

Quantifying Governance of Online Communities at Web Scale

General Examination - April 23, 2024

Galen Cassebeer Weld

Supervising Committee: Tim Althoff, Amy X. Zhang, Yulia Tsvetkov, Benjamin Mako Hill, Sanjay Kairam

**Online communities are
ubiquitous.**



Discord

91%+ of the U.S. population regularly uses an online community!

Statista - Social Media Usage in the United States - Statistics & Facts

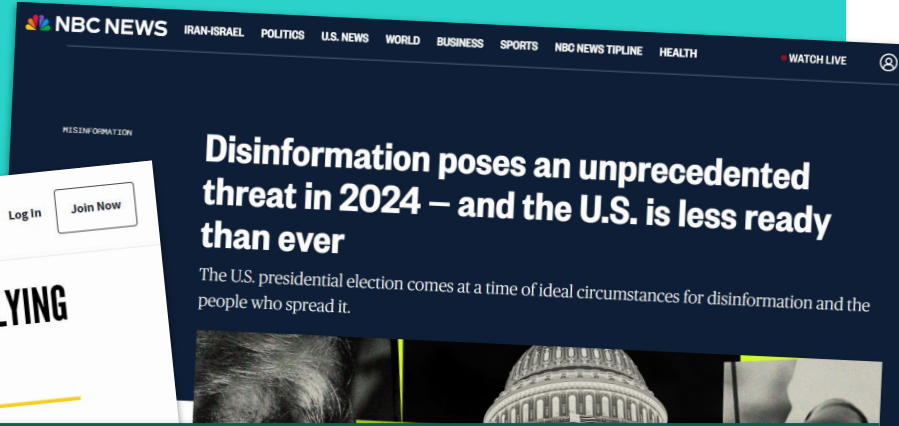


TikTok

Online communities bring people together...

- Connection for marginalized groups
- Easy access to information
- New forms of culture and media

...but also have real harms.



Social Media and Youth Mental Health

2023

The U.S. Surgeon General's Advisory



Communities rely on some form of governance to increase the good while decreasing the bad.

Community Governance

Moderators with special privileges take actions to improve the community as they see fit.

Moderators (AKA admins, mods)

- Volunteer or paid
- Visible or hidden
- Community members or outsiders
- Elected or appointed

Actions

- Set rules
- Enforce rules by..
- Remove content
- Ban users



Why Reddit?

- 500M+ monthly users*
- Thousands of communities, each with **their own topic, mods, rules**
- Relative ease of data access

Near to my Heart ❤️

- 8+ years of experience as an active moderator
- /r/photocritique has grown from 37K to 1.7M subscribers since I began moderating
- 2+ year member of the Reddit Moderator Council
- Computational Social Science Fellow at reddit since January


Community Governance varies wildly!

reddit for mountain bikers mountainbiking

new.reddit.com/r/MTB/




r/MTB

Advertise

 **reddit for mountain bikers** Joined 🔔

r/MTB

Posts Wiki Guides

 Create Post  

🔥 Hot ⚙️ New 🏠 Top ⋮

↑ 35 📄 **RULE #3 REMINDER, PLEASE READ IT**
Posted by u/Awesom3RedKite 5010-V3 1 month ago 🟢 🟢


↶ 💬 60 Comments ➦ Share ⋮

reddit for mountain bikers mountainbiking

new.reddit.com/r/mountainbiking/




r/mountainbiking

Advertise

 **mountainbiking** Join

r/mountainbiking

Posts

 Create Post  

🔥 Hot ⚙️ New 🏠 Top ⋮

↑ 28 **Pennyflake/Orange bikes** i.redd.it/haowef... Bike Picture/NBD
Posted by u/kebslcn 5 hours ago

↶ 💬 16 Comments ➦ Share ⋮

↑ **What does a highly advanced civilization have to do to get noticed around here? For 400 years we've been trying to make contact, to no avail. Then we made a huge**

Community Governance varies wildly!

The screenshot shows the 'Moderators of r/MTB' page. The browser address bar indicates the URL is 'new.reddit.com/r/MTB/about/moderators/'. The page title is 'R/MTB / MODERATORS'. Below the title is a search bar labeled 'Search for a user'. A list of four moderators is displayed, each with a profile picture, name, join date, and a 'Everything' link.

Moderator Name	Join Date	Permissions
AutoModerator	3 years ago	Everything
Viffer98	3 years ago	Everything
Awesom3RedKite	2 years ago	Everything
itskohler	1 year ago	Everything

The screenshot shows the 'Moderators of r/mountainbiking' page. The browser address bar indicates the URL is 'new.reddit.com/r/mountainbiking/about/moderators/'. The page title is 'R/MOUNTAINBIKING / MODERATORS'. Below the title is a search bar labeled 'Search for a user'. A list of one moderator is displayed, with a profile picture, name, join date, and a 'Everything' link.

Moderator Name	Join Date	Permissions
MTBiker_Boy	5 years ago	Everything

Community Governance varies wildly!

reddit for mountain bikers x mountainbiking x +

new.reddit.com/r/MTB/about/rules/

R/MTB / RULES

These are rules that visitors must follow to participate. They can be used as reasons to report or ban posts, comments, and users. Communities can have a maximum of 15 rules.

- 1 Be cool to each other!
- 2 Need help choosing a bike?
- 3 Photos should be of people riding mountain bikes.
- 4 No fundraising, karma-baiting, spam posts, or cryptic post titles.
- 5 No image macros, Rage Comics, or memes.
- 6 Links to other social media sites (Instagram, Facebook, etc.) are not allowed.
- 7 This not a Buy/Sell/Trade sub.

reddit for mountain bikers x mountainbiking x +

new.reddit.com/r/mountainbiking/about/rules/

R/MOUNTAINBIKING / RULES

These are rules that visitors must follow to participate. They can be used as reasons to report or ban posts, comments, and users. Communities can have a maximum of 15 rules.

- 1 Be Nice
- 2 No Self-Promotion Spam
- 3 Follow Rediquette
- 4 Stay appropriate
- 5 Don't overuse the modmail
- 6 English Only
- 7 Gore should be tagged as NSFW

The Question

How do we make data-driven decisions to best govern online communities?

The Answer

Quantifying the diversity of communities' governance practices and outcomes yields insights to make online communities better.

/r/science

- Governance
- Outcomes

/r/politics

- Governance
- Outcomes

/r/seattle

- Governance
- Outcomes

/r/nfl

- Governance
- Outcomes

/r/throwers

- Governance
- Outcomes

The Solution

Quantifying the diversity of communities' governance practices and outcomes yields insights to make online communities better.

This has several advantages over previous approaches.

#1

Previous work focuses on specific harms

Holistic approach to community health

#2

Collecting new data is expensive and slow

Leverage existing and historical data

#3

Qualitative work is difficult to scale

Generalize from thousands of communities

Research Activity

Defining

Assessing

Informing

Deploying

Research Focus

Governance
Practices

Community
Outcomes

Causal
Inference
Methods

Taxonomy of
Community
Values
(ICWSM 2024)

Community Perceptions of Moderators
(arXiv 2024)

Moderation
Practice
& Engagement
(in progress)

Quantitative Study
of Community
Values
(ICWSM 2022)

Analyses of News
Sharing Behavior
on Reddit
(🏆 ICWSM 2021)

LLM Coaching to
Improve
Discussion
(in progress)

Semi-Synthetic
Evaluation
Framework for CI
Methods
(ICWSM 2022)

SHERBERT
Method
for Causal
Inference
(ICWSM 2022)

Connecting Community Governance to Outcomes: Research Activities

Defining

What outcomes are important to communities?

Assessing

What current practices seem most promising?

Deploying

How can we have real world impact?

Foundational

Applied



Research Activities

Defining

Assessing

Deploying

Community Values (ICWSM '22 & '24)

What outcomes are important?

How do they vary?

Perceptions of Mods (arXiv '24)

Measuring moderation practices at scale

News Sharing Behavior (🏆 ICWSM '21)

Community affordances to improve trust

Causal Inference Methods (ICWSM '22)

Robust observational studies

Ongoing & Proposed Work

Maximize real world impact

1. Community Values

(ICWSM 2022 & ICWSM 2024)

2. Perceptions of Mods

(arXiv 2024)

3. News Sharing Behavior

( ICWSM 2021)

4. Causal Inference Methods

(ICWSM 2022)

