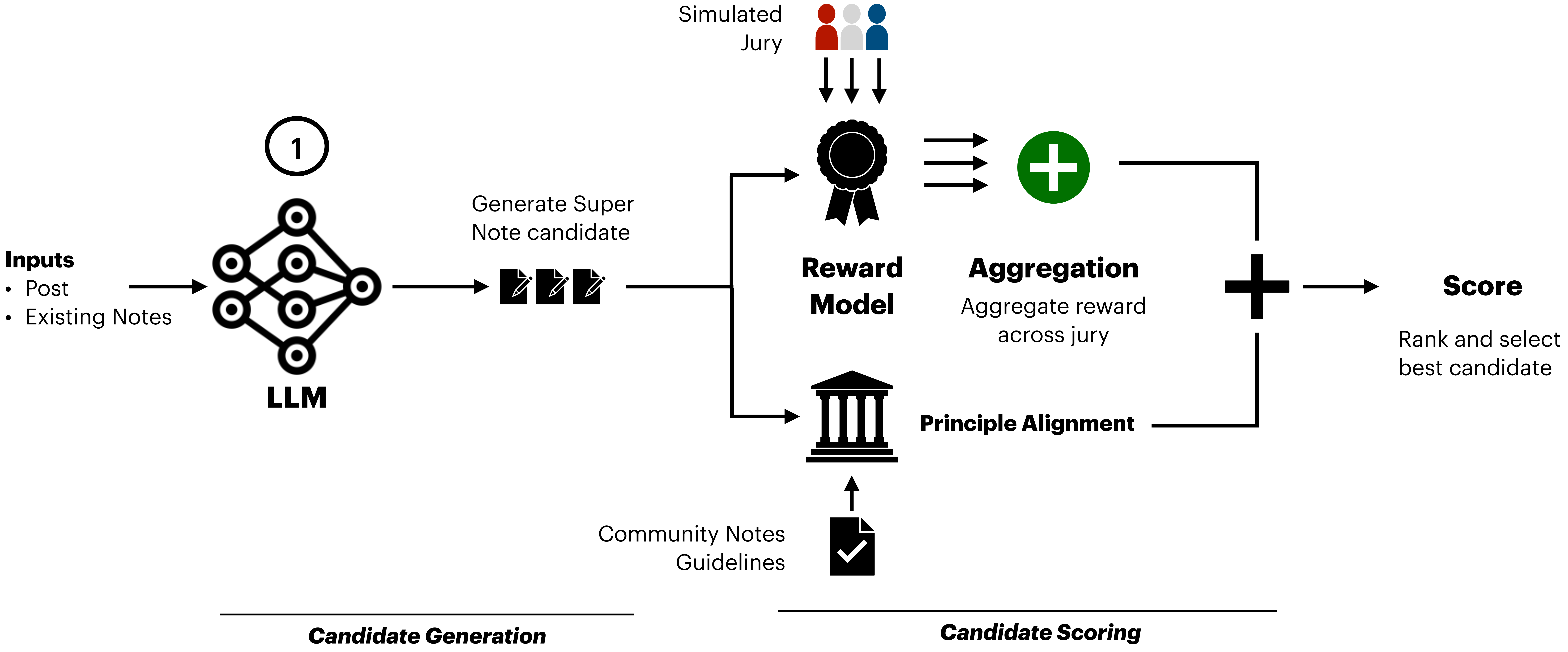
This tweet has a few notes on it, but **none of them are currently rated helpful**

We attempt to generate a **Supernote** - the summary of existing notes predicted to be most helpful to users.

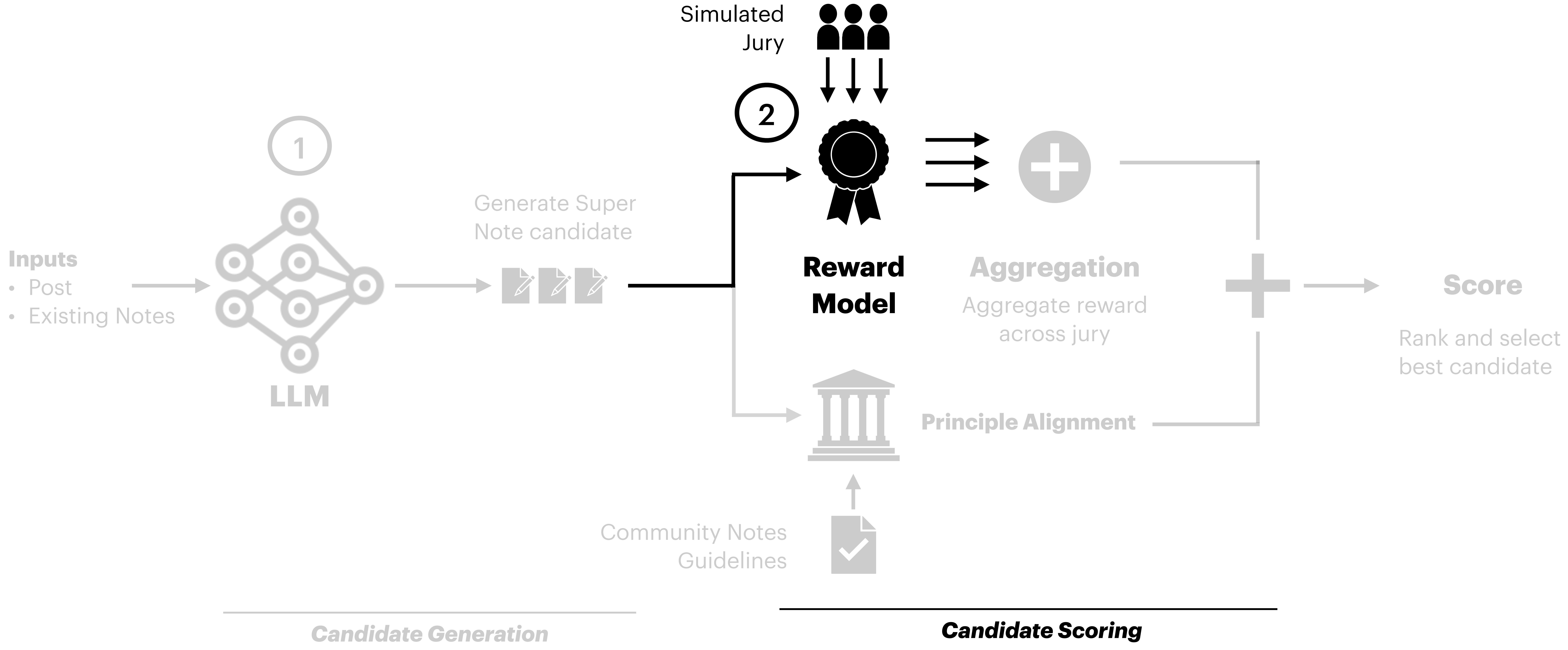
If the Supernote is expected to be **more helpful than any existing note**, it is shown on the platform alongside other candidate notes.

So, how to make a Supernote?

