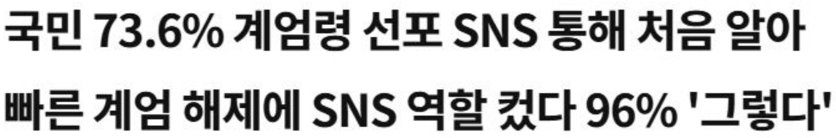
Cleanity

The Social Harmony

20180532 Heechan Lee
20200433 Gyuho Lee
20210277 Hojun Park

Communication inside Community

Internet community showed us its strength by being the information network when the proclamation of martial law was announced [1].



국민 73.6% 계엄령 선포 SNS 통해 처음 알아

빠른 계엄 해제에 SNS 역할 컸다 96% '그렇다'

73.6% of the people first learned through social media that martial law was declared.
96% answered that SNS played a large role in the rapid end of martial law.

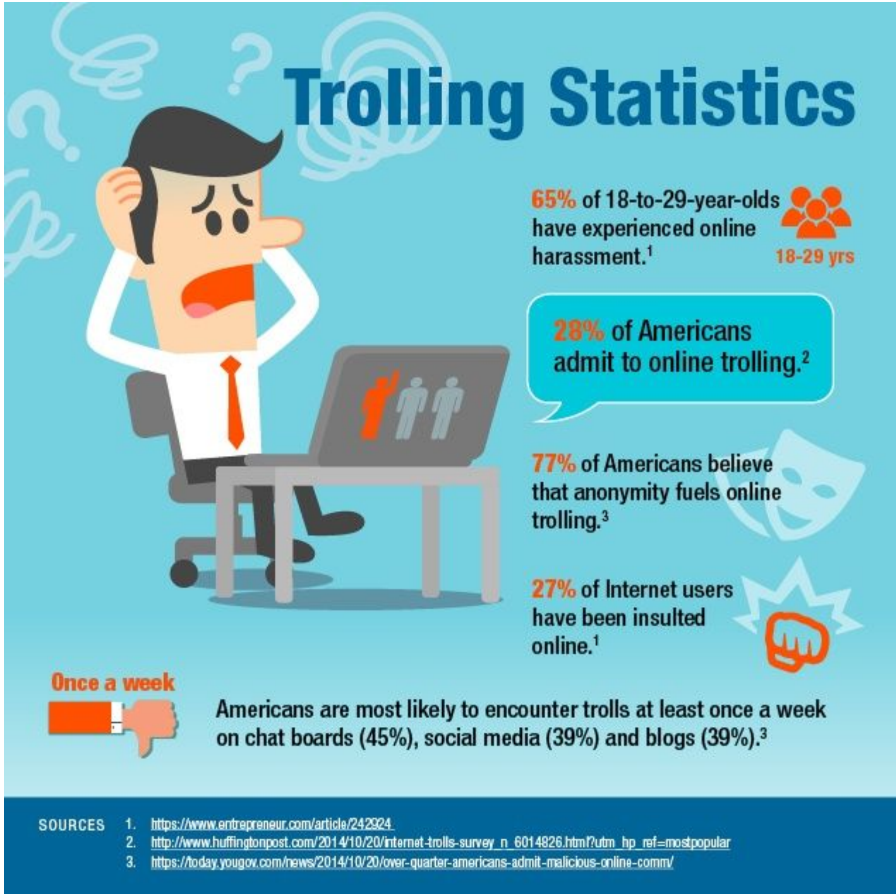
This quick delivery of information and presentation of opinions is the reason for the existence of the community, but problems sometimes arise.

[1] https://www.hankyung.com/article/202412068632g

Trolls, Everywhere

Trolling is a major problem that occurs online that cause several problems in maintaining proper communication.

[1]https://www.marketwatch.com/story/online-trolls-are-ruining-social-media-marketing-2016-07-13



Trolling Statistics

65% of 18-to-29-year-olds have experienced online harassment.[1]

28% of Americans admit to online trolling.[2]

77% of Americans believe that anonymity fuels online trolling.[3]

27% of Internet users have been insulted online.[1]

Americans are most likely to encounter trolls at least once a week on chat boards (45%), social media (39%) and blogs (39%).[3]

SOURCES
1. https://www.entrepreneur.com/article/242924
2. http://www.huffingtonpost.com/2014/10/20/internet-trolls-survey_n_6014826.html?utm_hp_ref=mostpopular
3. https://today.yougov.com/news/2014/10/20/over-quarter-americans-admit-malicious-online-comm/

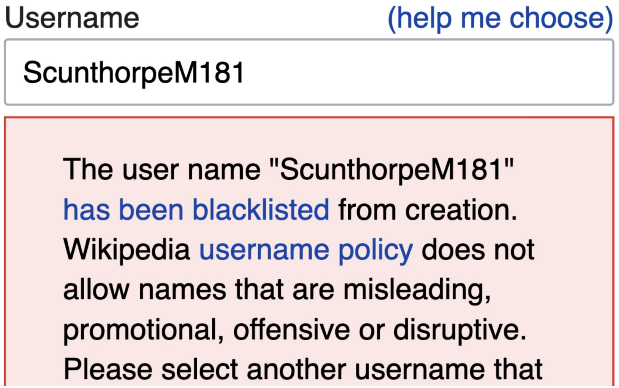
Problems on Existing Solutions with Current ChatBots

1. Detecting banned words → Scunthorpe problem [1]

2. Complex rules depending on different community

ex) community that handles [KBO vs Hanhwa Eagles]

3. Trolling is hard to block fundamentally



ScunthorpeM181

The user name "ScunthorpeM181" has been blacklisted from creation. Wikipedia username policy does not allow names that are misleading, promotional, offensive or disruptive. Please select another username that

→ Due to these complexities in managing the community, widely used methods were to manually scan the writings one by one and delete/block them.