5. Ongoing & Proposed Work

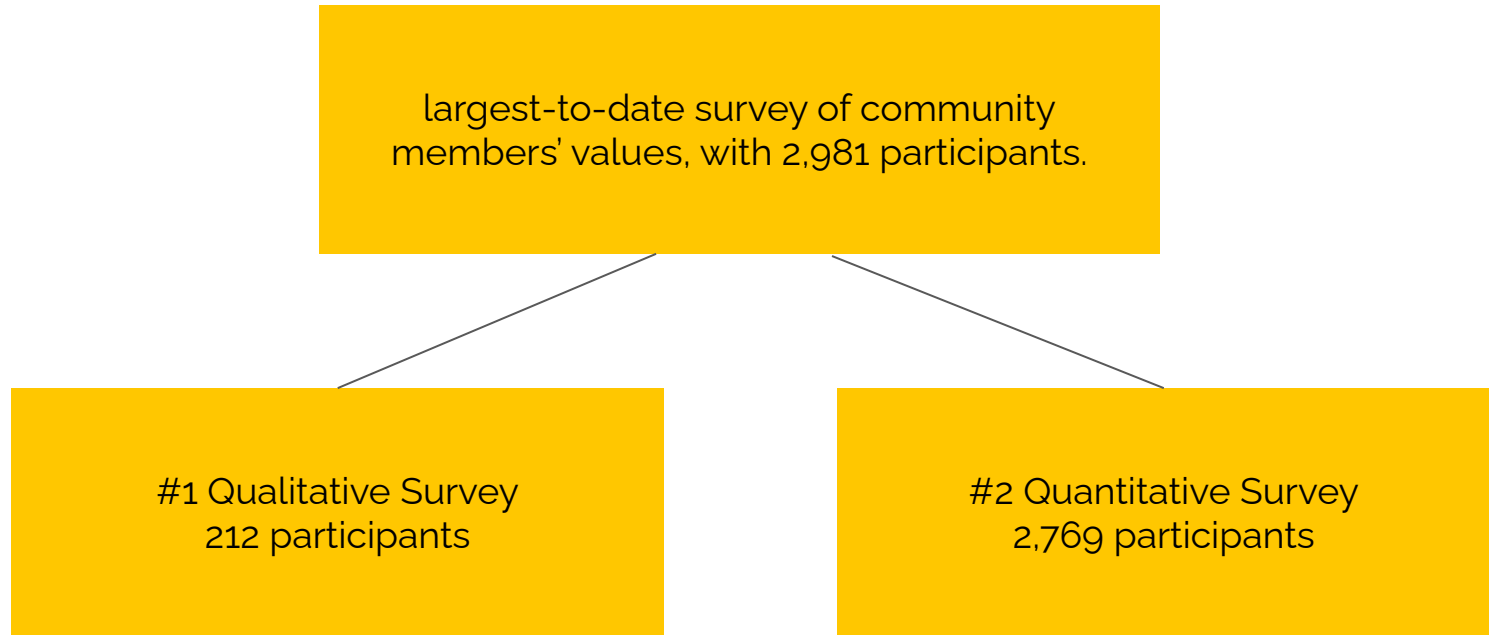
(Ongoing)

1. Community Values

(ICWSM 2022 & ICWSM 2024)

What do we mean when we say we want to make online communities 'better'?

We asked community members about their values.



We asked community members about their values.

largest-to-date survey of community members' values, with 2,981 participants.

```
graph TD; A[largest-to-date survey of community members' values, with 2,981 participants.] --- B[#1 Qualitative Survey  
212 participants]; A --- C[#2 Quantitative Survey  
2,769 participants];
```

#1 Qualitative Survey
212 participants

#2 Quantitative Survey
2,769 participants



9 values

largest-to-date survey of community members' values, with 2,981 participants.

```
graph TD; A[largest-to-date survey of community members' values, with 2,981 participants.] --- B[#1 Qualitative Survey  
212 participants]; A --- C[#2 Quantitative Survey  
2,769 participants];
```

#1 Qualitative Survey
212 participants

#2 Quantitative Survey
2,769 participants

3 dimensions

Rank (9-point scale)

how *important* is the value?

Current State (11-point scale)

how is this value *right now*?

Desired Change (3-point scale)

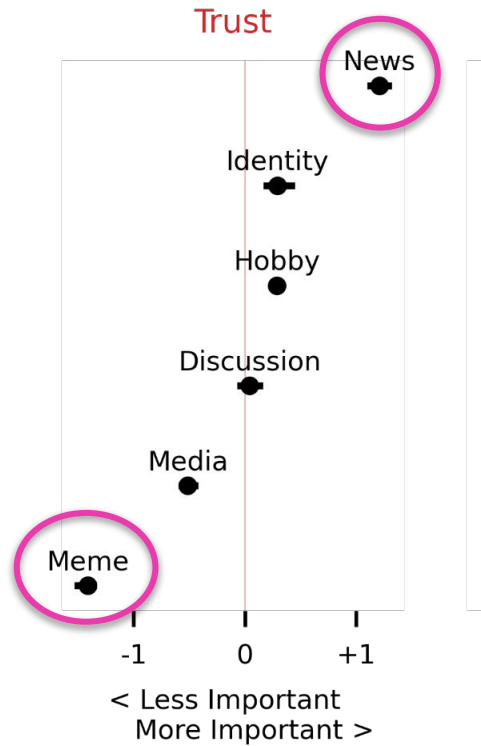
how would you like this value to *change*?

1.

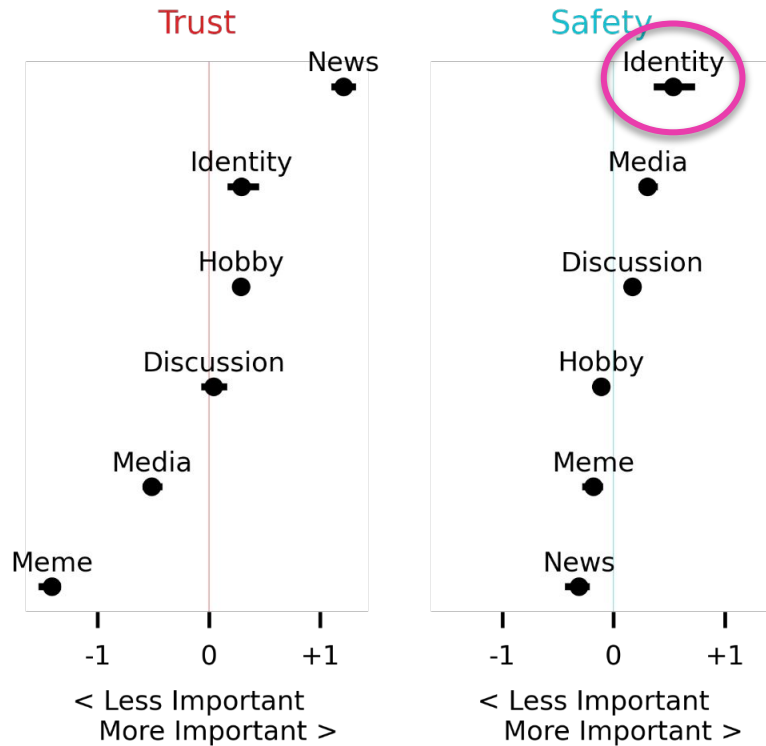
How do values vary
between communities?

How do values vary across different types of communities?

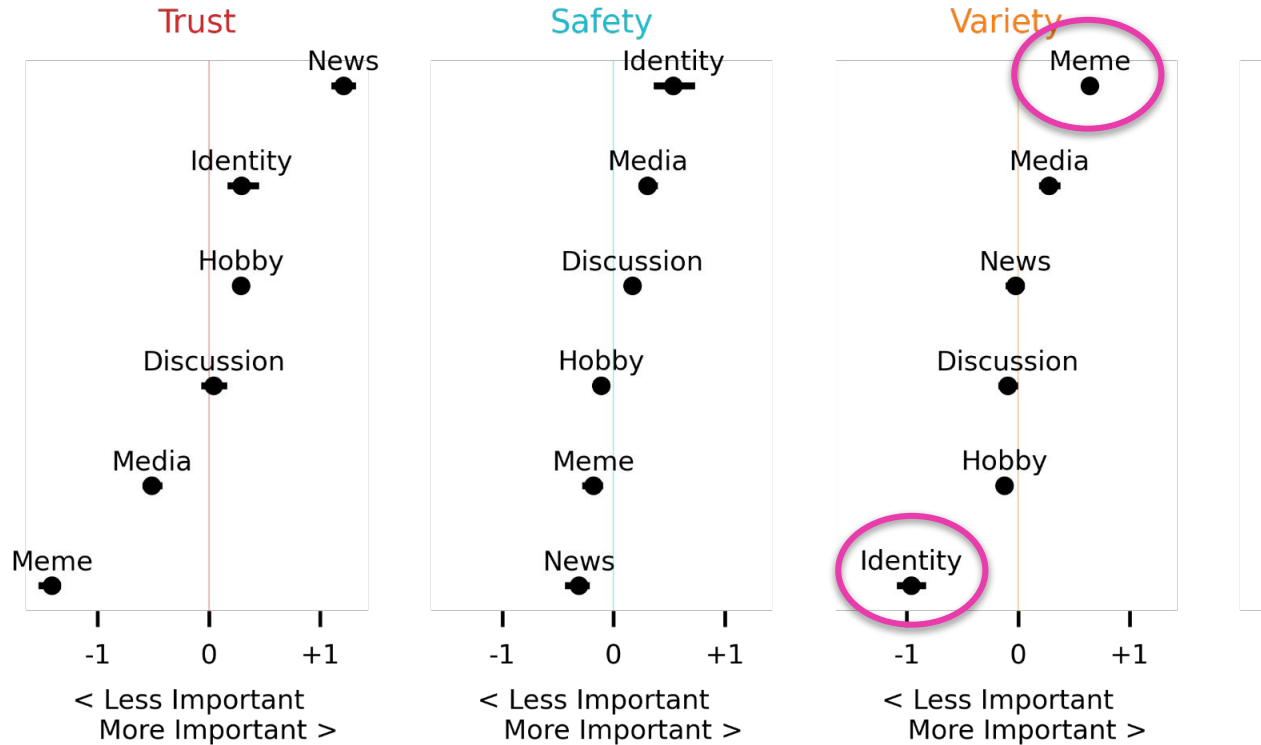
How do values vary across different types of communities?