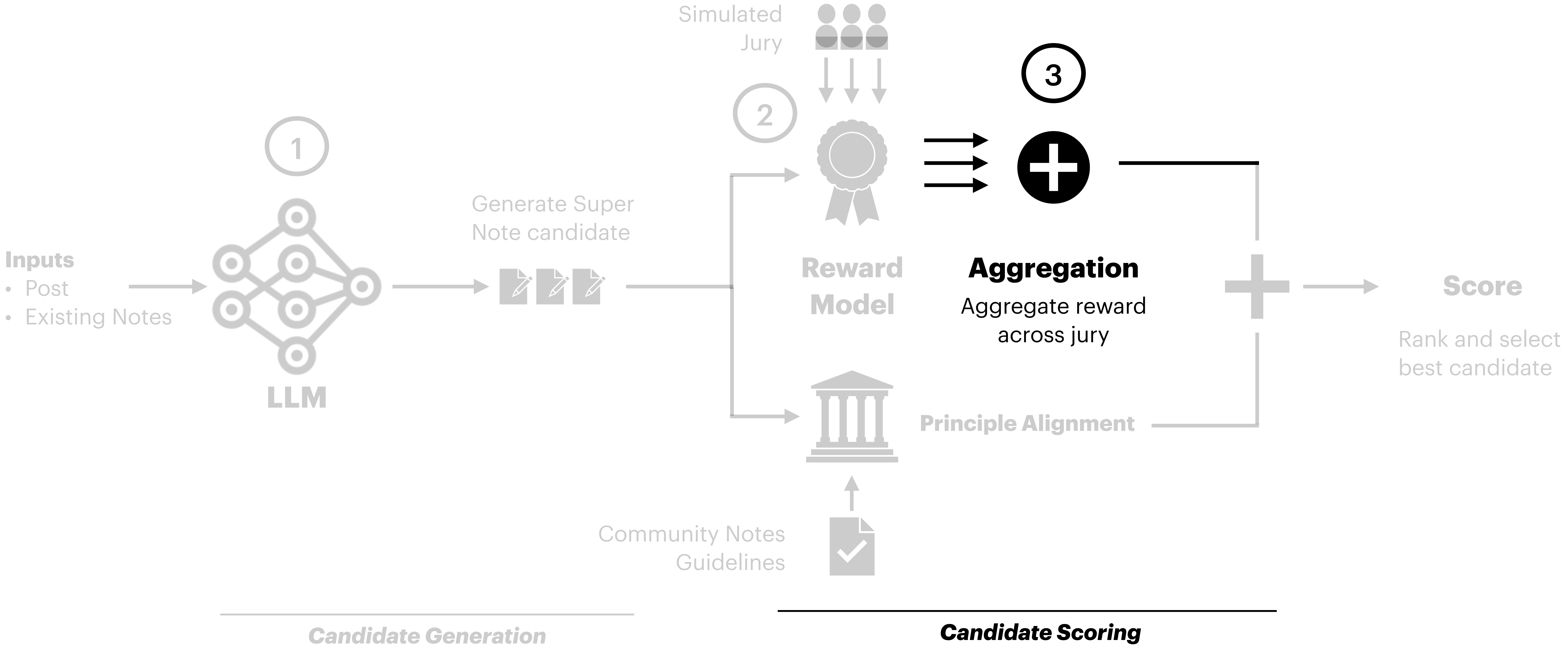
Supernote Generation Pipeline



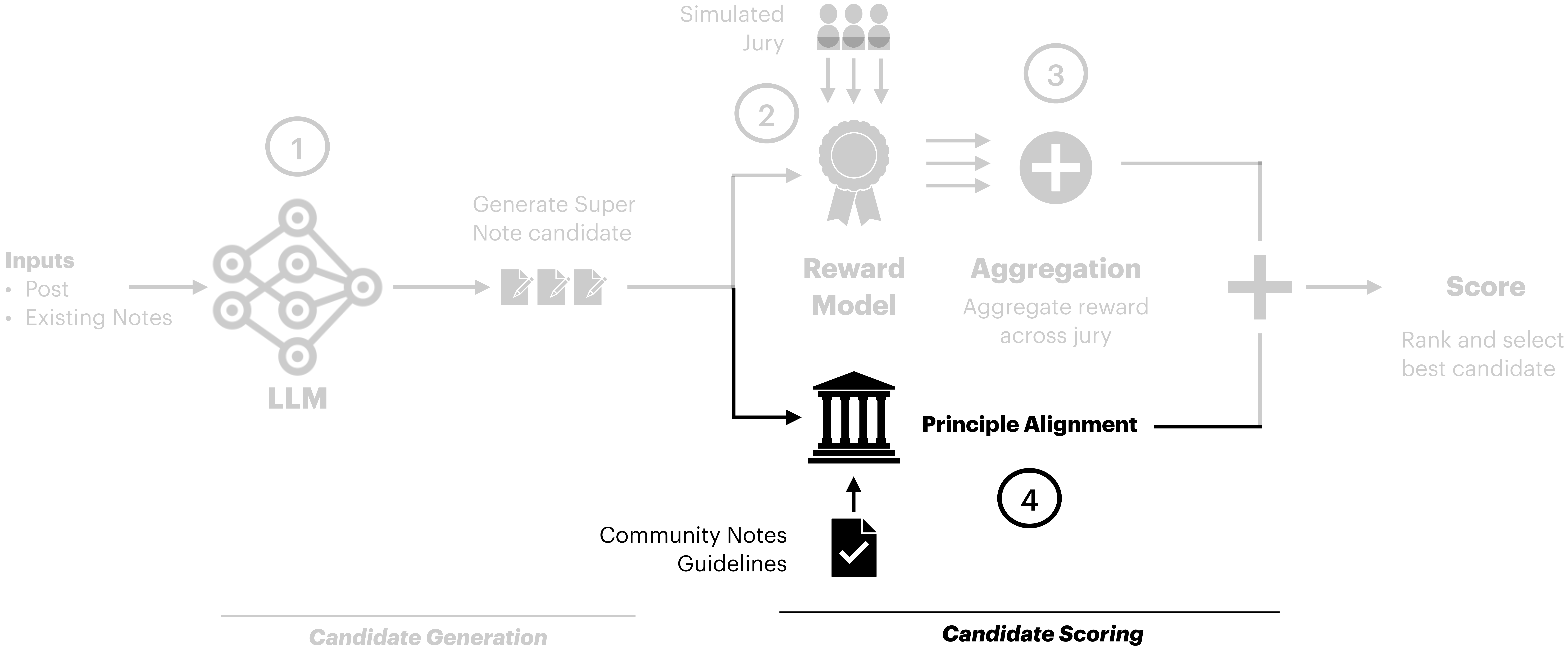
Supernote Generation Pipeline



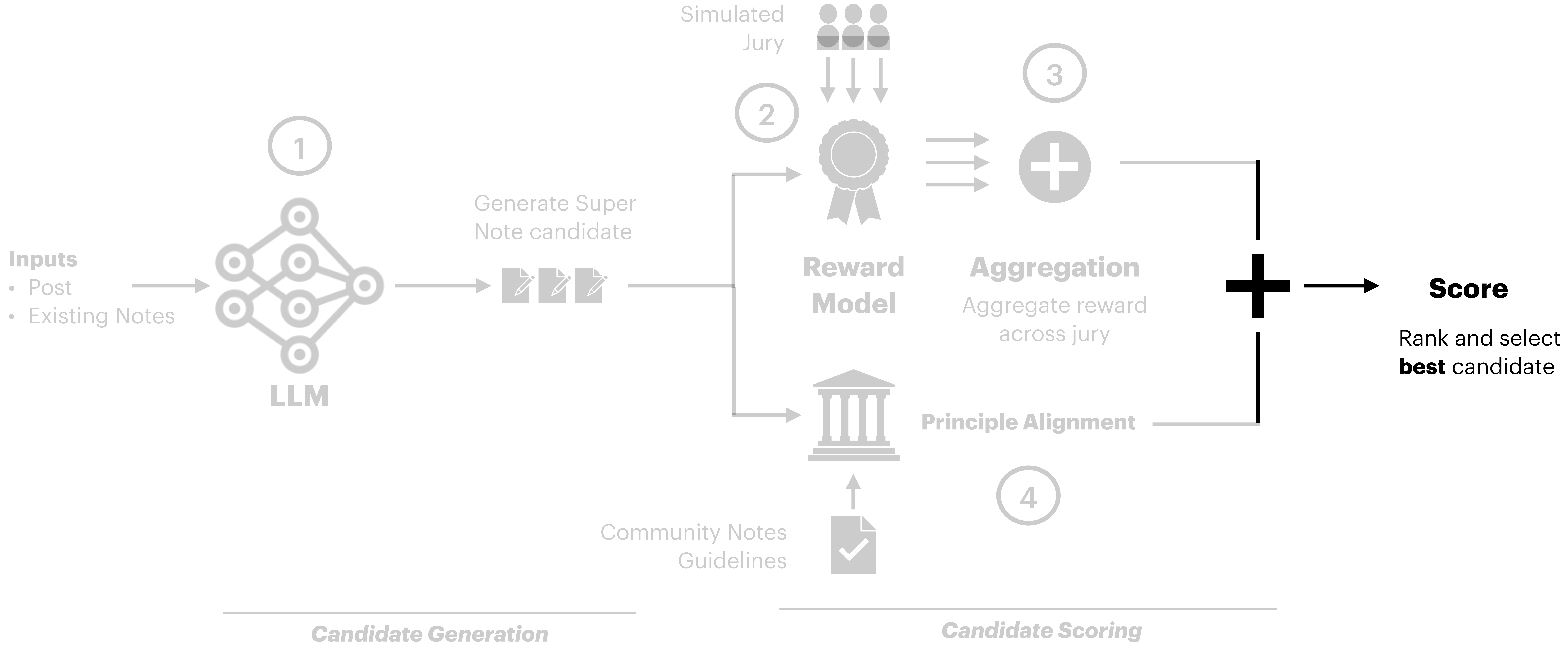
Supernote Generation Pipeline



Supernote Generation Pipeline

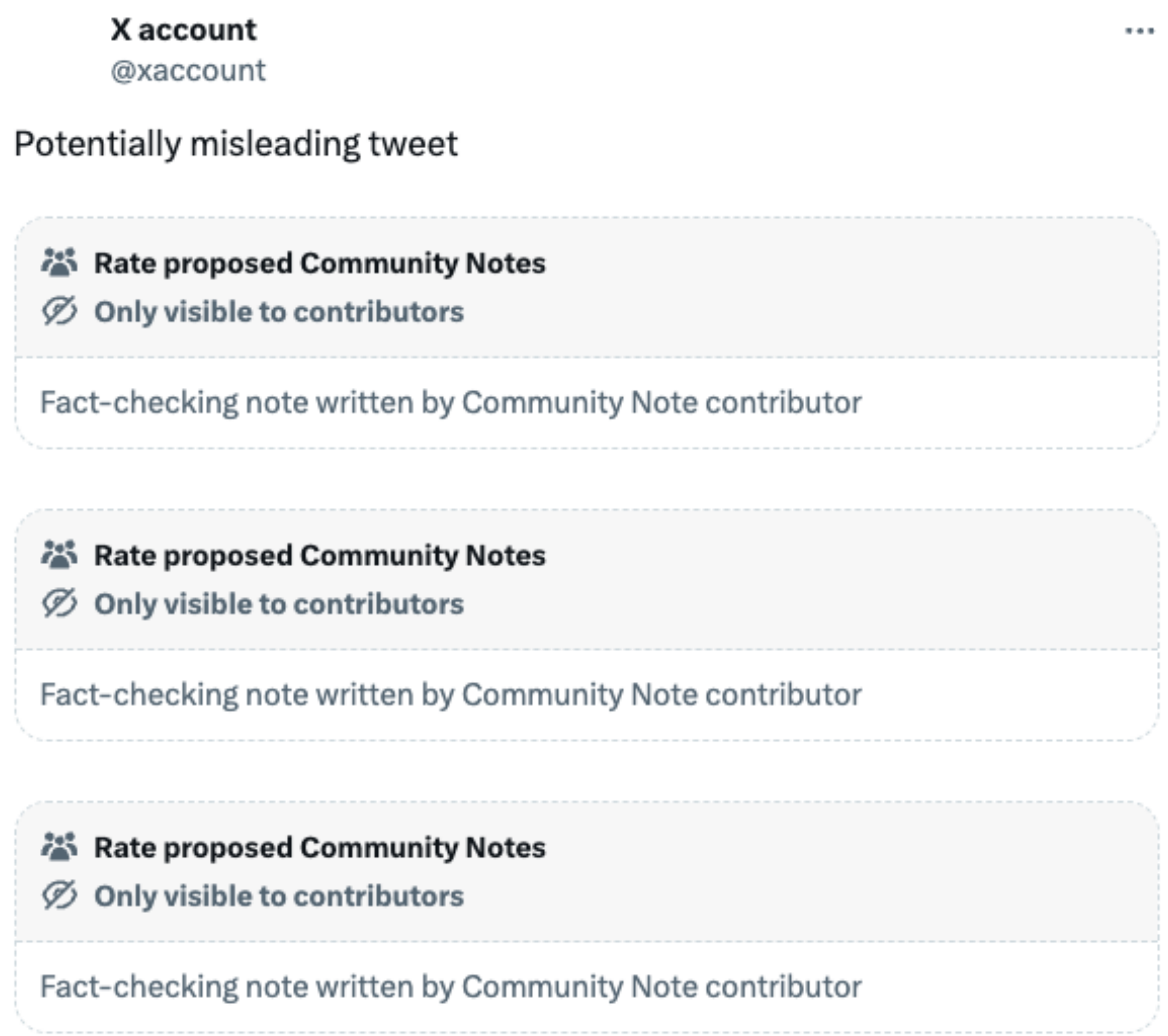


Supernote Generation Pipeline