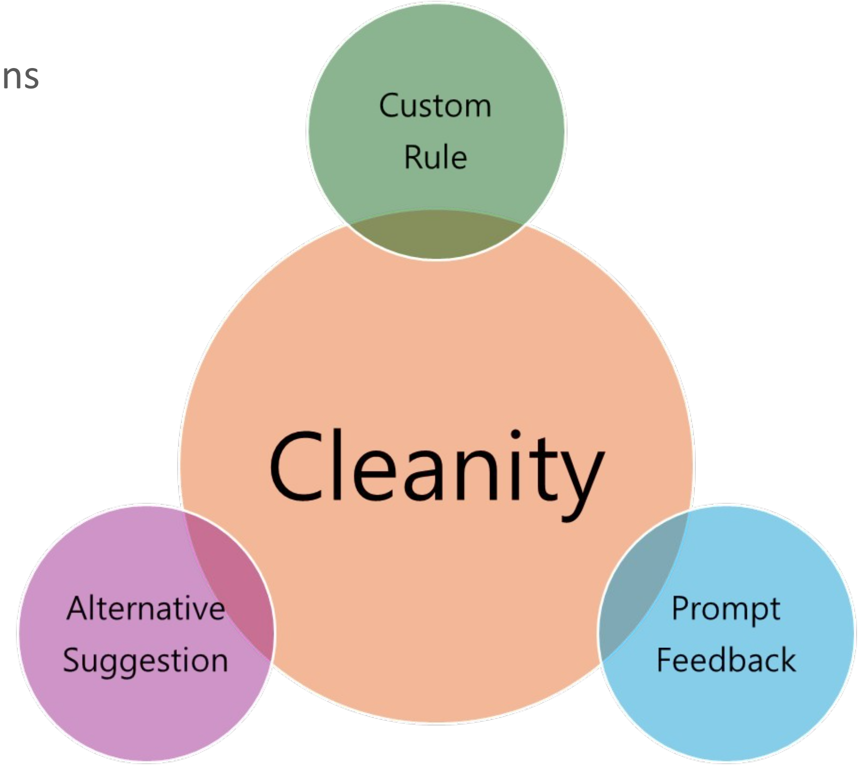
[1] https://en.wikipedia.org/wiki/Scunthorpe_problem#/media/File:Scunthorpe_problem_(cropped).png

Suggestion

- **Custom Rule**
  - Adjustable rules to diverse and specific situations

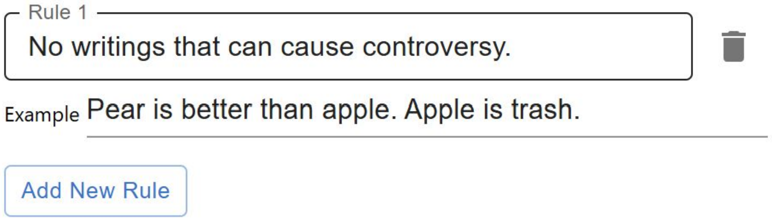
- **Alternative Suggestion**
  - Replace contents to recommended ver.

- **Prompt Feedback**
  - Flexible regulation and suggestion



Custom Rules

Moderator can set the rule by entering the content of the rule and an example of breaking the rule, according to the community they belong.



Rule 1

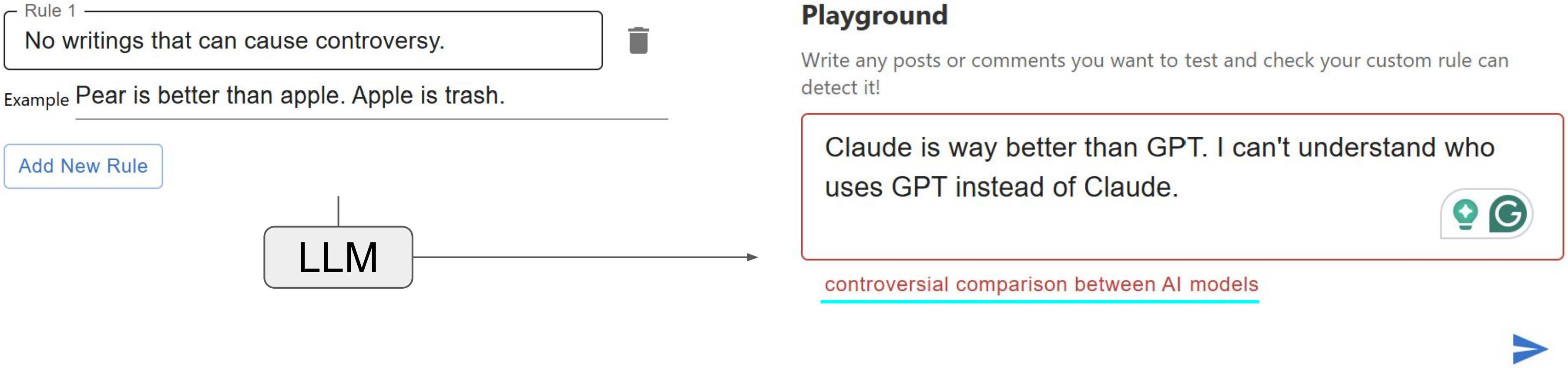
No writings that can cause controversy.

Example Pear is better than apple. Apple is trash.

The rule and example are included as shots in the LLM together to help the rule work as intended by the modulator.

Custom Rules

Moderator can set the rule by entering the content of the rule and an example of breaking the rule, according to the community they belong.



Rule 1

No writings that can cause controversy.

Example Pear is better than apple. Apple is trash.

Playground

Write any posts or comments you want to test and check your custom rule can detect it!

Claude is way better than GPT. I can't understand who uses GPT instead of Claude.