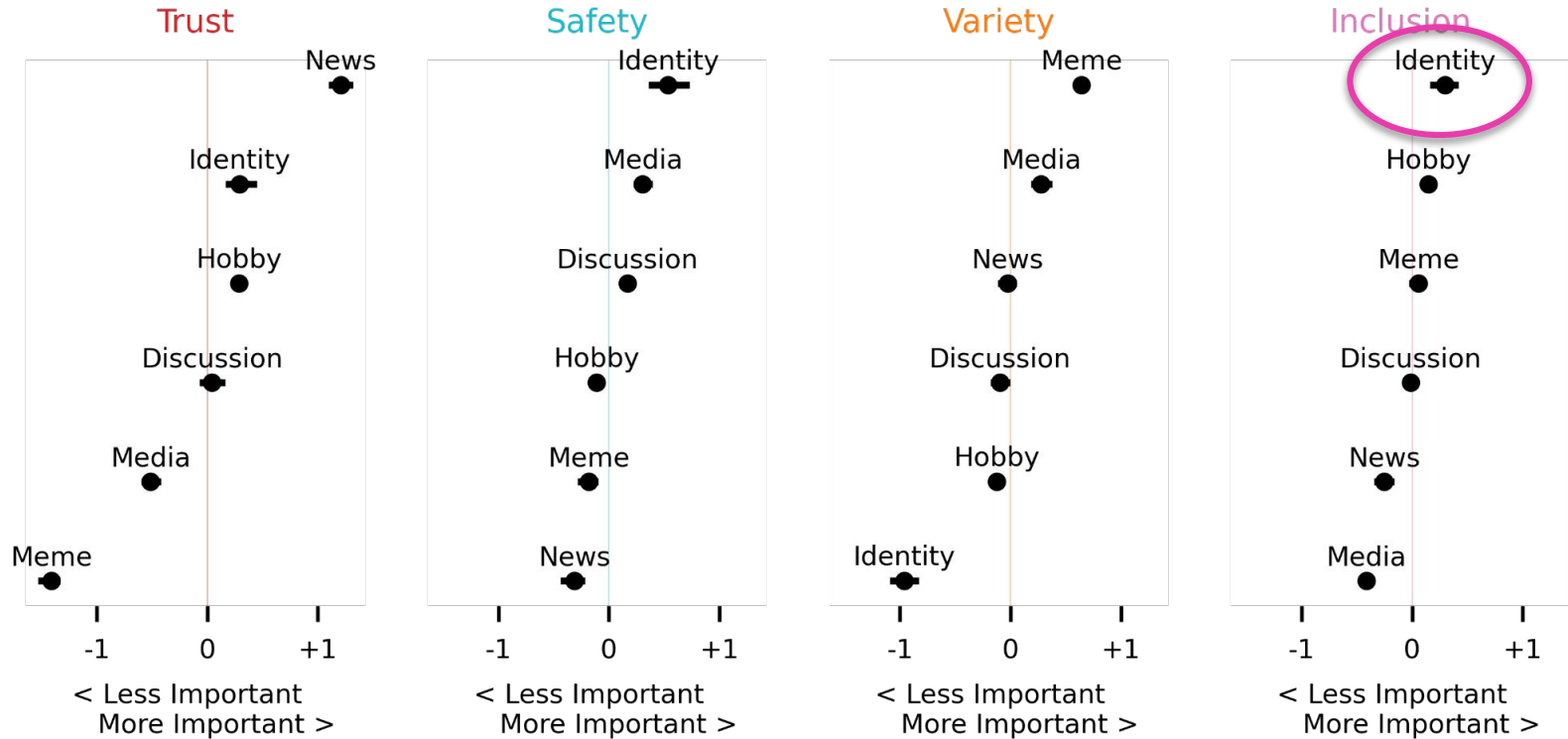
How do values vary across different types of communities?



How do values vary across different types of communities?



How do values vary across different types of communities?



There is no “one size fits all”
set of values.

There is lots of variance between communities!

There is still
substantial structure
in how values vary.

Prediction Tasks

14 easily quantifiable features

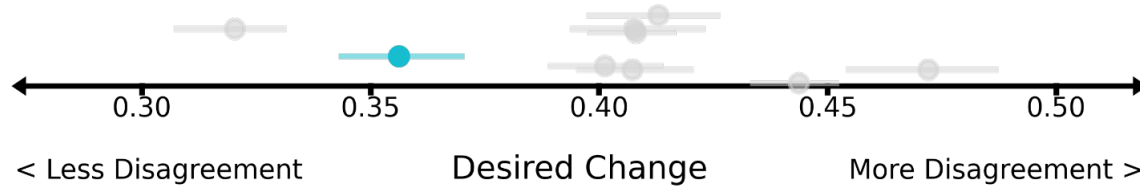
ROC AUC 0.667 averaged across all
tasks

Some values are easier to predict
than others! (ROC AUC > 0.9)

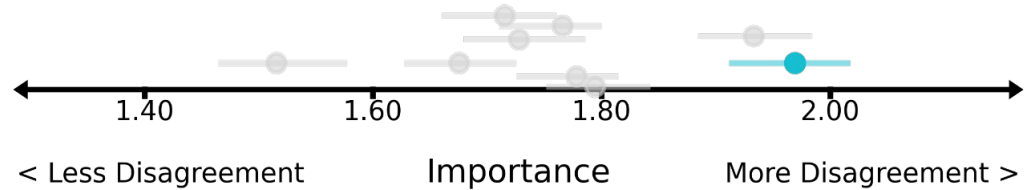
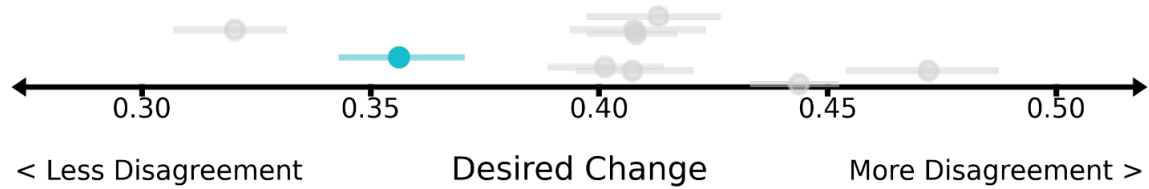
2.

How do values vary
within communities?

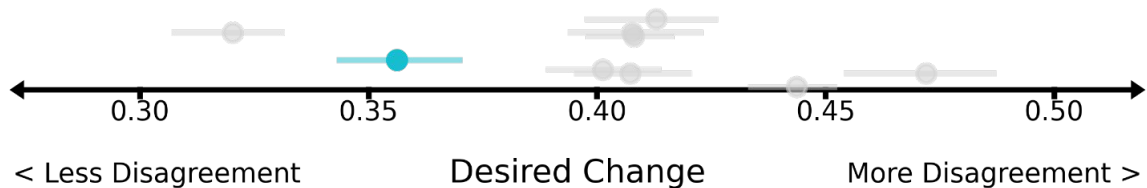
Disagreement over Safety within communities



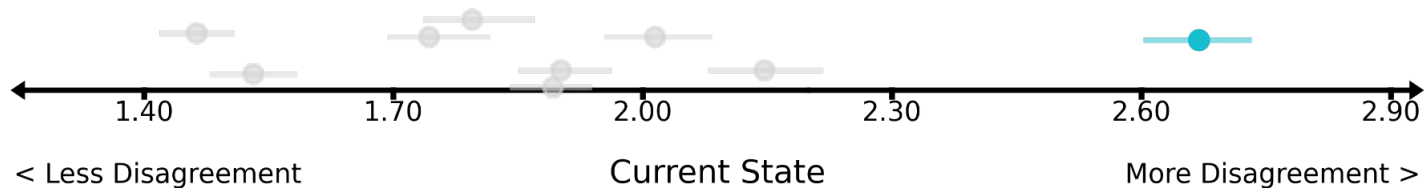
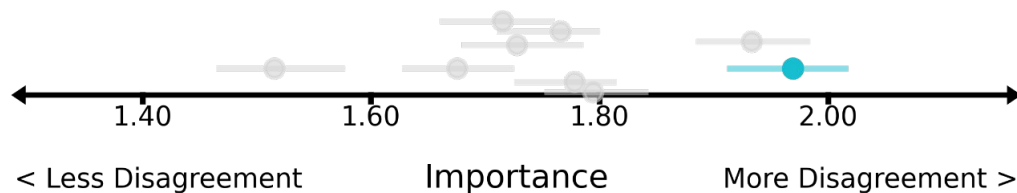
Disagreement over Safety within communities



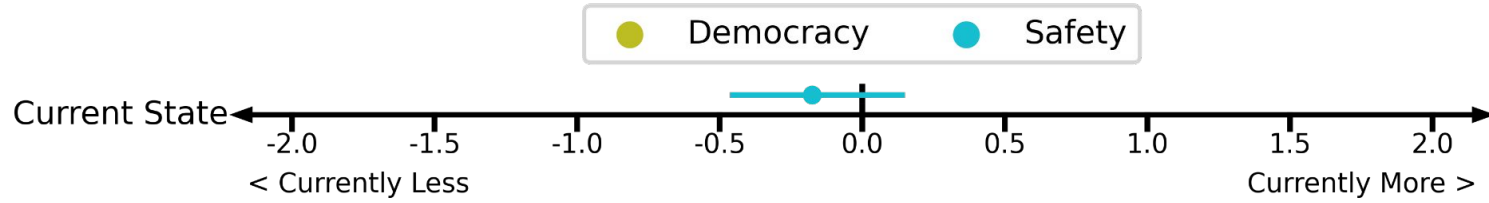
Disagreement over Safety within communities