1 LLM Summarization

Generates 100 candidate supernotes by summarizing existing notes



X account
@xaccount

Potentially misleading tweet

Rate proposed Community Notes
Only visible to contributors

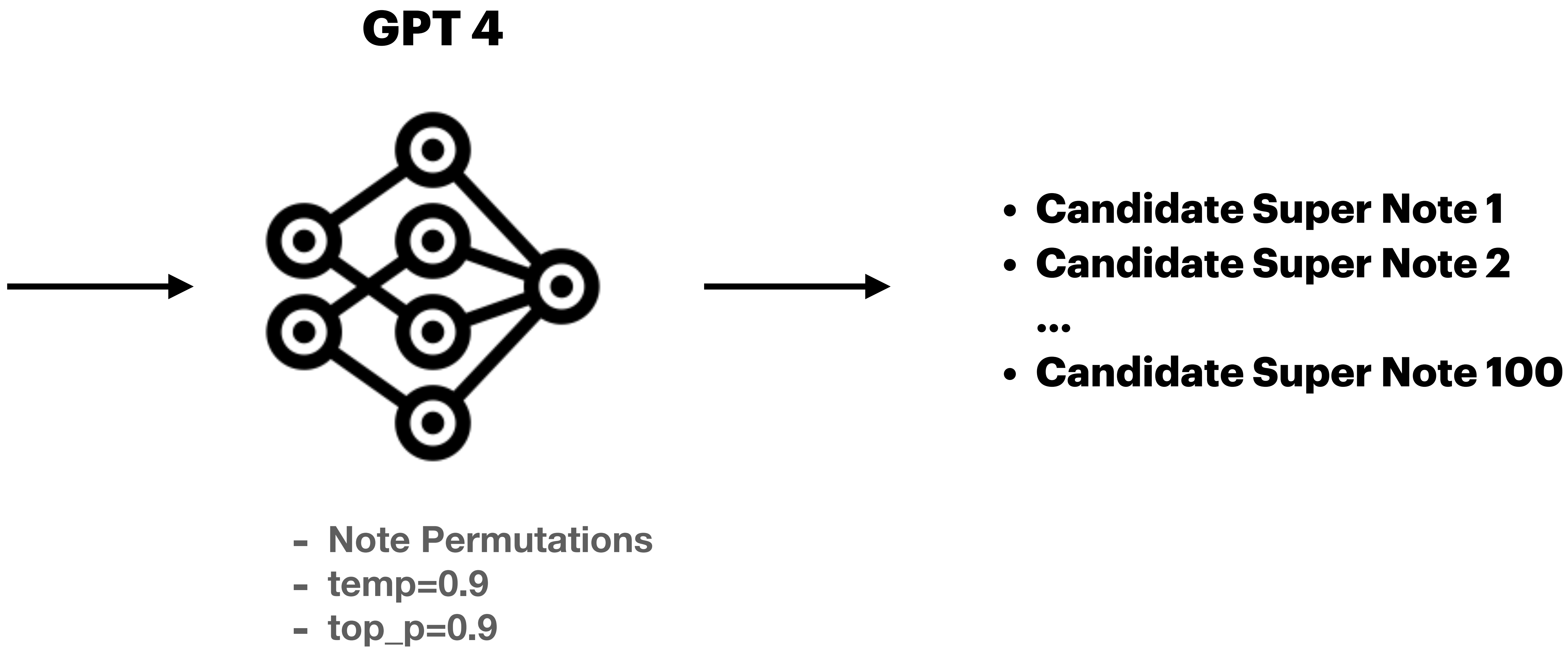
Fact-checking note written by Community Note contributor