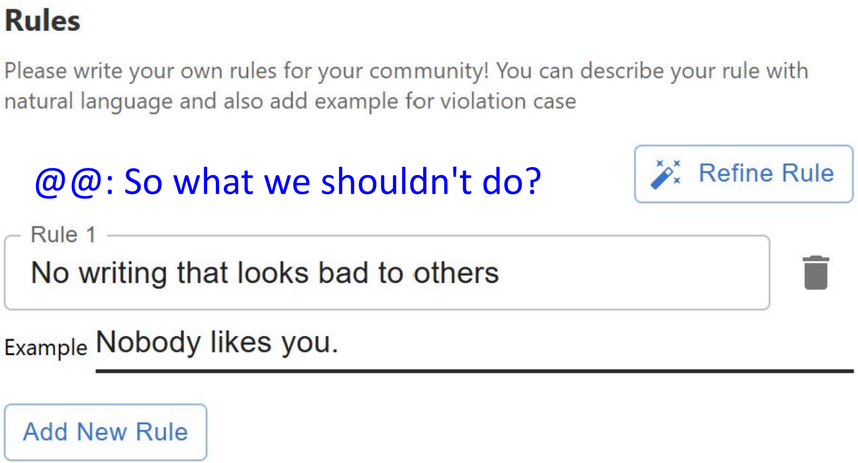
controversial comparison between AI models

There's also a Playground where you can make sure that the rules you set are set up well.

Custom Rules

However, after testing several times, we find out…

- annoying to make/update rules every time
- rules can be set too naively sometimes



Rules

Please write your own rules for your community! You can describe your rule with natural language and also add example for violation case

@@: So what we shouldn't do?

Refine Rule

Rule 1

No writing that looks bad to others

Example Nobody likes you.

Add New Rule

Mrs. Mutner liked to go over a few of her rules on the first day of school.

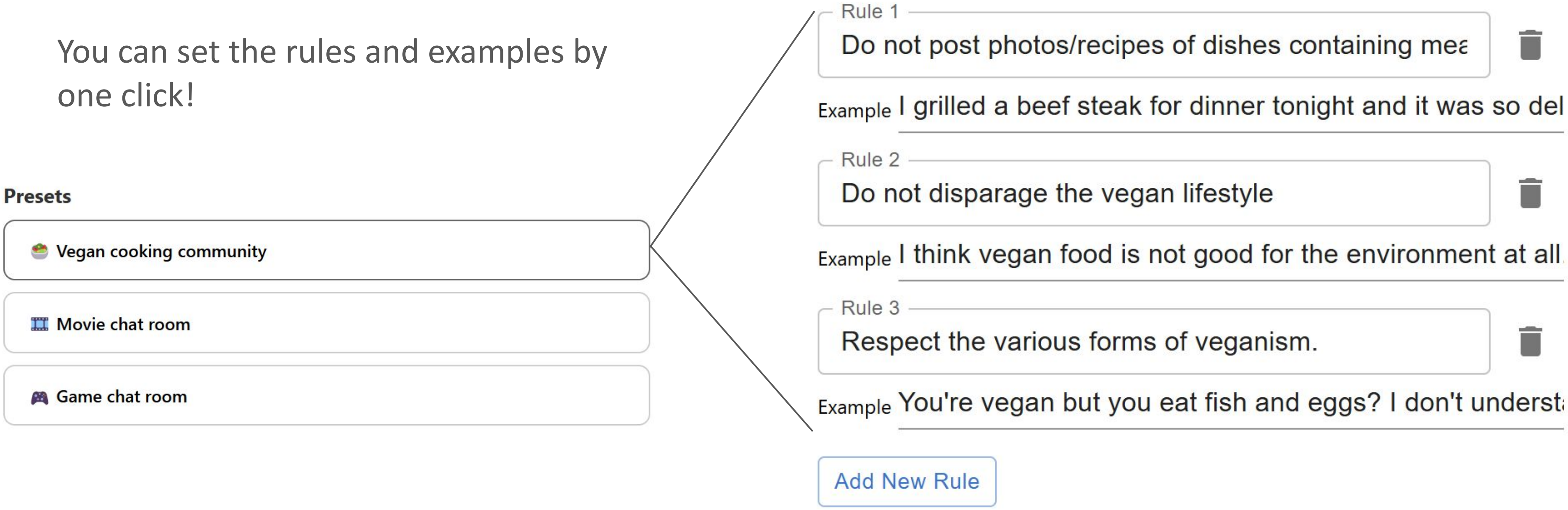
[1] https://ar.pinterest.com/pin/640566746971828108/

Custom Rules

So we made extra features!

1. Rule Preset

You can set the rules and examples by one click!



Presets

- 🥗 Vegan cooking community
- 🎞 Movie chat room
- 🎮 Game chat room

Rule 1: Do not post photos/recipes of dishes containing mea

Example: I grilled a beef steak for dinner tonight and it was so del

Rule 2: Do not disparage the vegan lifestyle

Example: I think vegan food is not good for the environment at all.

Rule 3: Respect the various forms of veganism.

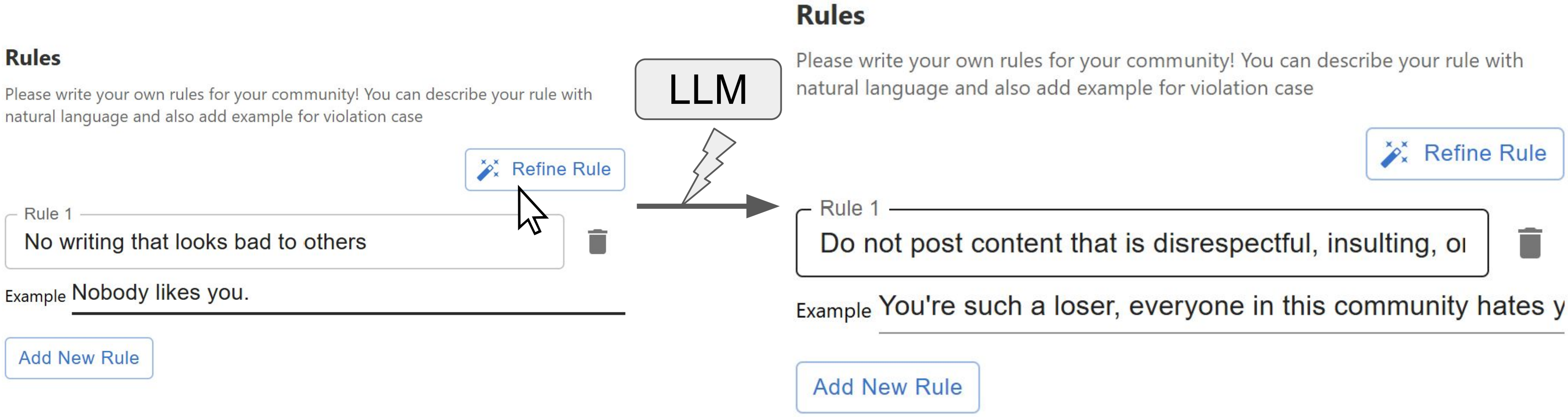
Example: You're vegan but you eat fish and eggs? I don't underst

Custom Rules

So we made extra features!