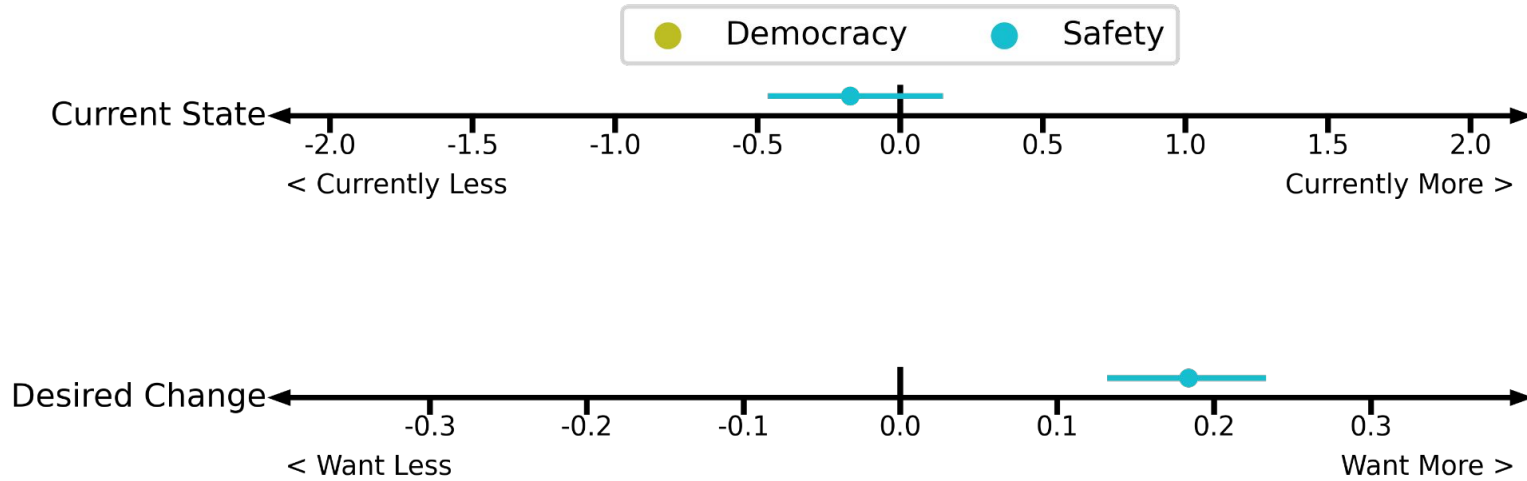
Safety is especially disagreed upon; we need to be careful to protect vulnerable minorities!



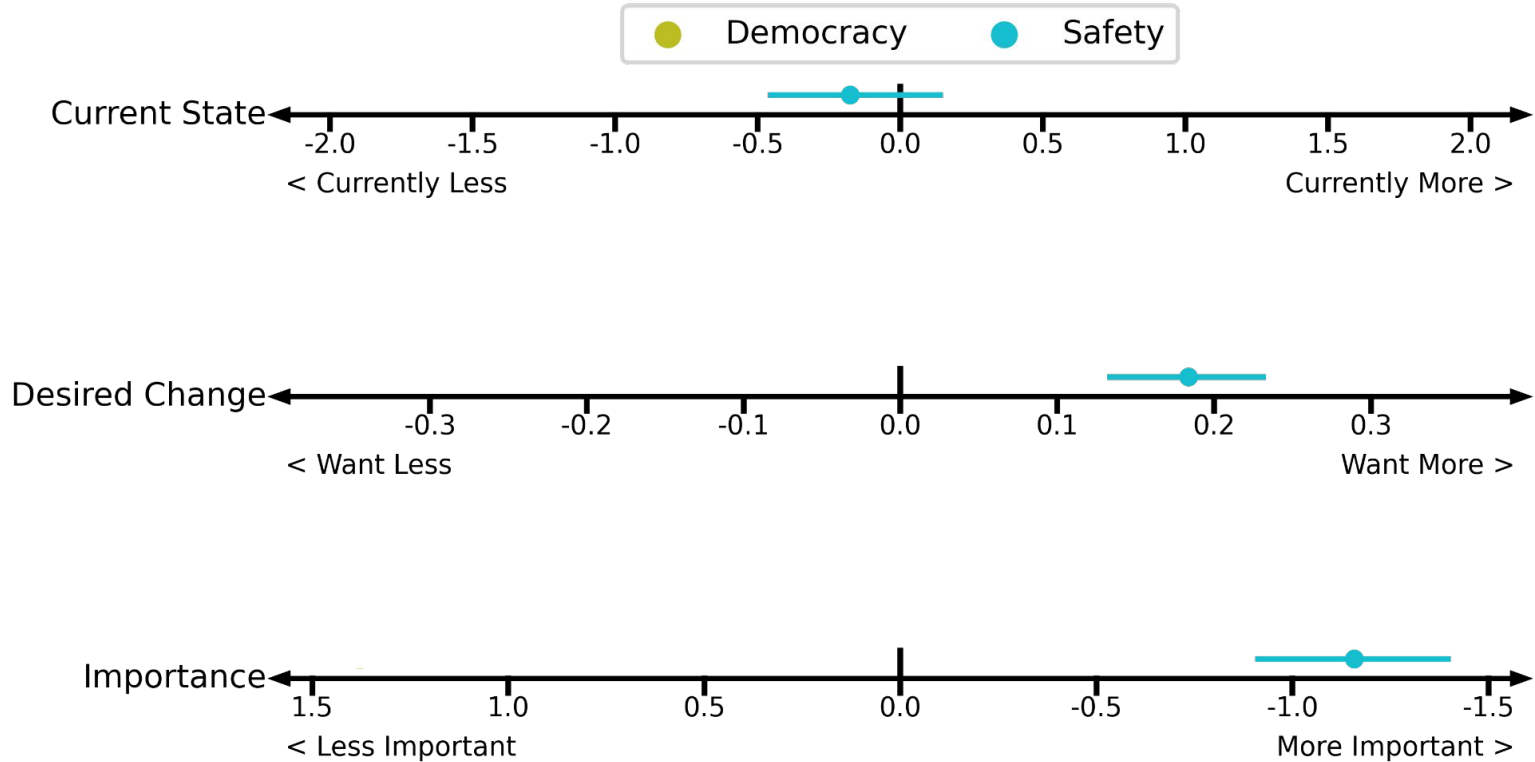
Moderator's values differ from non-moderator community members



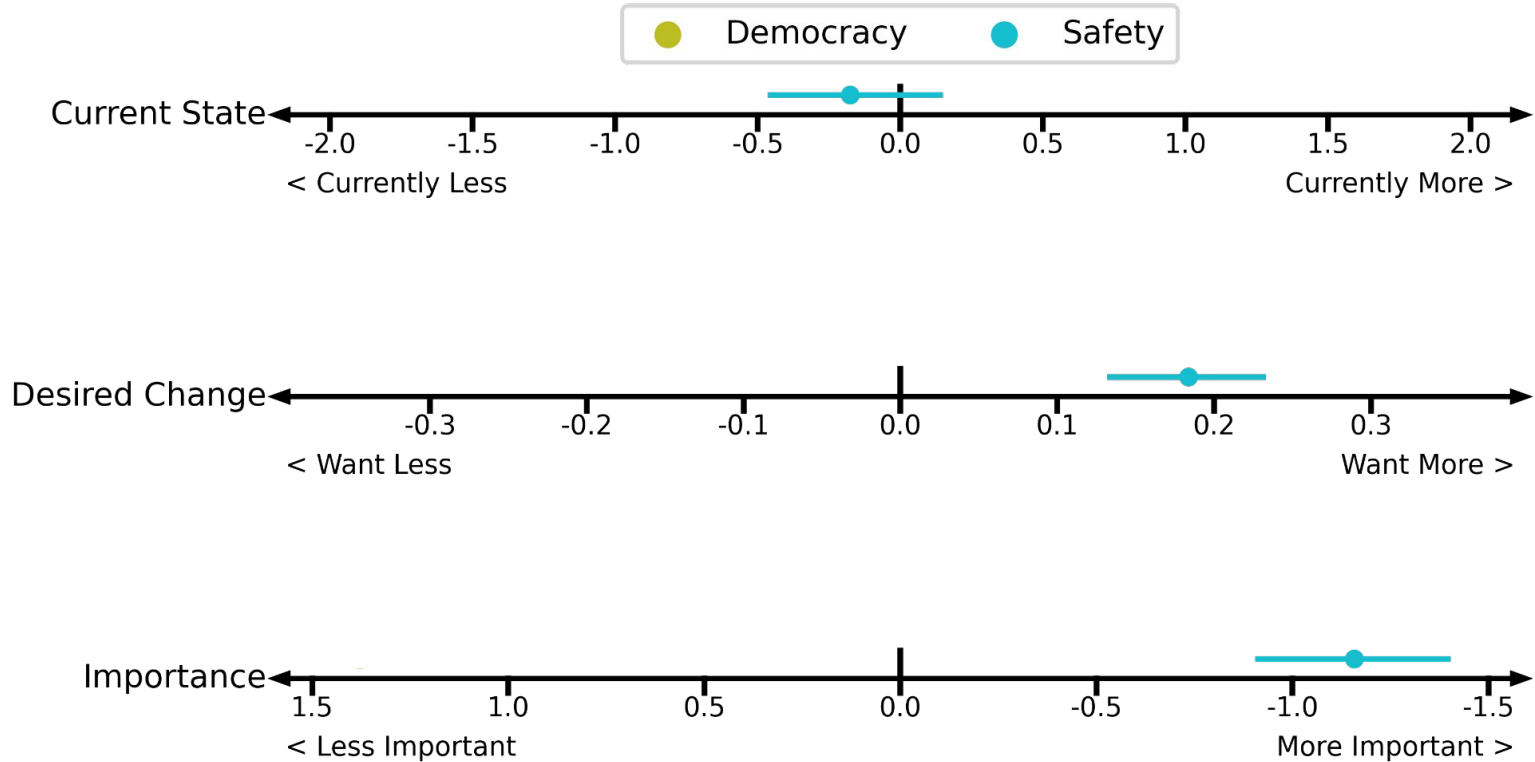
Moderator's values differ from non-moderator community members



Moderator's values differ from non-moderator community members



Moderator's values differ from non-moderator community members



1. Community Values

(ICWSM 2022 & ICWSM 2024)

There is lots of variety in values, and
no 'one size fits all'
approach to making
communities better!