Rate proposed Community Notes
Only visible to contributors

Fact-checking note written by Community Note contributor

Rate proposed Community Notes
Only visible to contributors

Fact-checking note written by Community Note contributor



GPT 4

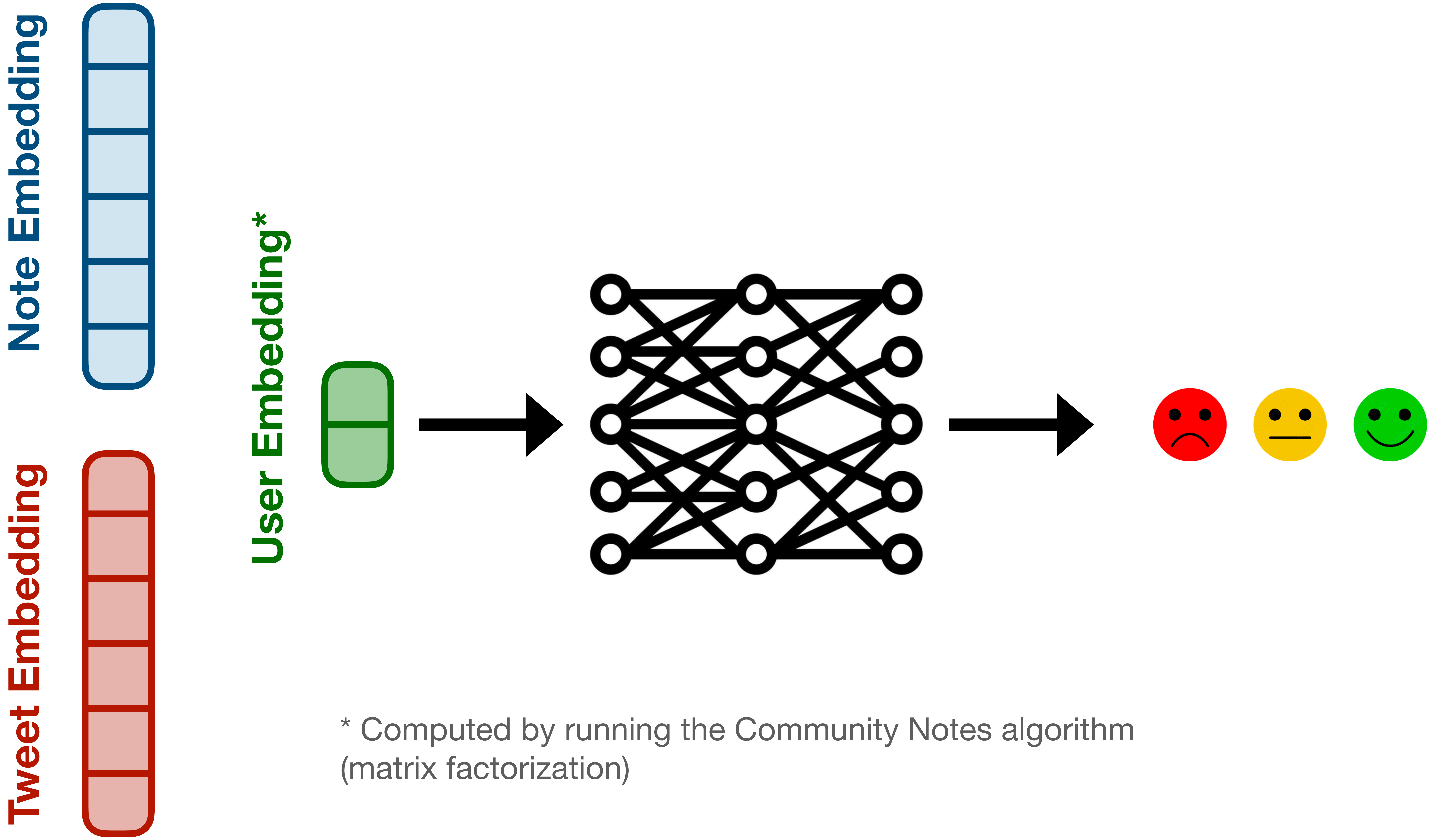
- Note Permutations
- temp=0.9
- top_p=0.9

- **Candidate Super Note 1**
- **Candidate Super Note 2**

  ...
- **Candidate Super Note 100**

2 Reward Model

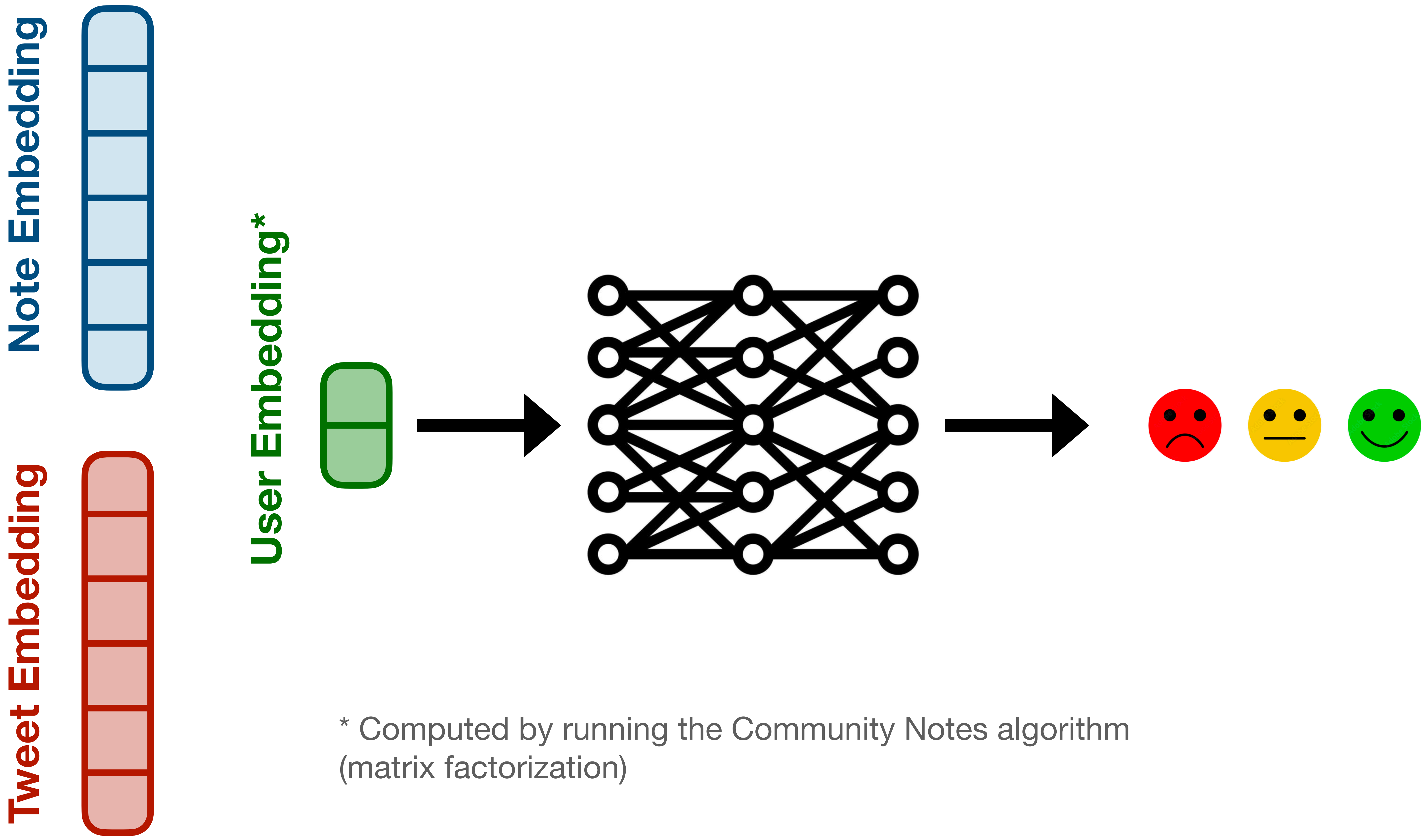
Rater-conditioned helpfulness predictions



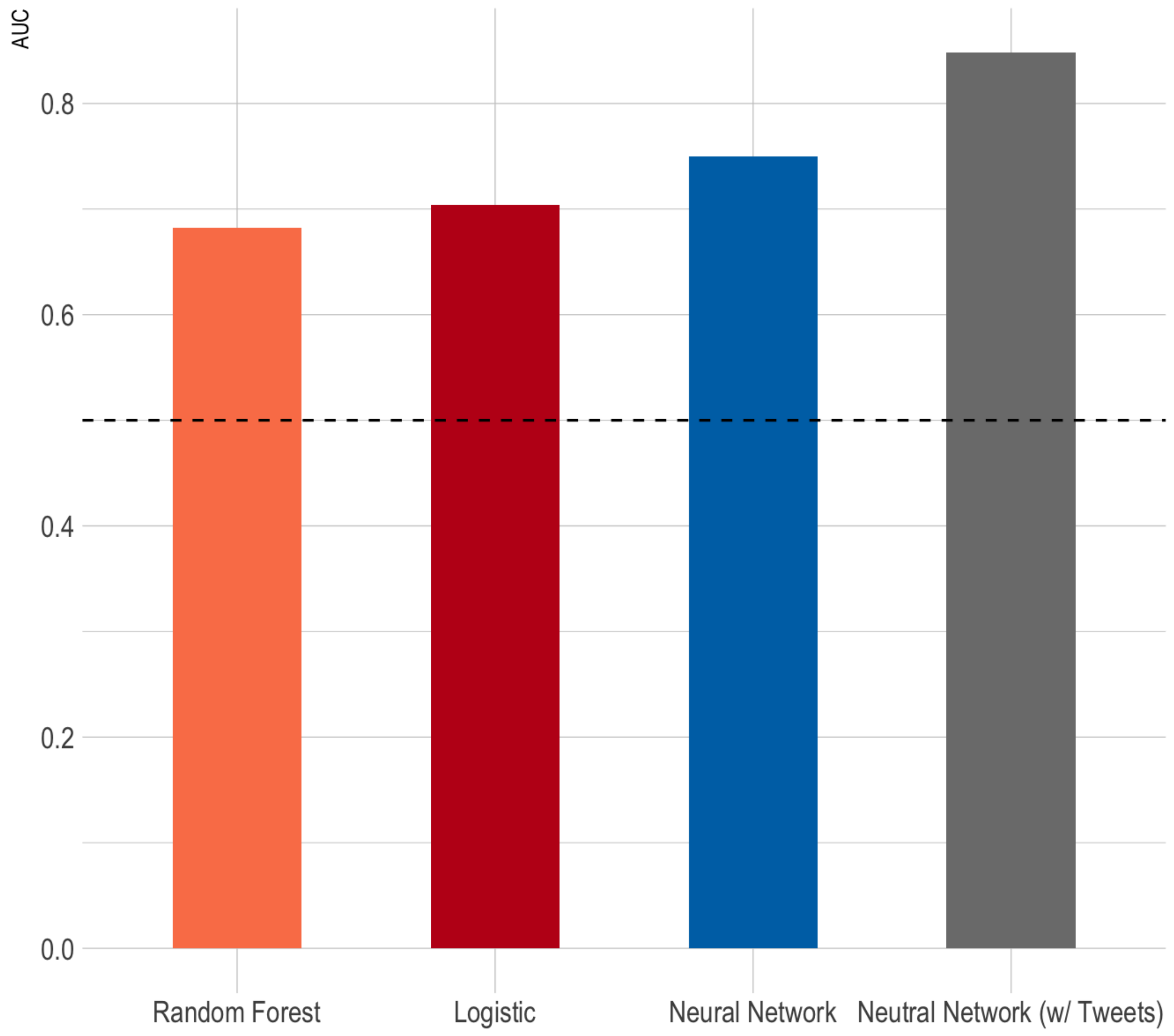
* Computed by running the Community Notes algorithm (matrix factorization)