2. Rule Refine

Rules and examples are included as shots in the LLM, and LLM refines it to better rule/example that’s more suitable for discrimination.



Rules

Please write your own rules for your community! You can describe your rule with natural language and also add example for violation case

Rule 1: No writing that looks bad to others

Example: Nobody likes you.

Rules

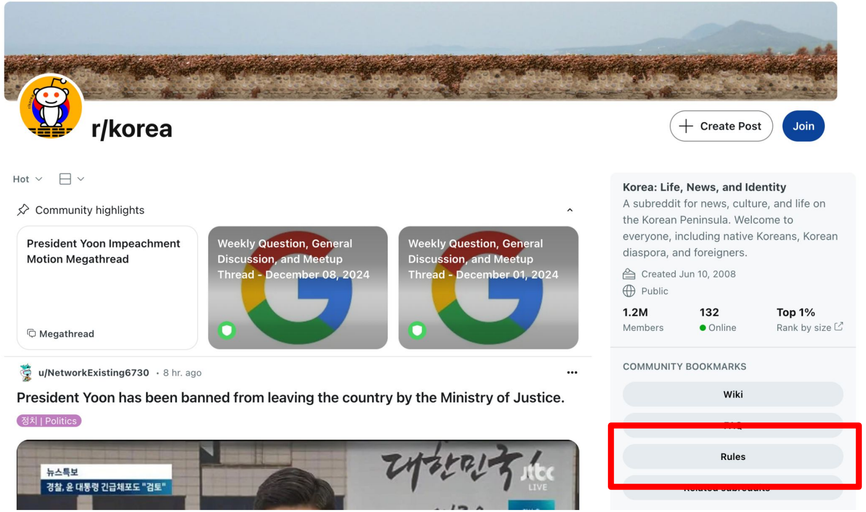
Please write your own rules for your community! You can describe your rule with natural language and also add example for violation case

Rule 1: Do not post content that is disrespectful, insulting, oı

Example: You're such a loser, everyone in this community hates y

Rule Awareness of Users

Even there are rules for communities, but usually, **users are not very aware of the rules**



r/korea

Community highlights

President Yoon Impeachment Motion Megathread

Megathread

Weekly Question, General Discussion, and Meetup Thread - December 08, 2024

Weekly Question, General Discussion, and Meetup Thread - December 01, 2024

u/NetworkExisting6730 · 8 hr. ago

President Yoon has been banned from leaving the country by the Ministry of Justice.

정치 | Politics

Korea: Life, News, and Identity

A subreddit for news, culture, and life on the Korean Peninsula. Welcome to everyone, including native Koreans, Korean diaspora, and foreigners.

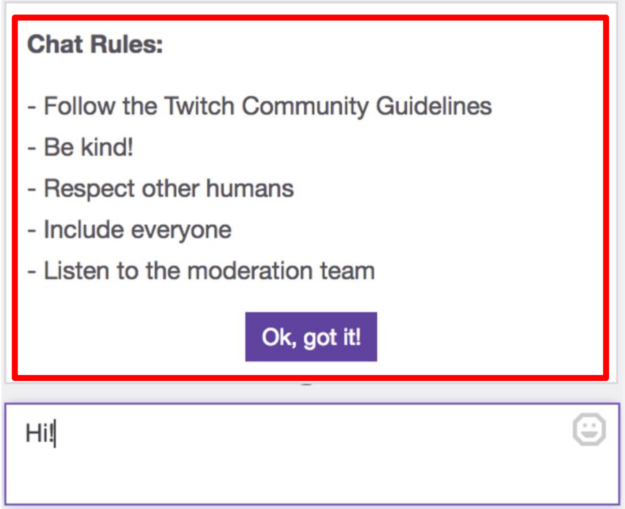
Created Jun 10, 2008

Public

1.2M	132	Top 1%
Members	Online	Rank by size

COMMUNITY BOOKMARKS

Each reddit community can make their own community **rules**



Chat Rules:

- Follow the Twitch Community Guidelines
- Be kind!
- Respect other humans
- Include everyone
- Listen to the moderation team

In twitch chat, **chat rules box** is popped up when user click the chat-box first time

Why Rule Awareness Important?

When not aware of rules;

- Users who are unintentionally abusing frequently come out
  - **Workload of Moderators increase!**
- Mods - Users or User - User fight frequently
- Some users might not agree with the rules that they didn’t know!

Make users aware of rules;

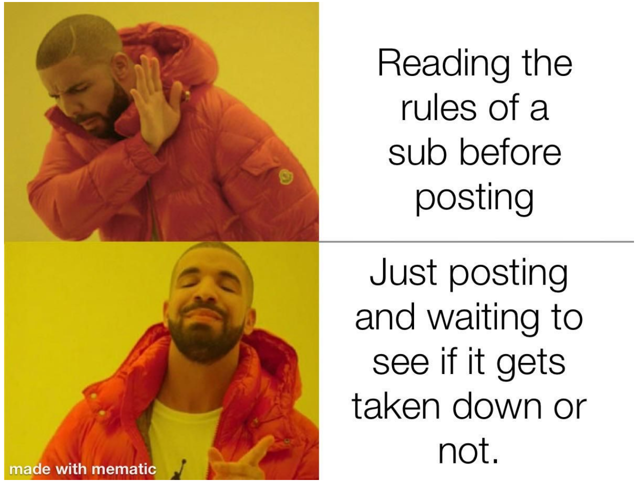
- Unintentional abusers decrease
- Users can **collaboratively improve and build the rules together**

=> Awareness can help the community to align well between users and improve the rules

How to make users aware of rules

1) Force users to read the rules

=> ***Almost impossible!***
Lots of communities are trying to make users read the rules but users just click the ‘I checked’ button and never read.