1. Community Values

(ICWSM 2022 & ICWSM 2024)

2. Perceptions of Mods

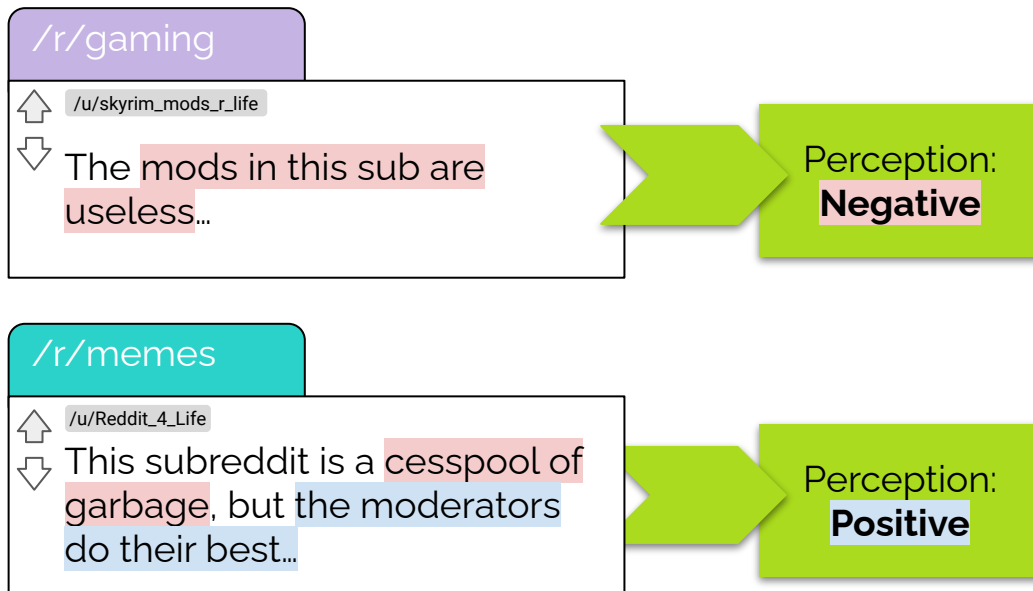
(arXiv 2024)

2. Perceptions of Mods

(arXiv 2024)

How can we measure communities' perceptions of mods to understand governance outcomes **at scale?**

Moderator discourse classifier measures community members' perceptions of moderators.



0. Prefiltering
1. Detection
2. Classification

Fine-tuned classifier with LLaMA2 and QLoRA **exceeds the performance of GPT-4**

Our approach enables measurement of governance outcomes at a massive scale

8.48K

unique subreddits

18+

months of data

1.89M

mod discourse
comments

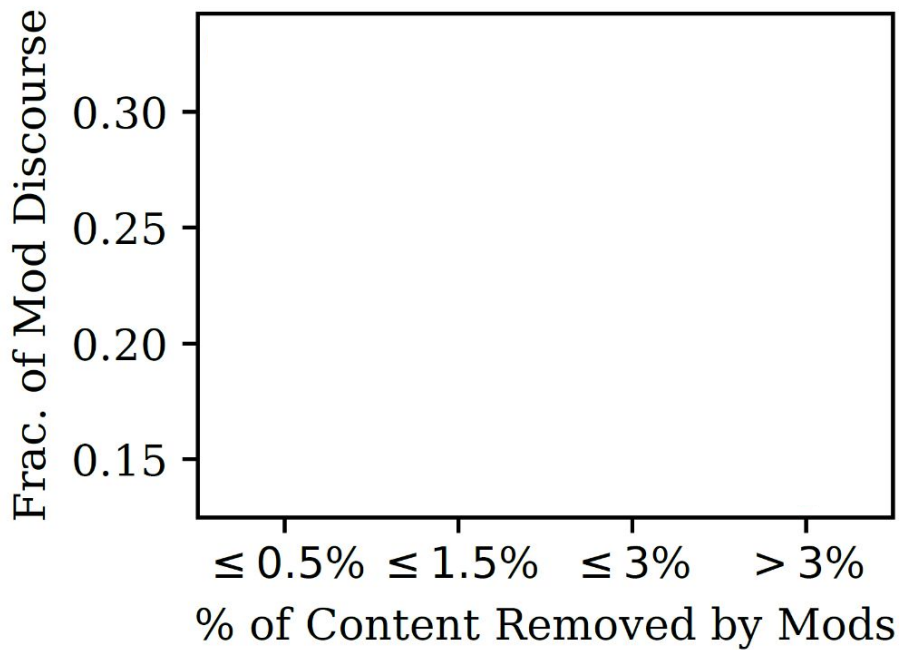
We make these data public!

RQ **1** What **rule enforcement strategies** are most associated with positive perceptions of moderators?

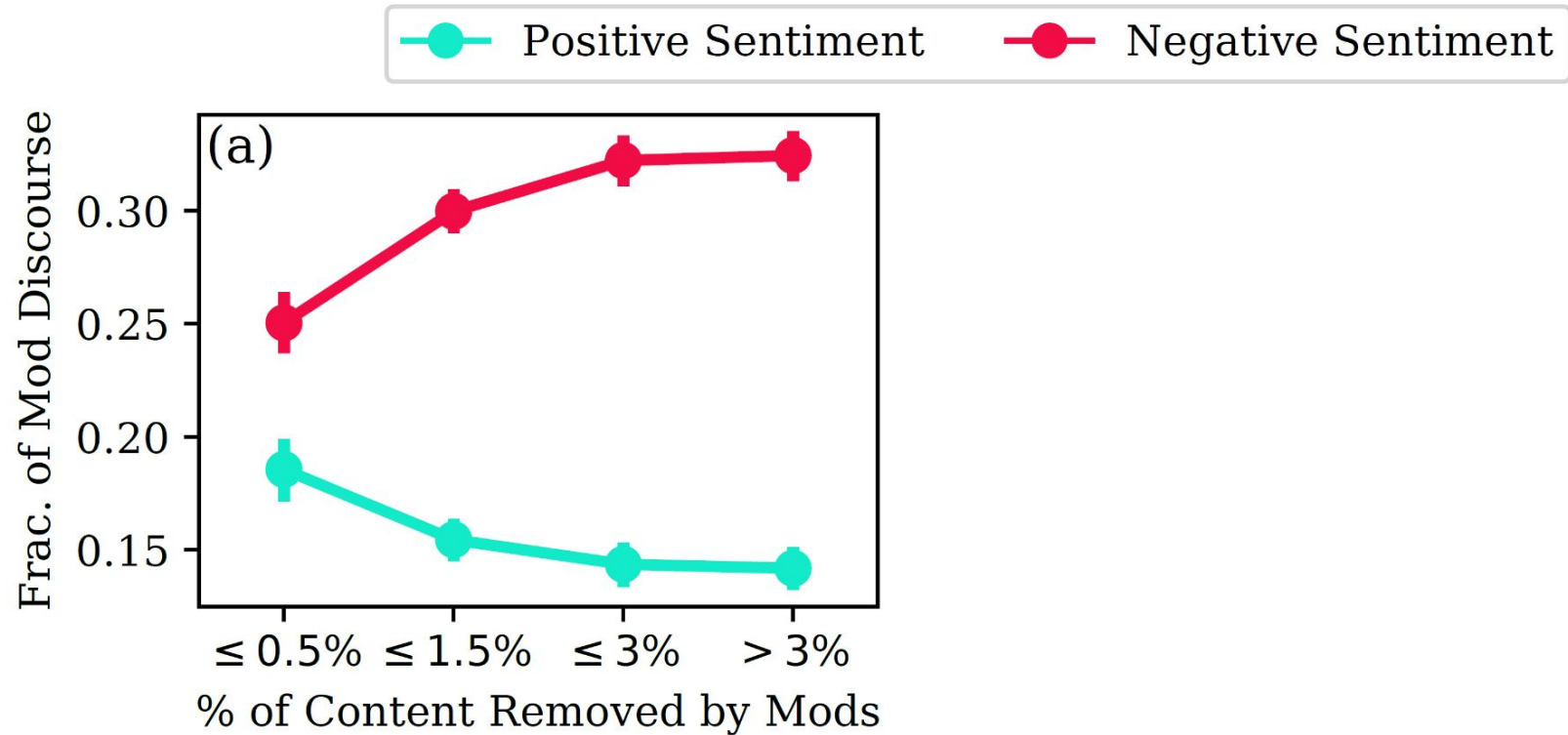
RQ **2** What **moderators and moderator teams** are most associated with positive perceptions of moderators?

RQ **1** What **rule enforcement strategies** are most associated with positive perceptions of moderators?

RQ **2** What **moderators and moderator teams** are most associated with positive perceptions of moderators?