2 Reward Model

Rater-conditioned helpfulness predictions

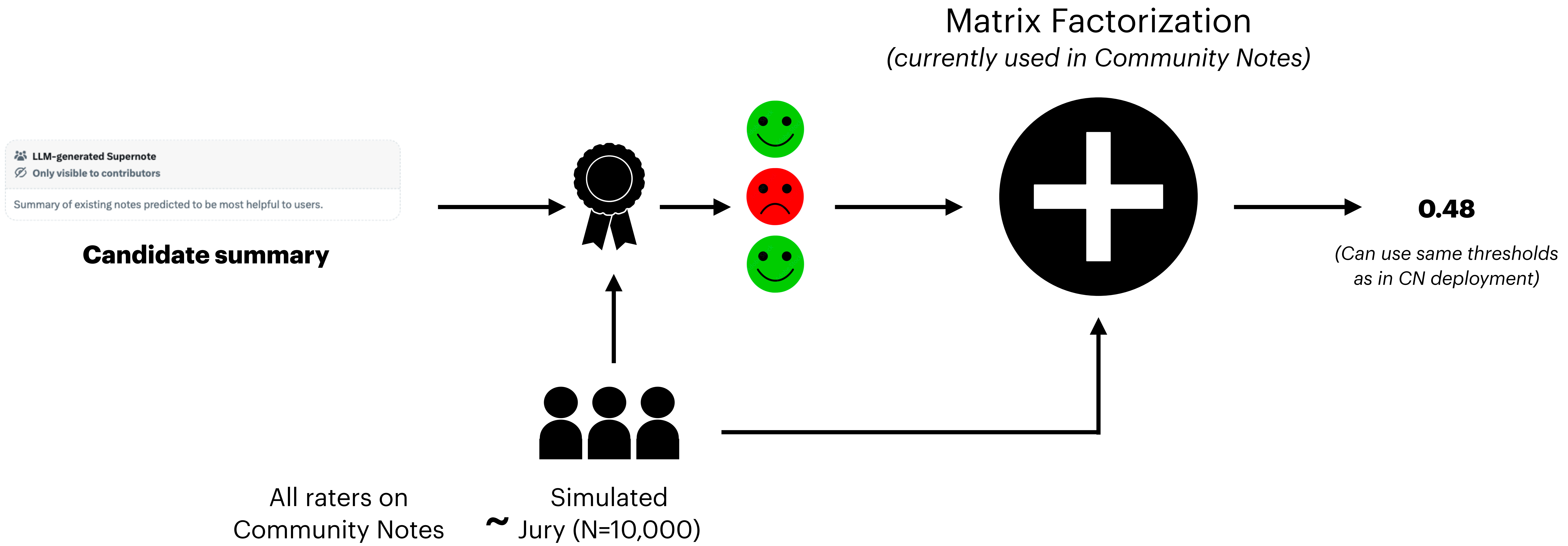


* Computed by running the Community Notes algorithm
(matrix factorization)



3 Aggregating Rewards

Get a single score for a candidate summary



LLM-generated Supernote

Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

Candidate summary

Matrix Factorization

(currently used in Community Notes)

0.48

(Can use same thresholds as in CN deployment)

All raters on Community Notes ~ Simulated Jury (N=10,000)

4 Principle Alignment

Reject candidates that do not follow principles

- Cites high-quality sources
- Easy to understand
- Directly addresses the post’s claim
- Provides important context
- Neutral or unbiased language

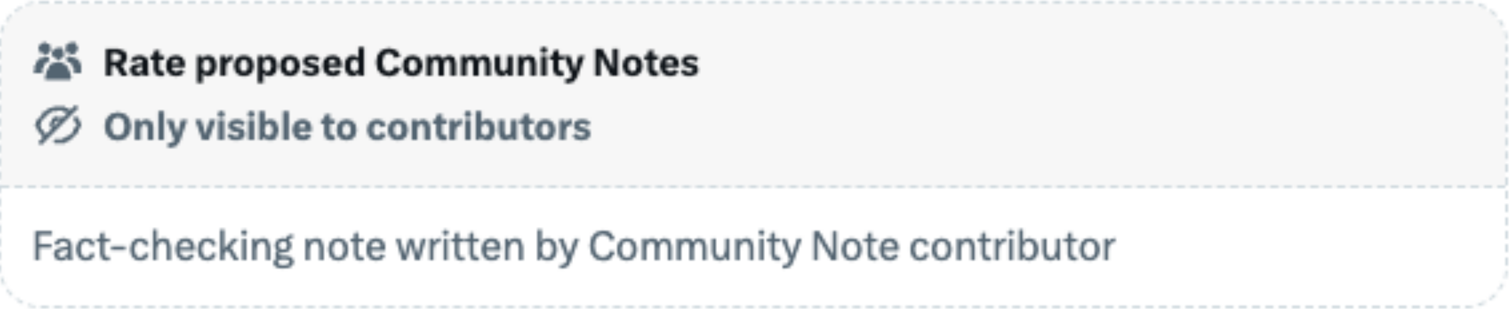
Source: https://communitynotes.x.com/guide/en/contributing/examples

Cool, does it work?