Reading the rules of a sub before posting
Just posting and waiting to see if it gets taken down or not.
made with mematic

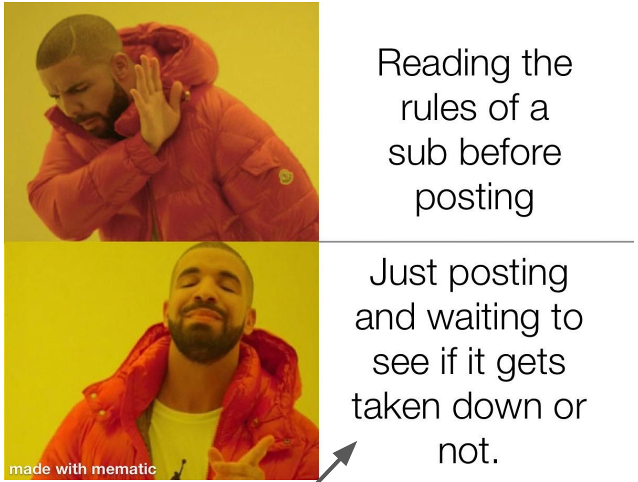
How to make users aware of rules

1) Force users to read the rules

=> ***Almost impossible!***
Lots of communities are trying to make users read the rules but users just click the ‘check’ button and never read.

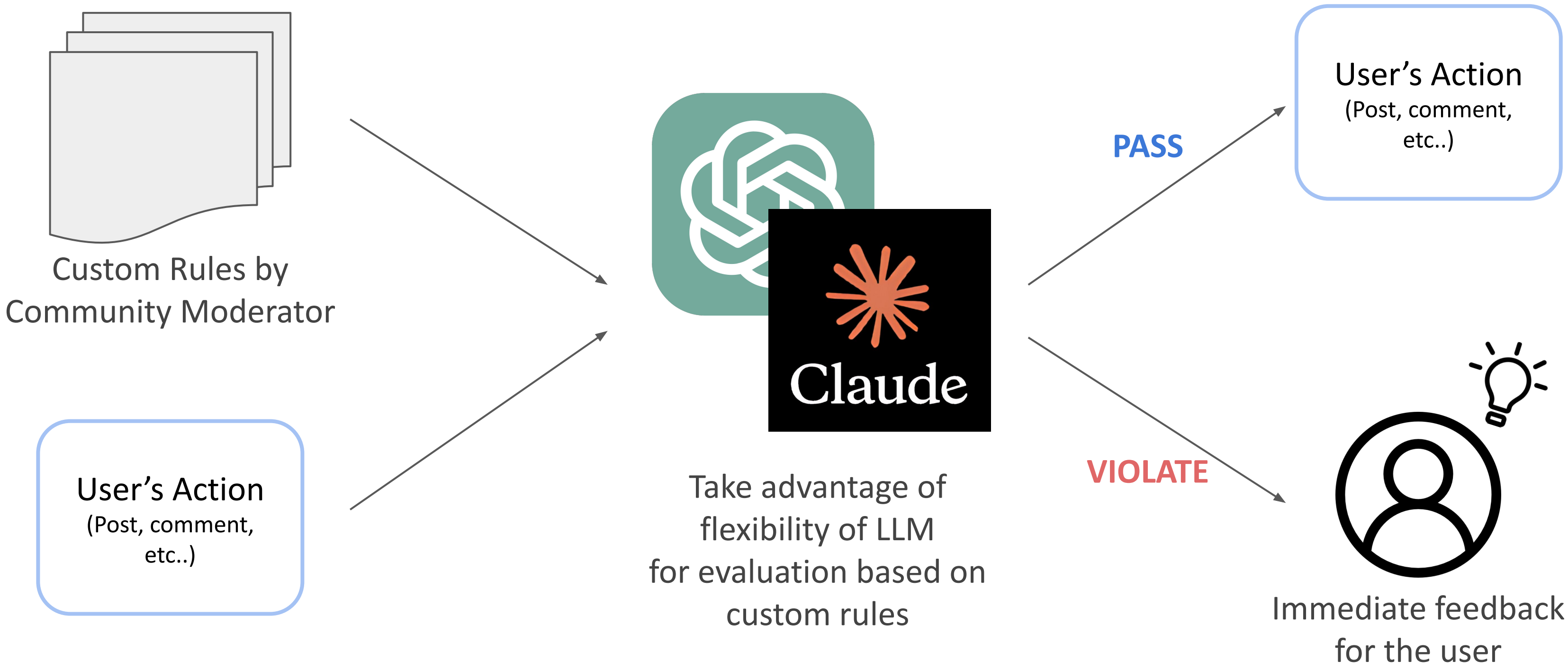
2) Naturally give feedback based on the rules when users do some action

Let users do what they want, and **give proper intervention if they break the rules**


Reading the rules of a sub before posting
Just posting and waiting to see if it gets taken down or not.
made with mematic

Interruption in this phase

Evaluate and Feedback Pipeline



Custom Rules by Community Moderator

User's Action
(Post, comment, etc..)

Take advantage of flexibility of LLM for evaluation based on custom rules

PASS

User's Action
(Post, comment, etc..)

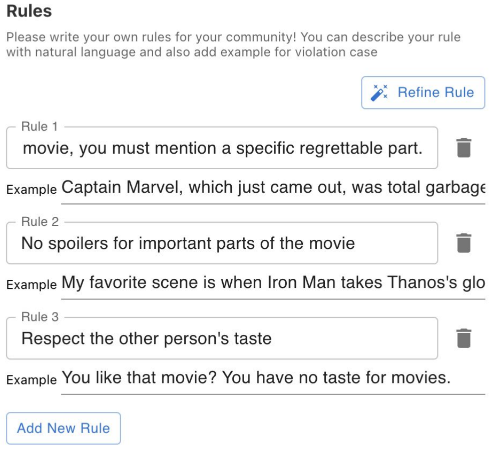
VIOLATE

Immediate feedback for the user

Our System View

Our demo targets providing *immediate / convenient / informative* feedback

- Immediate: Respond quickly before inappropriate content affects others
- Convenient: **Decrease the load of moderators** who should surveillance community all day
- Informative: Feedback can **make users aware of the rules well**



Rules

Please write your own rules for your community! You can describe your rule with natural language and also add example for violation case

Rule 1: movie, you must mention a specific regrettable part.