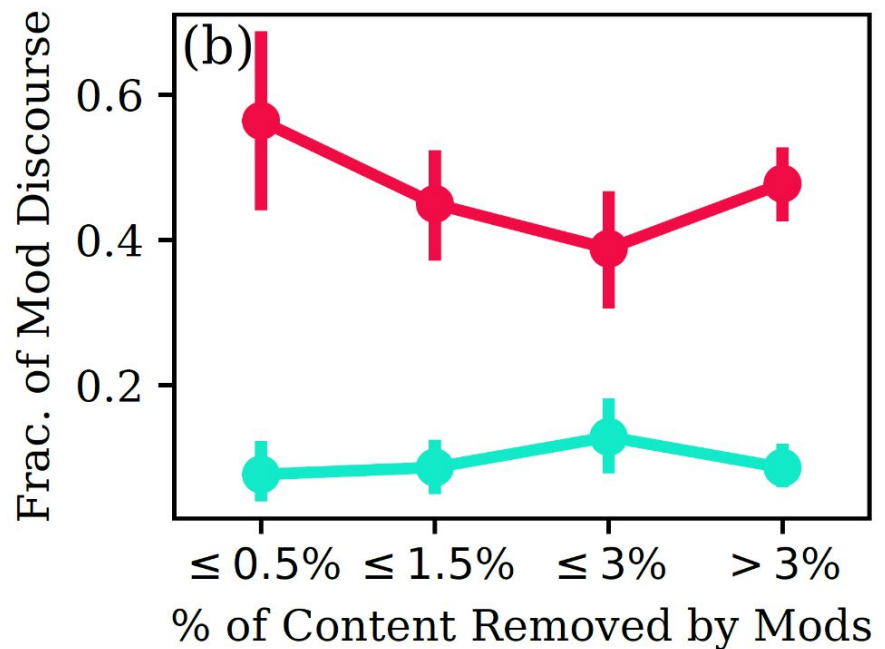
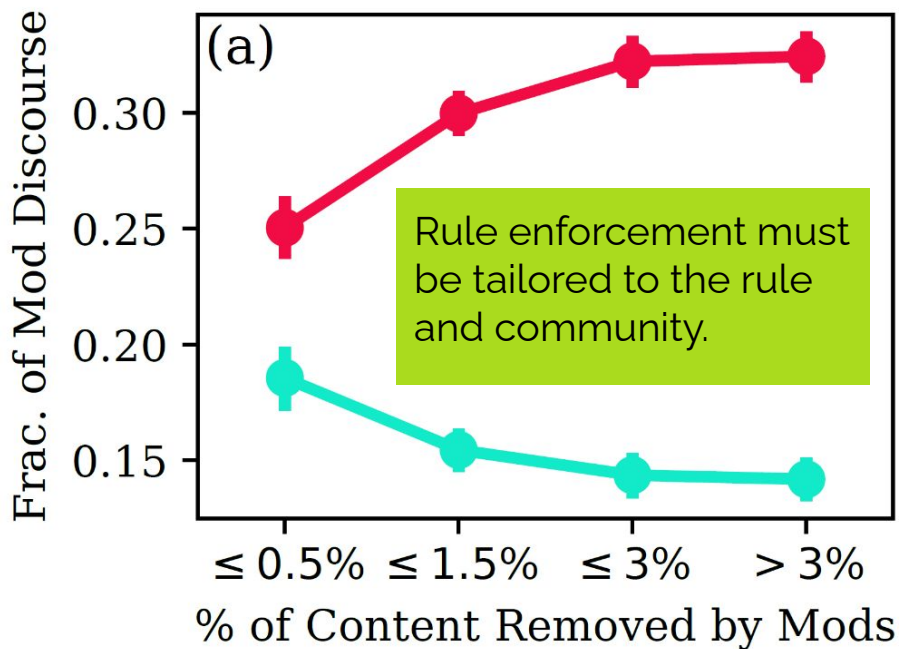
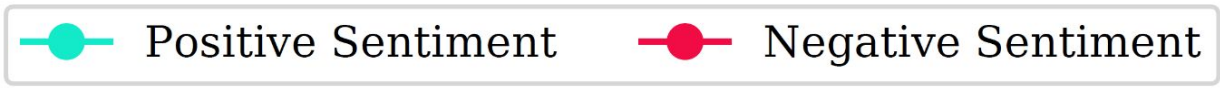
Stricter rule enforcement is associated with more negative perceptions of mods...



...except in News Sharing communities!