Human Evaluation

We run a pilot (n=15) to evaluate Supernotes against existing notes

A

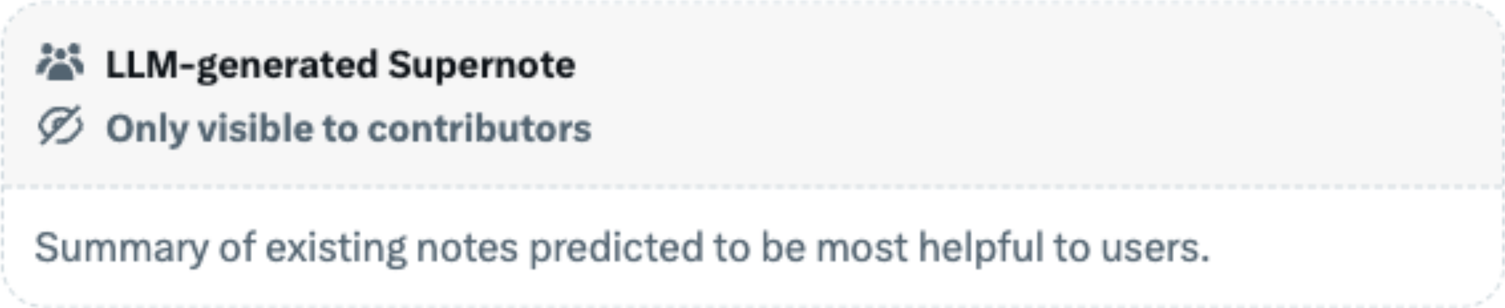


Rate proposed Community Notes

Only visible to contributors

Fact-checking note written by Community Note contributor

B



LLM-generated Supernote

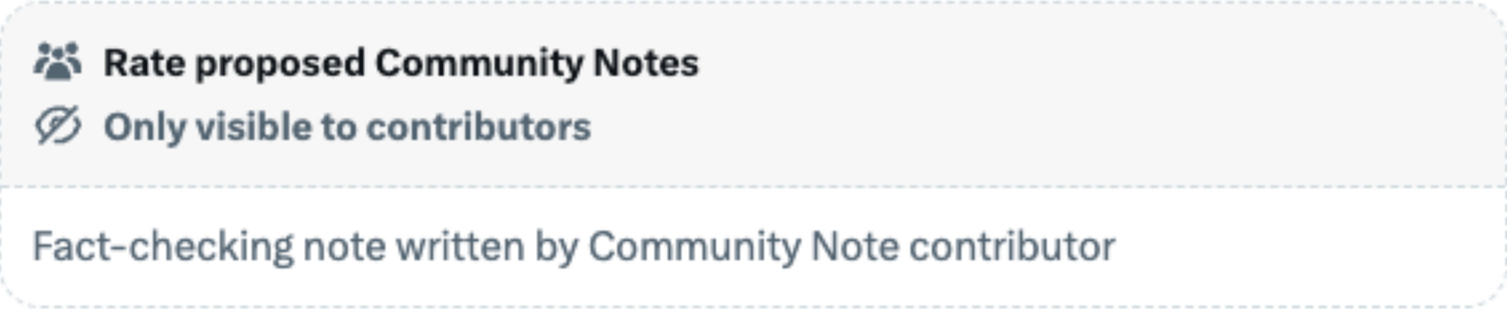
Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

Human Evaluation

We run a pilot (n=15) to evaluate Supernotes against existing notes

A



Rate proposed Community Notes

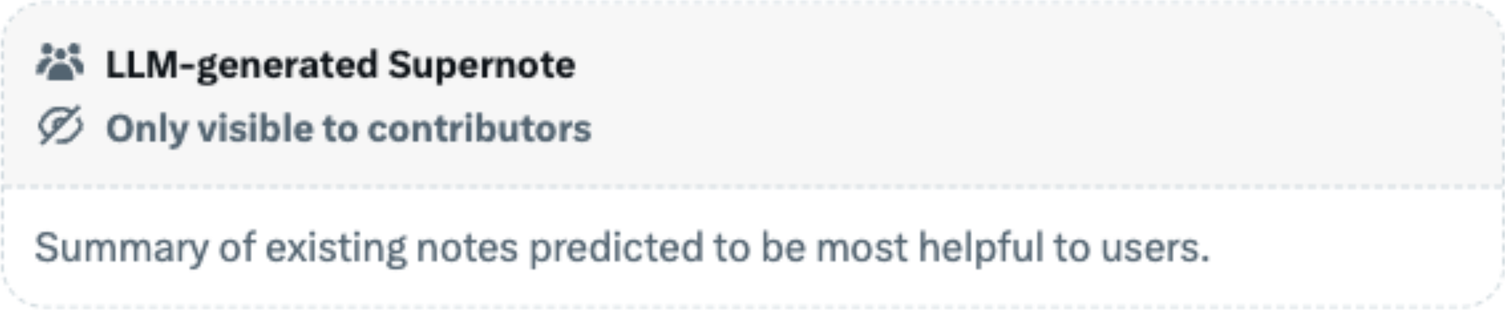
Only visible to contributors

Fact-checking note written by Community Note contributor

1. Is this note helpful?

🙁 😐 🙂

B



LLM-generated Supernote

Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

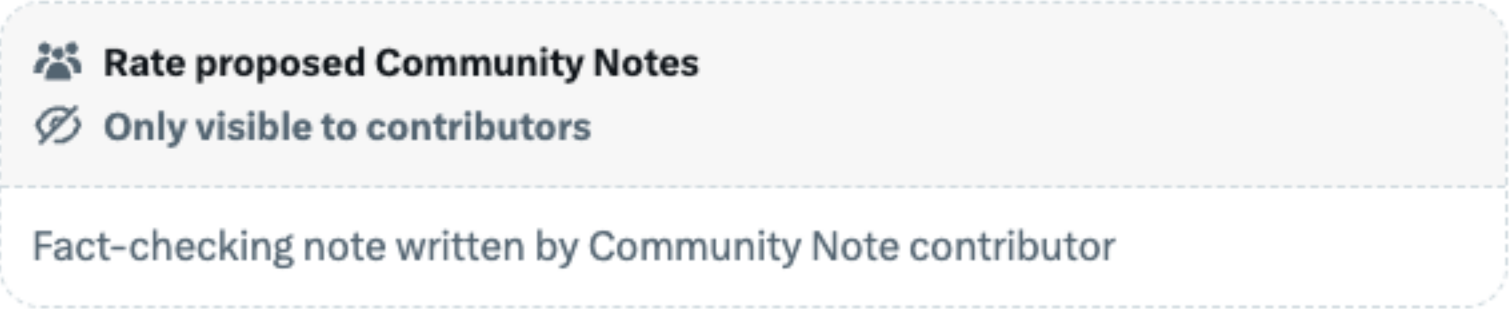
1. Is this note helpful?

🙁 😐 🙂

Human Evaluation

We run a pilot (n=15) to evaluate Supernotes against existing notes

A



Rate proposed Community Notes

Only visible to contributors

Fact-checking note written by Community Note contributor

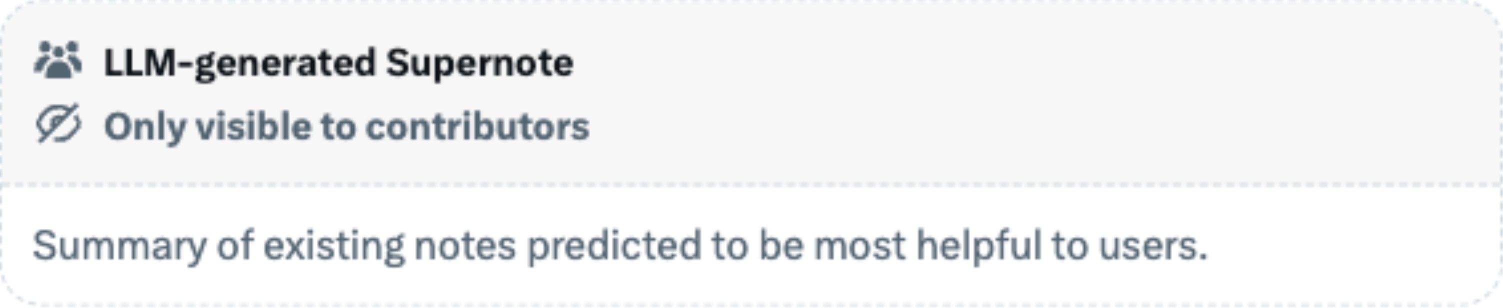
1. Is this note helpful?

☹ 😐 ☺

2. Agree/Disagree:

a. High quality sources ●—●—●—●—●
b. Clarity ●—●—●—●—●
c. Comprehensive ●—●—●—●—●
d. Context ●—●—●—●—●
e. Non argumentative ●—●—●—●—●

B



LLM-generated Supernote

Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

1. Is this note helpful?

☹ 😐 ☺

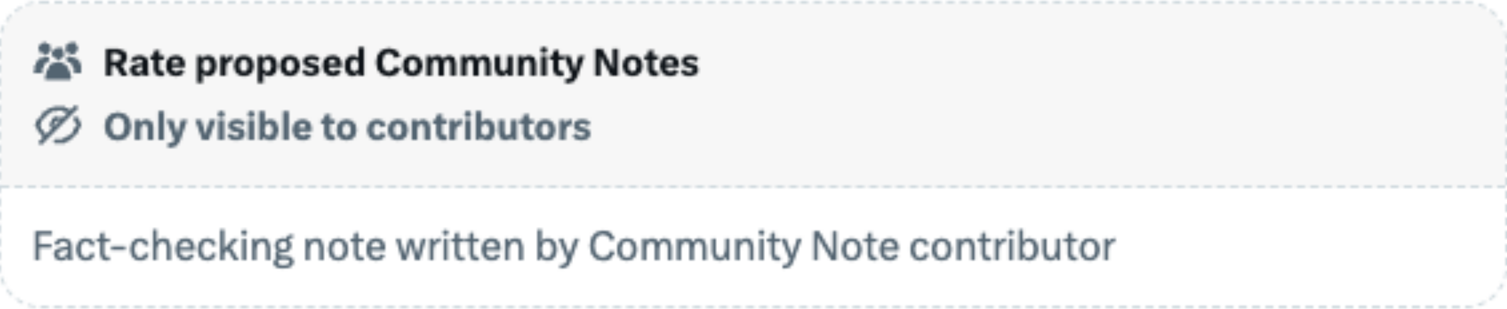
2. Agree/Disagree:

a. High quality sources ●—●—●—●—●
b. Clarity ●—●—●—●—●
c. Comprehensive ●—●—●—●—●
d. Context ●—●—●—●—●
e. Non argumentative ●—●—●—●—●