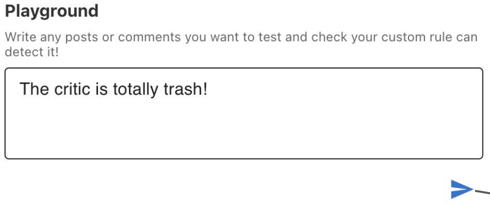
Example: Captain Marvel, which just came out, was total garbage

Rule 2: No spoilers for important parts of the movie

Example: My favorite scene is when Iron Man takes Thanos's glo

Rule 3: Respect the other person's taste

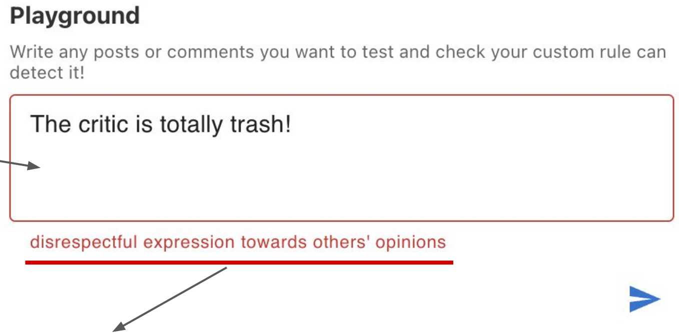
Example: You like that movie? You have no taste for movies.



Playground

Write any posts or comments you want to test and check your custom rule can detect it!

The critic is totally trash!



Playground

Write any posts or comments you want to test and check your custom rule can detect it!

The critic is totally trash!

disrespectful expression towards others' opinions

Immediate feedback after LLM evaluation

So Far…

We provided immediate/convenient/informative feedback to our users.

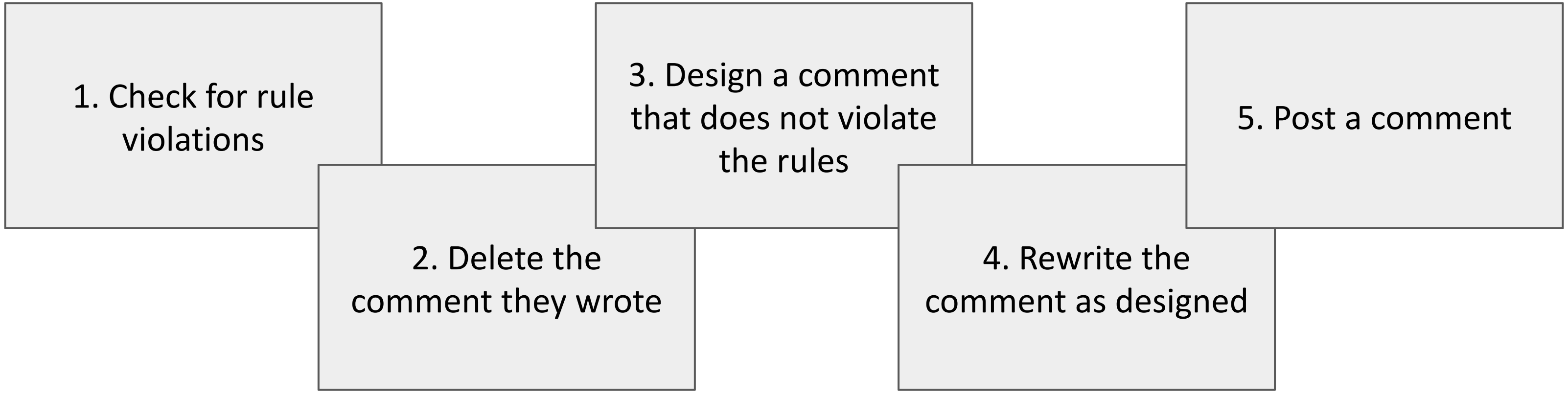
- But is this enough?
- Will users really behave appropriately just by receiving feedback?

We need **Additional features** to help users to leave proper comments

- **Censor/Limit** users to prevent inappropriate comments (X)
- **Encourage** users to write proper comments (O)

But, how to encourage them?