Other Communities

News Sharing Communities

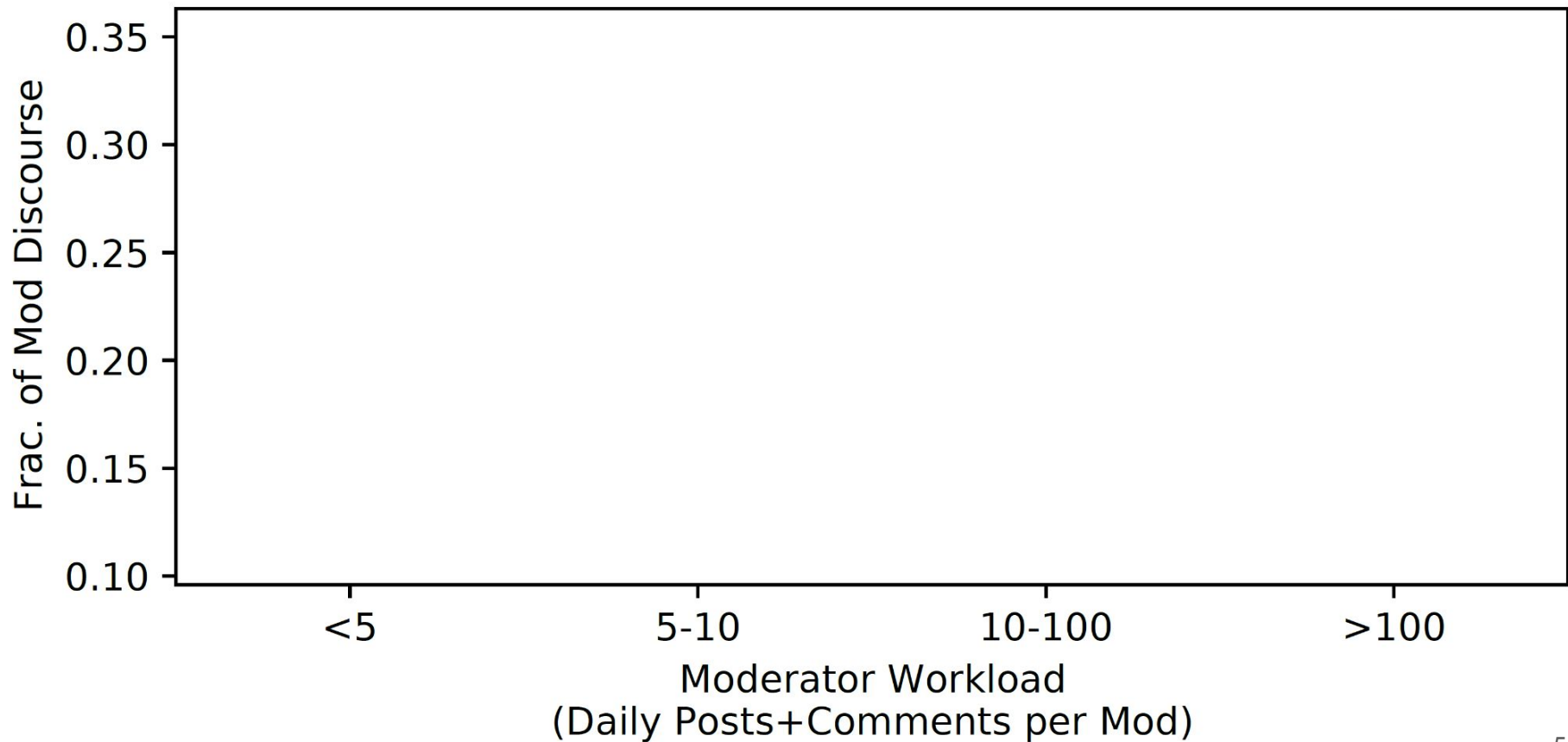


RQ **1**

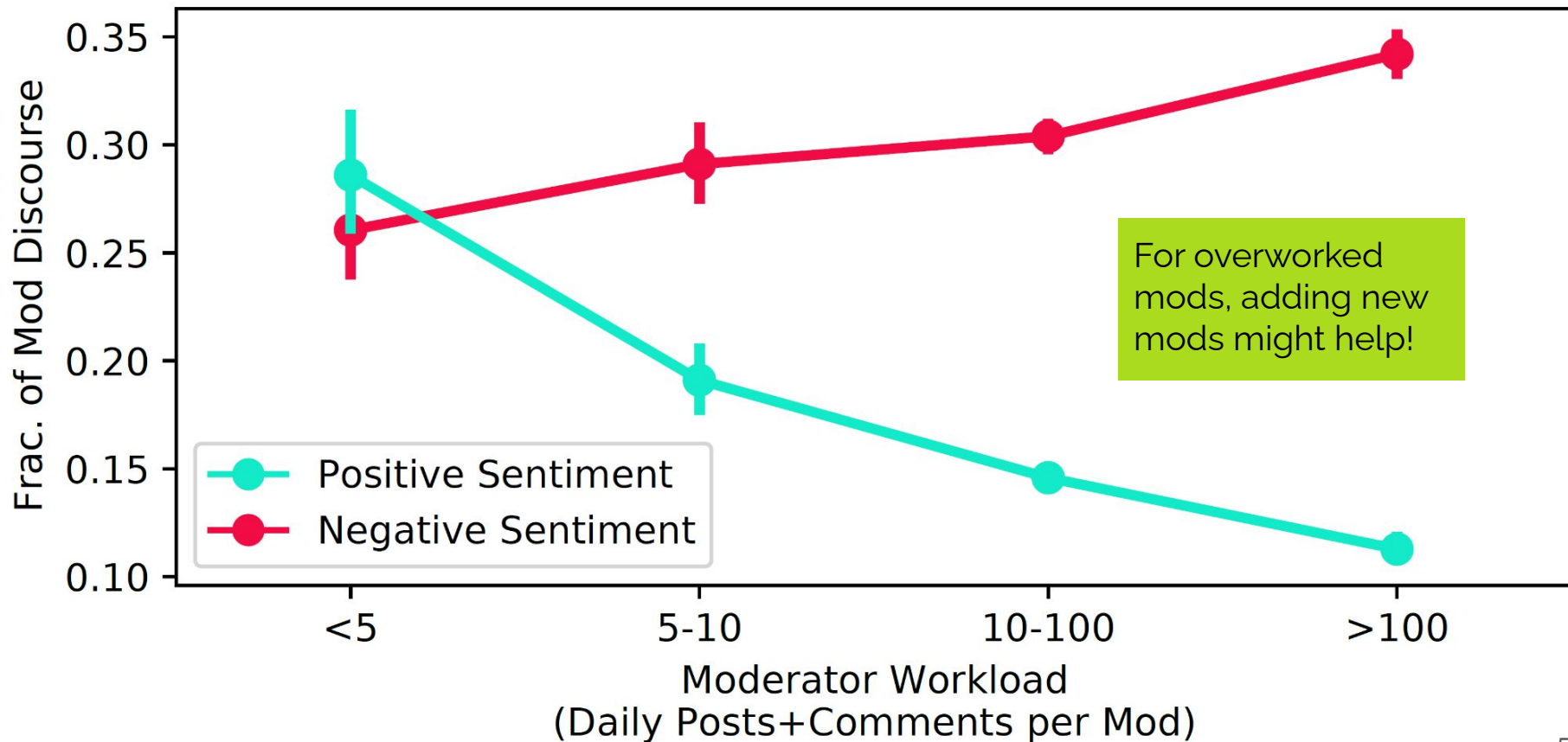
What **rule enforcement strategies** are most associated with positive perceptions of moderators?

RQ **2**

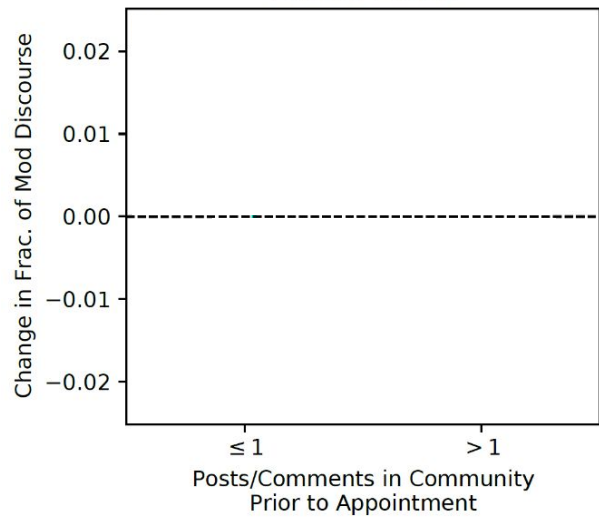
What **moderators and moderator teams** are most associated with positive perceptions of moderators?



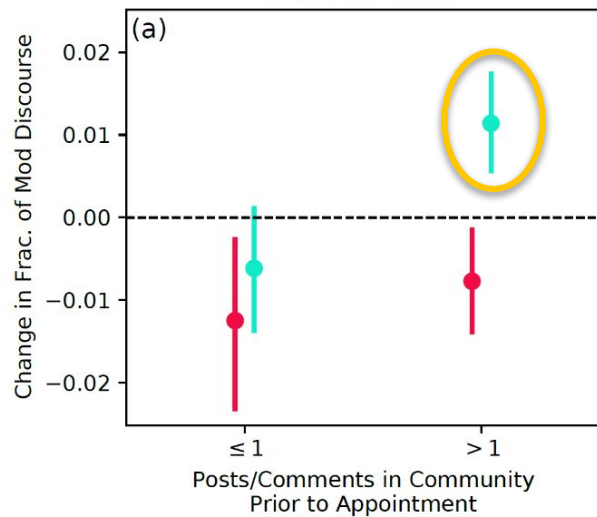
Subreddits with lower mod workloads have more positive perceptions of mods



Mod's Engagement Before Tenure

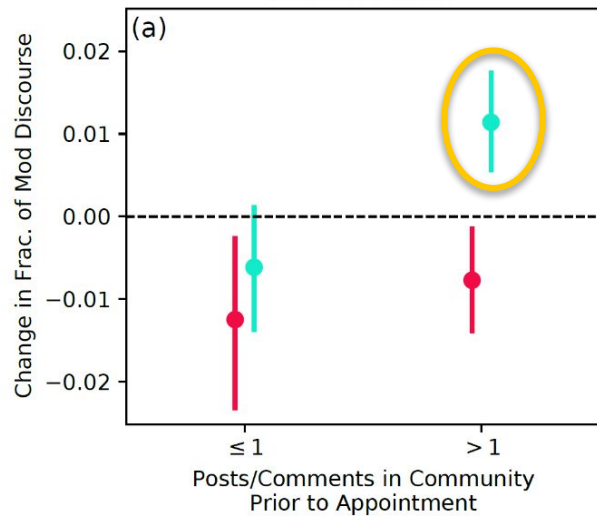


Mod's Engagement Before Tenure

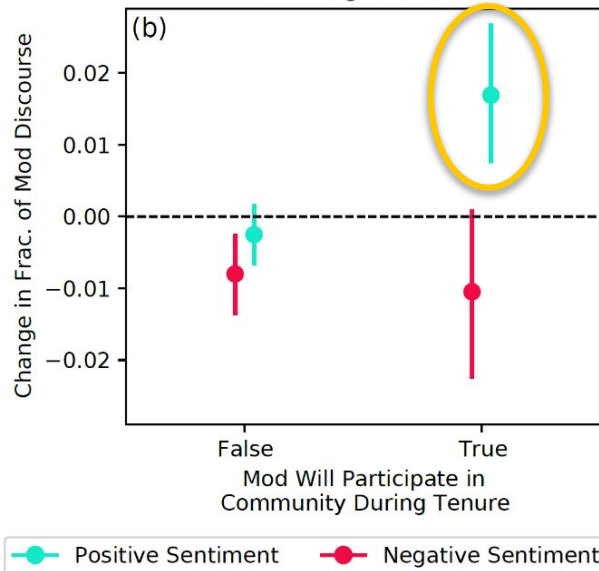


—●— Positive Sentiment —●— Negative Sentiment

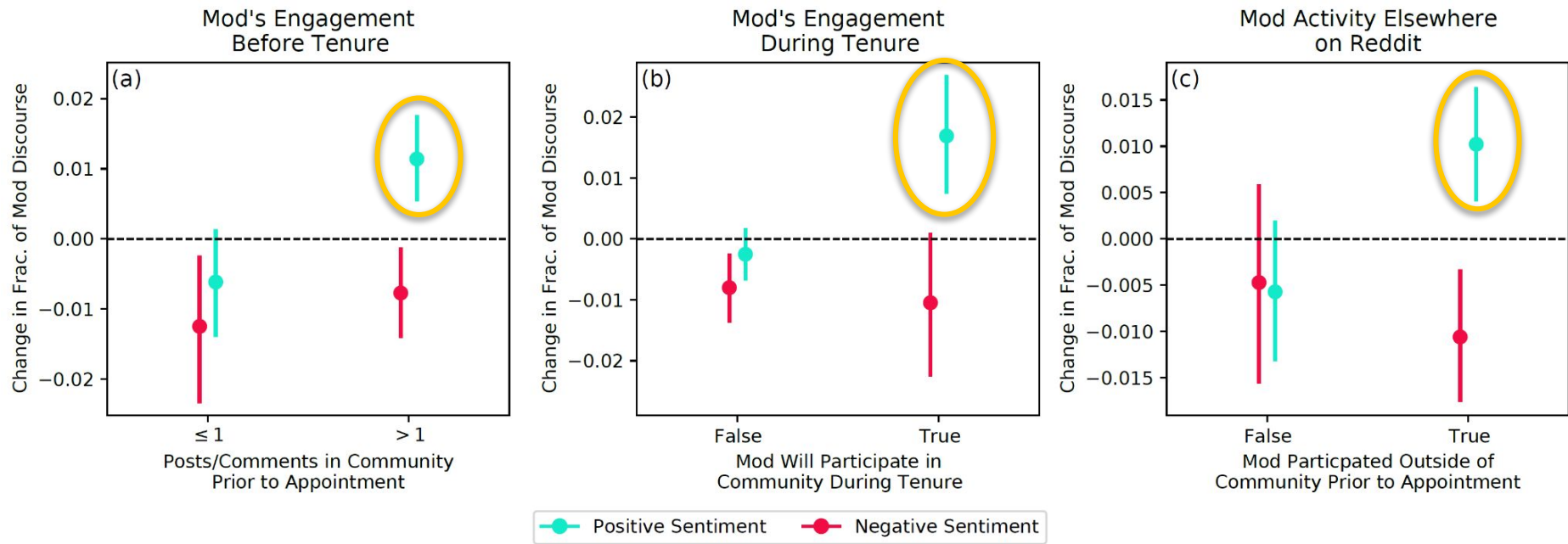
Mod's Engagement
Before Tenure



Mod's Engagement
During Tenure



Mods who are engaged before and during their tenure, and are active in other subreddits, are most associated with improved community perceptions



2. Perceptions of Mods

(arXiv 2024)

- Communities' perceptions of mods are a useful signal that can be measured.
- Moderators seem most positively received when they
 - enforce rules appropriately for their community,
 - are not overworked, and
 - are engaged with the community before and during their tenures

1. Community Values

(ICWSM 2022 & ICWSM 2024)

2. Perceptions of Mods

(arXiv 2024)

3. News Sharing Behavior

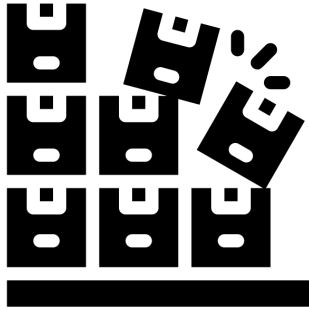
( ICWSM 2021)

3. News Sharing Behavior

( ICWSM 2021)

What community affordances improve the trustworthiness of news shared online?