Human Evaluation

We run a pilot (n=15) to evaluate Supernotes against existing notes

A



Rate proposed Community Notes

Only visible to contributors

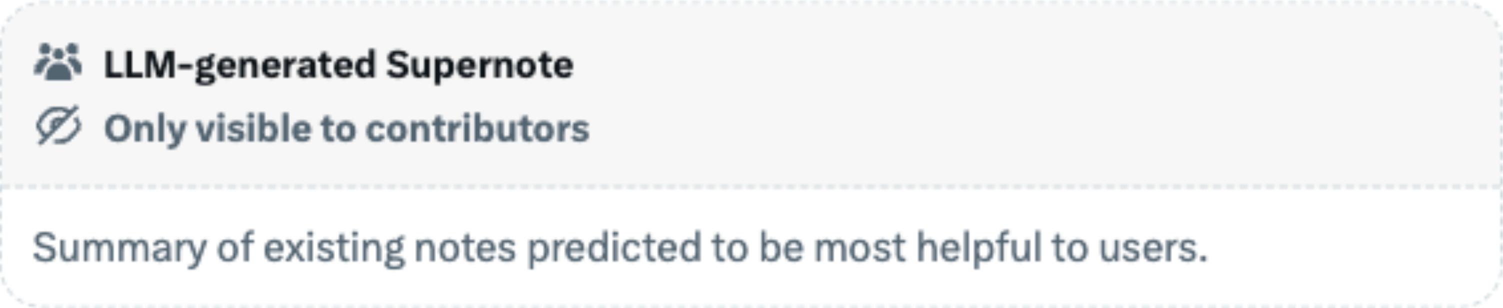
Fact-checking note written by Community Note contributor

1. Is this note helpful?

☹ 😐 ☺

2. Agree/Disagree:
   a. High quality sources ●—●—●—●—●
   b. Clarity ●—●—●—●—●
   c. Comprehensive ●—●—●—●—●
   d. Context ●—●—●—●—●
   e. Non argumentative ●—●—●—●—●

B



LLM-generated Supernote

Only visible to contributors

Summary of existing notes predicted to be most helpful to users.

1. Is this note helpful?

☹ 😐 ☺

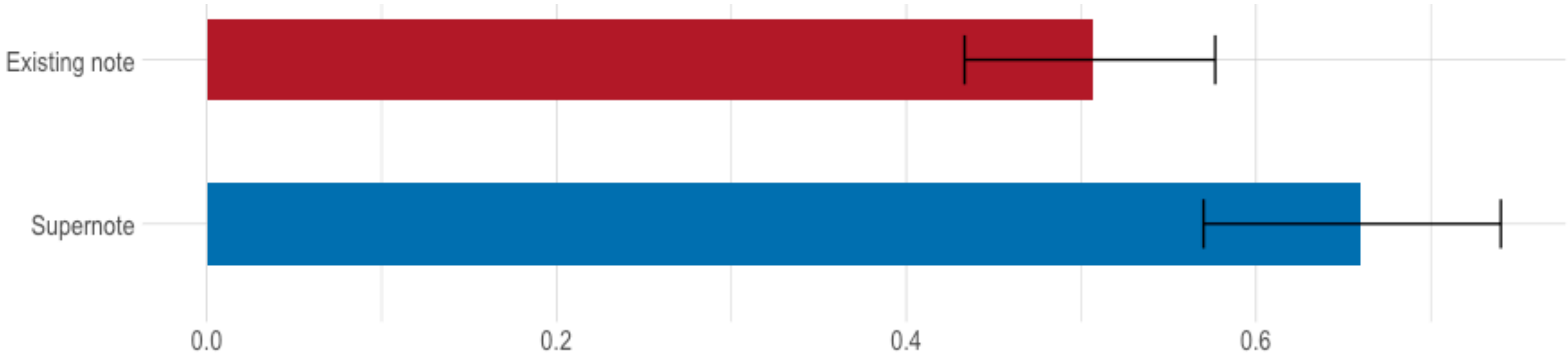
2. Agree/Disagree:
   a. High quality sources ●—●—●—●—●
   b. Clarity ●—●—●—●—●
   c. Comprehensive ●—●—●—●—●
   d. Context ●—●—●—●—●
   e. Non argumentative ●—●—●—●—●

3. Which one is better?

Ⓐ Ⓑ

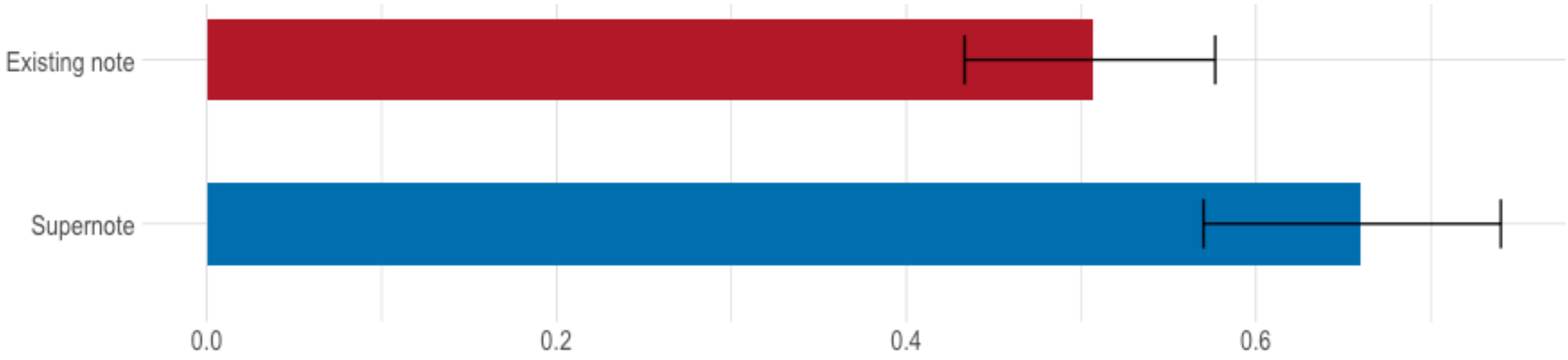
Supernotes vs. Existing Notes

1. Supernotes are more helpful than existing notes

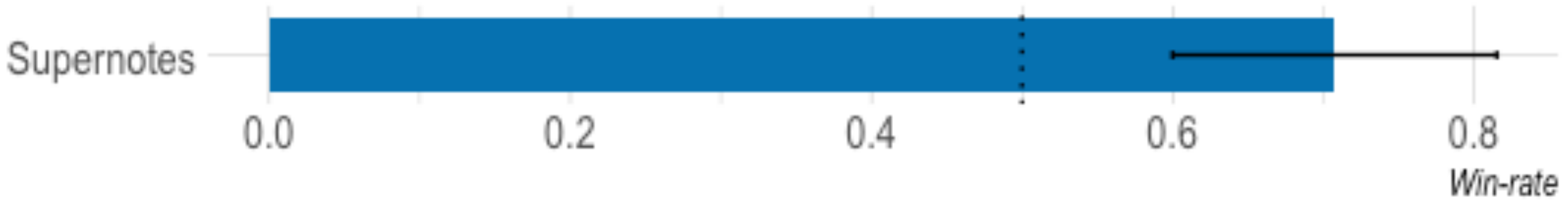


Supernotes vs. Existing Notes

1. **Supernotes are more helpful than existing notes**

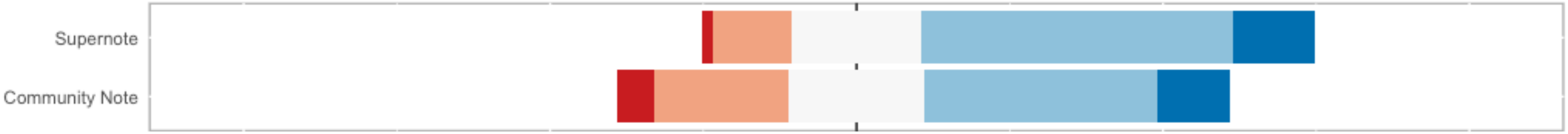


2. **Even when an existing note is helpful, a Supernote is preferred**



Supernotes Attributes

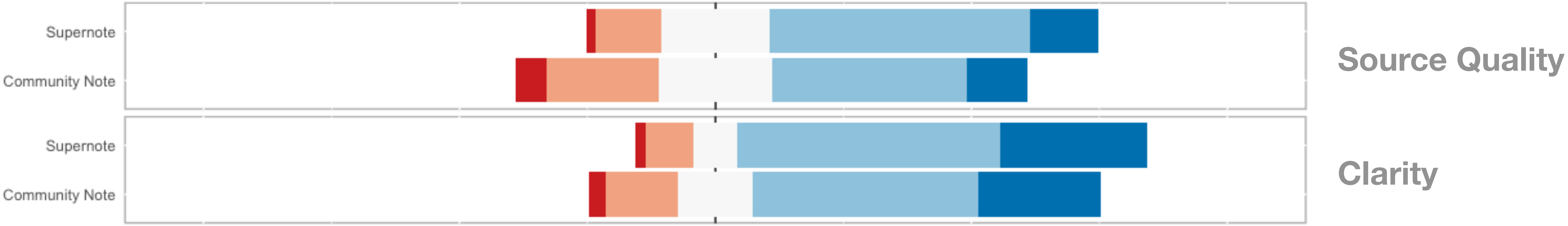
3. They follow guidelines for good fact-checking