Process of Editing Original Comment



1. Check for rule violations

2. Delete the comment they wrote

3. Design a comment that does not violate the rules

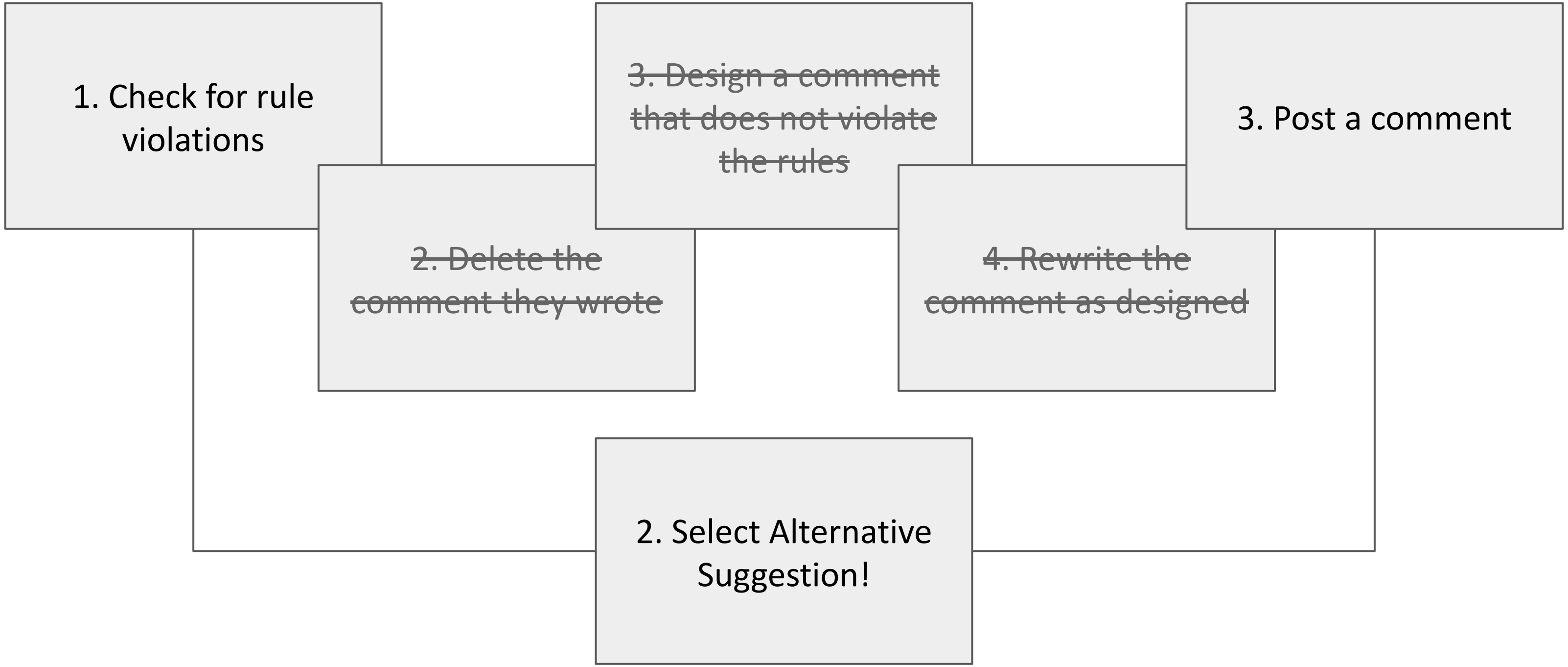
4. Rewrite the comment as designed

5. Post a comment

We found that the process of users editing their comments correctly was more cumbersome than expected.

- Have to delete what they have wrote.
- Have to write something different from their natural original thoughts.
- Have to rewrite everything.

What If We Make it Easier?



1. Check for rule violations

~~2. Delete the comment they wrote~~

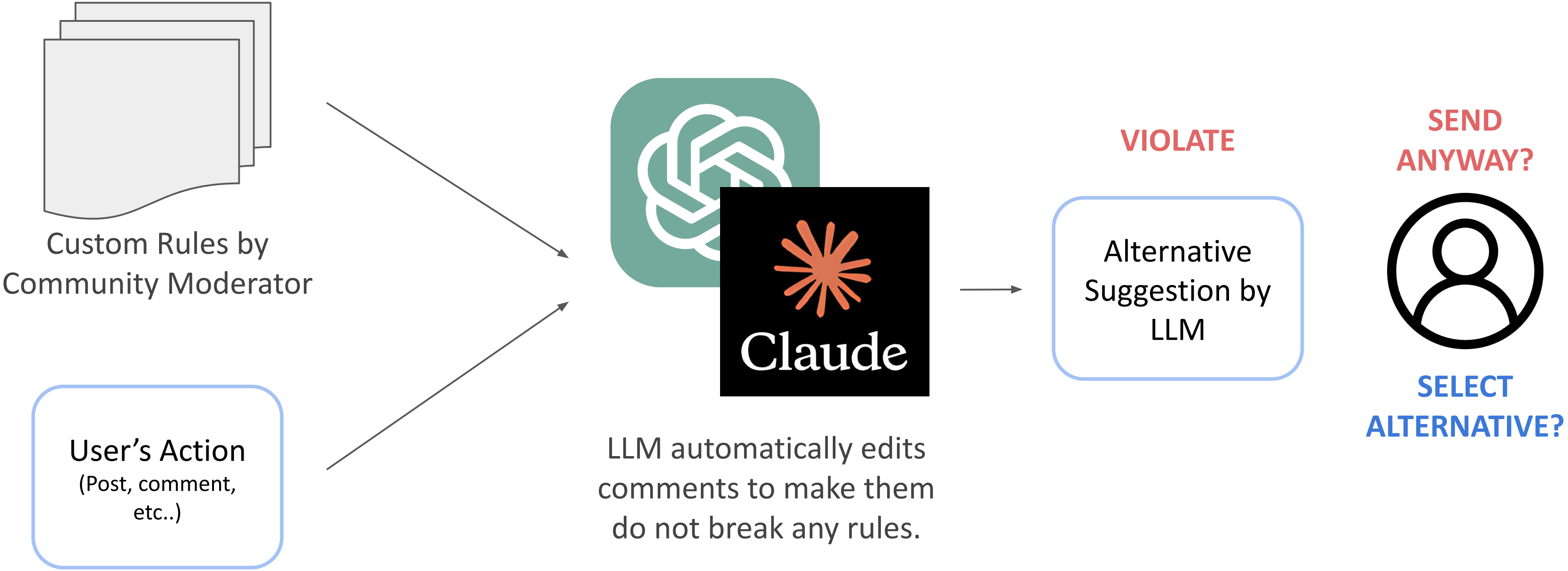
~~3. Design a comment that does not violate the rules~~

~~4. Rewrite the comment as designed~~

3. Post a comment

2. Select Alternative Suggestion!

Alternative Suggestion Process



Custom Rules by Community Moderator

User's Action
(Post, comment, etc..)

LLM automatically edits comments to make them do not break any rules.

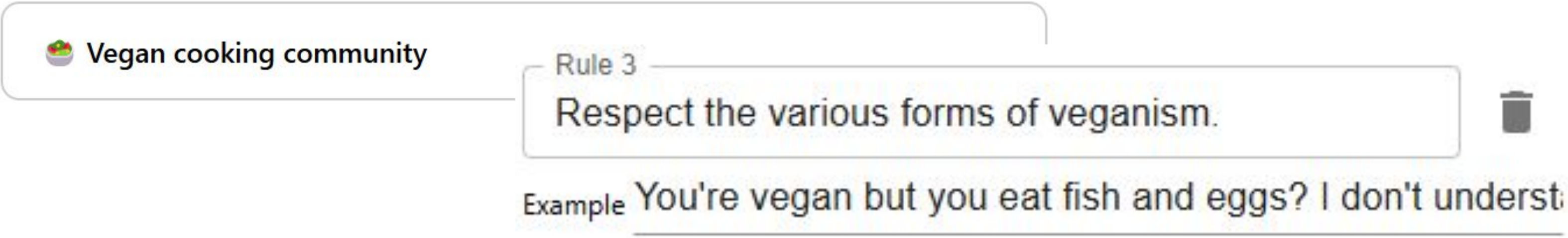
VIOLATE

Alternative Suggestion by LLM

SEND ANYWAY?

SELECT ALTERNATIVE?

Alternative Suggestion Example

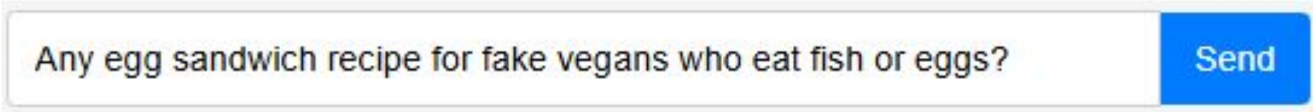


Vegan cooking community

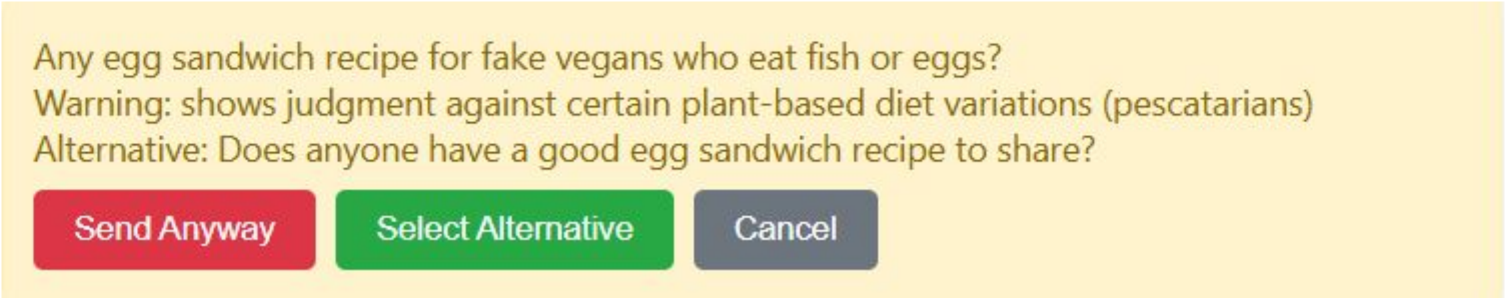
Rule 3

Respect the various forms of veganism.

Example You're vegan but you eat fish and eggs? I don't underst

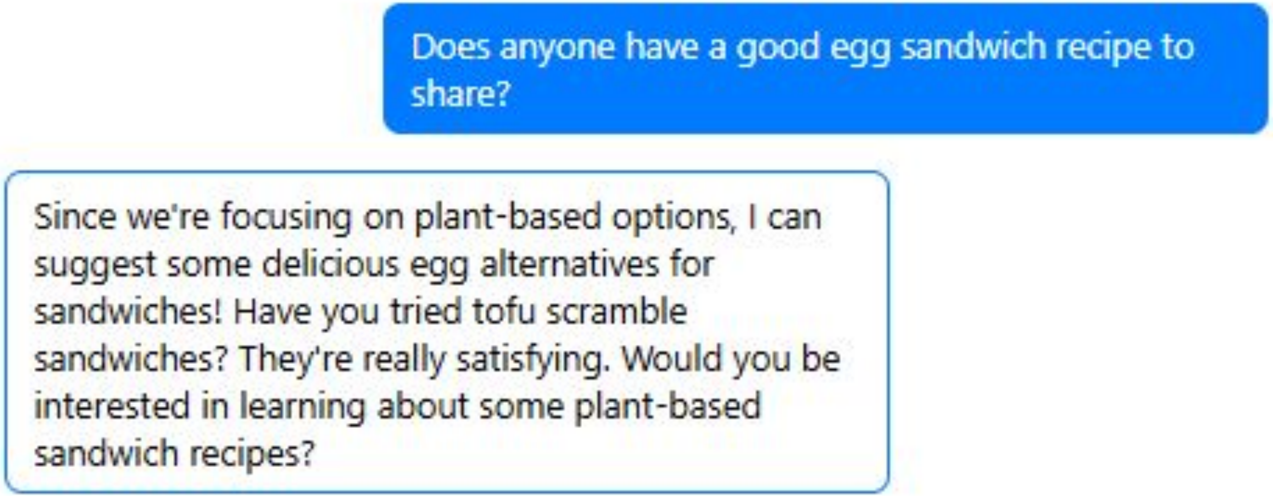


Any egg sandwich recipe for fake vegans who eat fish or eggs?



Any egg sandwich recipe for fake vegans who eat fish or eggs?
Warning: shows judgment against certain plant-based diet variations (pescatarians)
Alternative: Does anyone have a good egg sandwich recipe to share?

Alternative Suggestion Example



Does anyone have a good egg sandwich recipe to share?

Since we're focusing on plant-based options, I can suggest some delicious egg alternatives for sandwiches! Have you tried tofu scramble sandwiches? They're really satisfying. Would you be interested in learning about some plant-based sandwich recipes?

Conclusion

We developed Cleanity to regulate inappropriate commenting behavior that varies across communities.

- Allow administrators to freely set custom rules
    - Assisted by Rule preset, Rule refine
- Feedback
    - Immediate, Convenient, Informative
- Suggest Alternative
    - Simplify the process of writing proper comments

With our system, we expect users to...

- To be better at learning the rules.
- To be more aware of inappropriate behavior.
- To be more likely to correct their behavior.