Source Quality

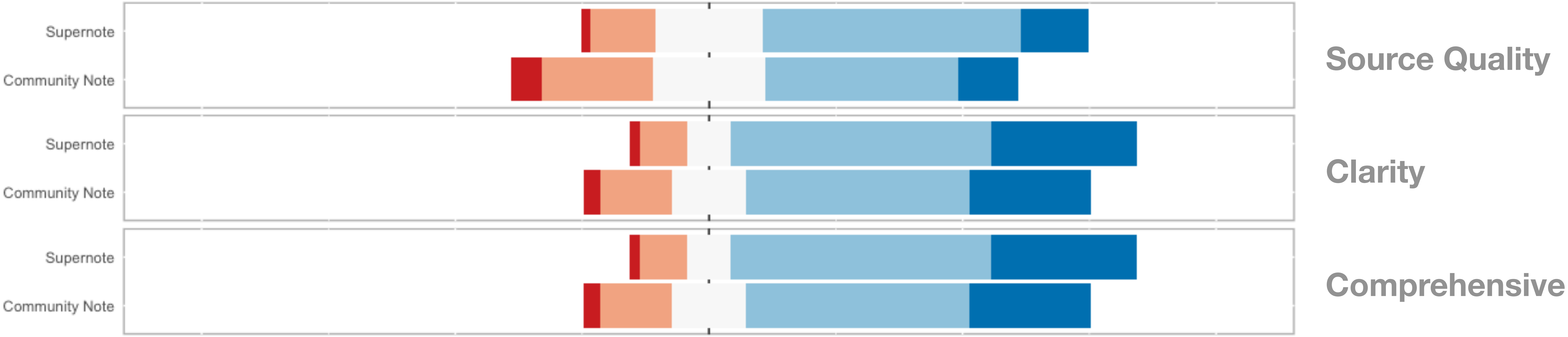
Supernotes Attributes

3. They follow guidelines for good fact-checking



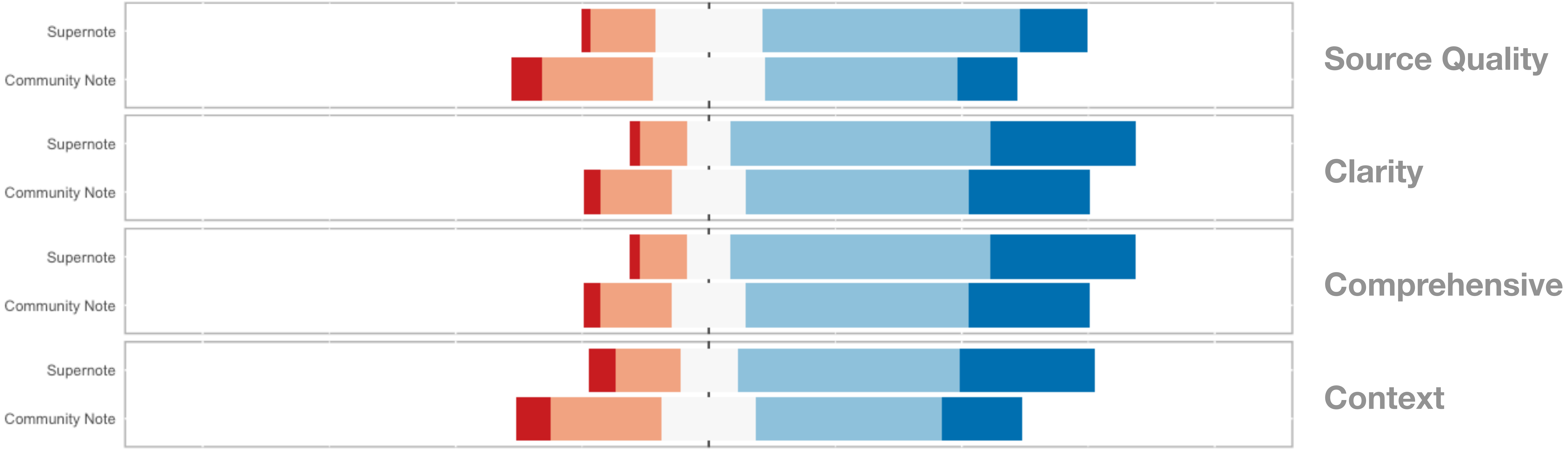
Supernotes Attributes

3. They follow guidelines for good fact-checking



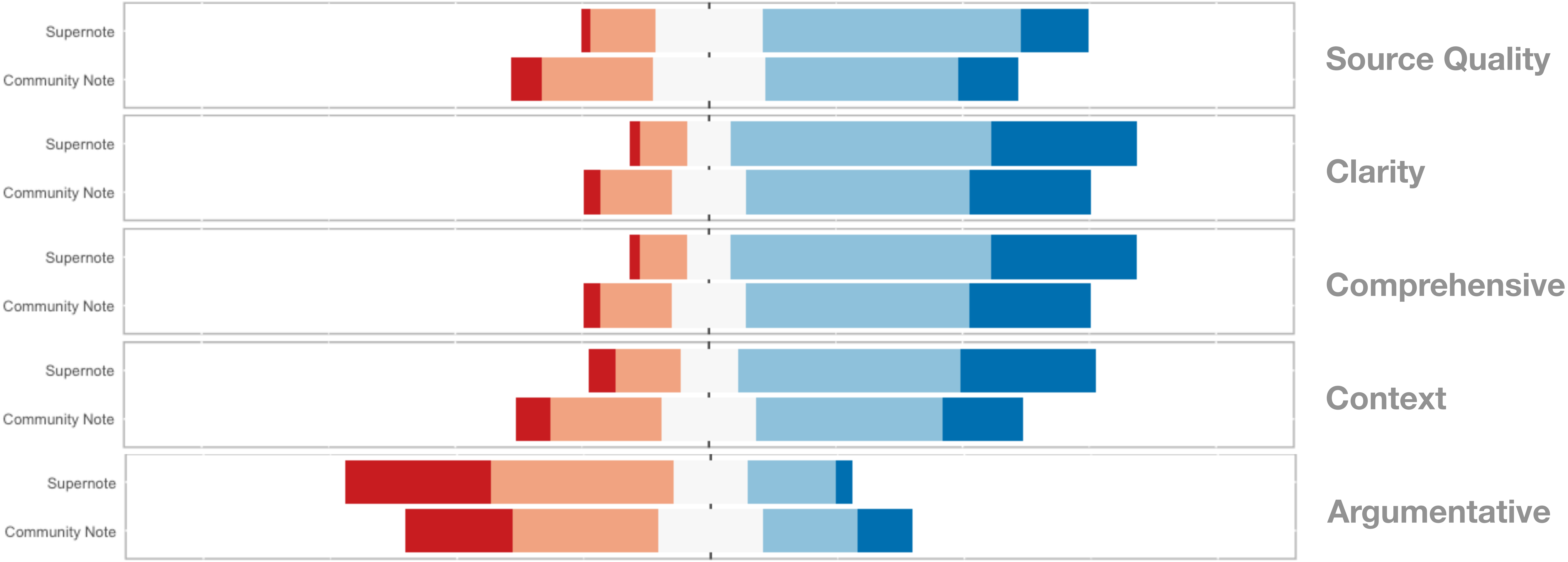
Supernotes Attributes

3. They follow guidelines for good fact-checking



Supernotes Attributes

3. They follow guidelines for good fact-checking



Ongoing Work

1. A larger study with more human participants
2. Ablation studies to measure impact of each part of our pipeline
3. Attribution issues - humans must get partial credit for writing notes that are used in a Supernote

The *Supernote*

Enhancing Crowd-sourced factchecking using LLM-driven consensus

Soham De | sohamde@uw.edu

IC^{2}S^2 2024