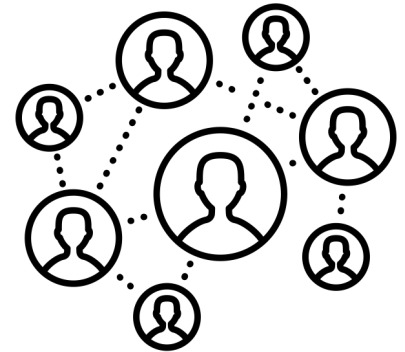
Understanding how news is shared is hard



scale of news
sharing is huge



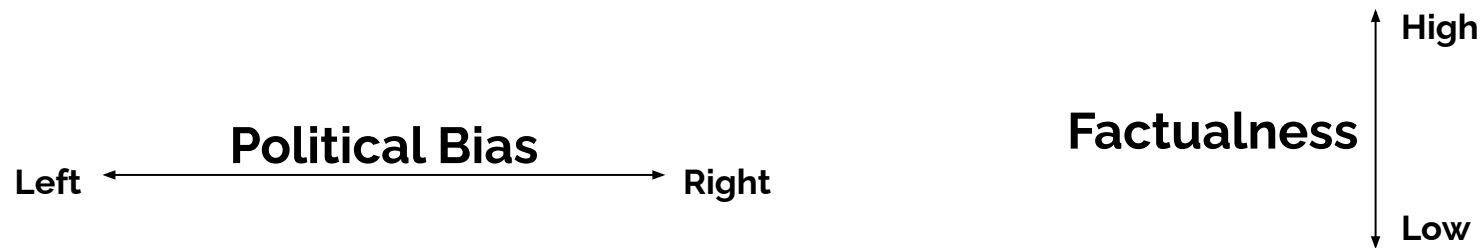
difficult to measure
bias and factualness



distribution involves many
people and communities

Our approach: use fact checking to label millions of links to news sources on reddit.

Labeling news sources: Media Bias/Fact Check



Labeling at the news source level enables massive scale that wouldn't be possible otherwise

Labeling news sources: Example

↑
21
↓



Why cats may have more to teach us about living the good life than Socrates (cbc.ca)

submitted 5 days ago by 

4 comments share save hide report [+c]

CBC News

Bias: **Left-Center**

Factualness: **High**

Largest study of news sharing on reddit to date

2015-

2019
date range

35M
news links

1.3M
users

135K
communities

Dataset is publicly available!

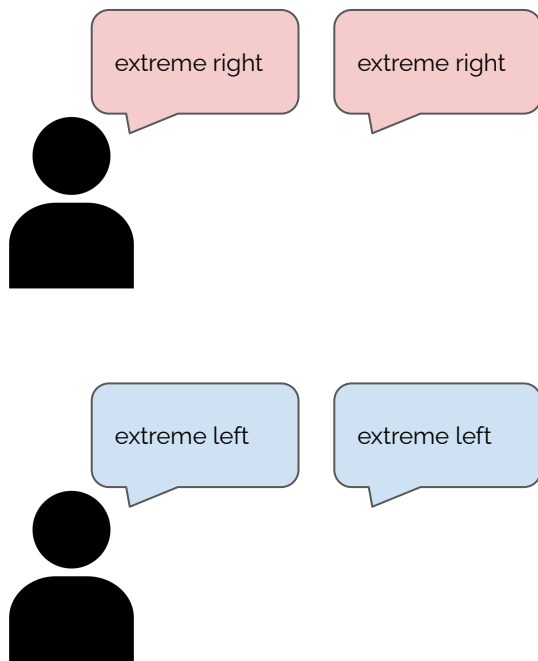
RQ **1** How is news shared differently across different communities?

RQ **2** How effective are existing curation and amplification mechanisms?

RQ **1** How is news shared differently across different communities?

RQ **2** How effective are existing curation and amplification mechanisms?

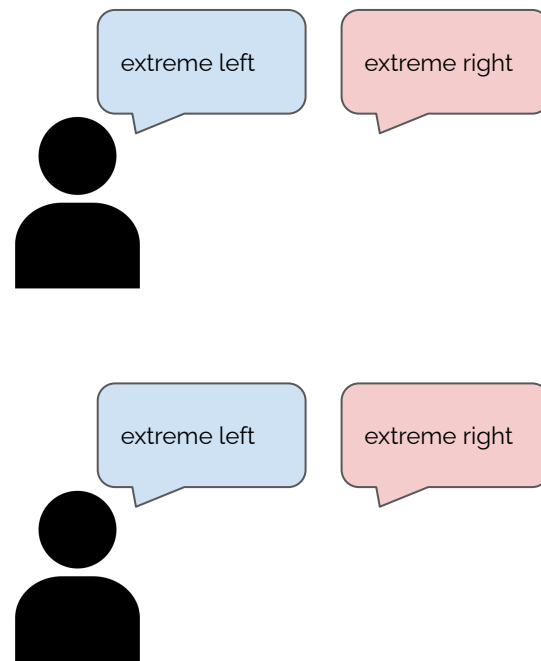
/r/we_love_news



group diversity:
high

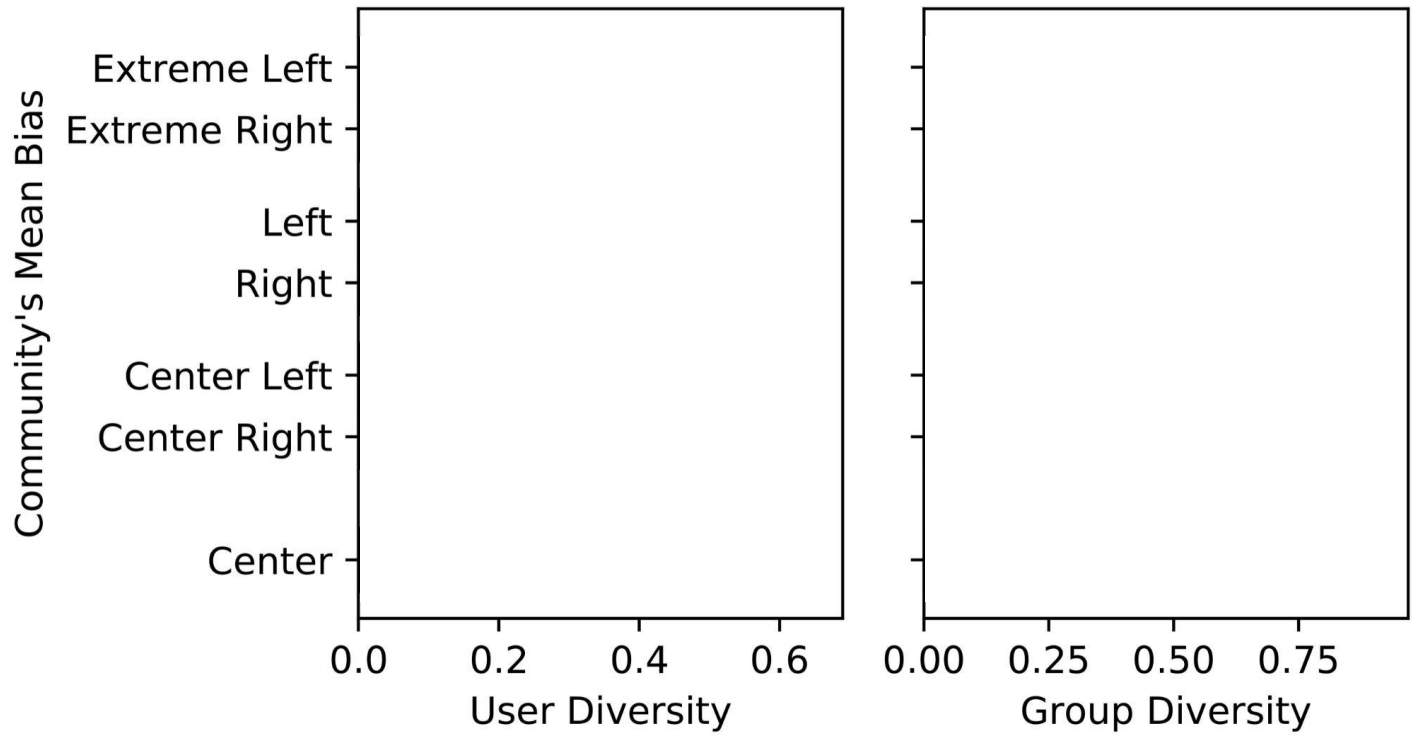
user diversity:
low

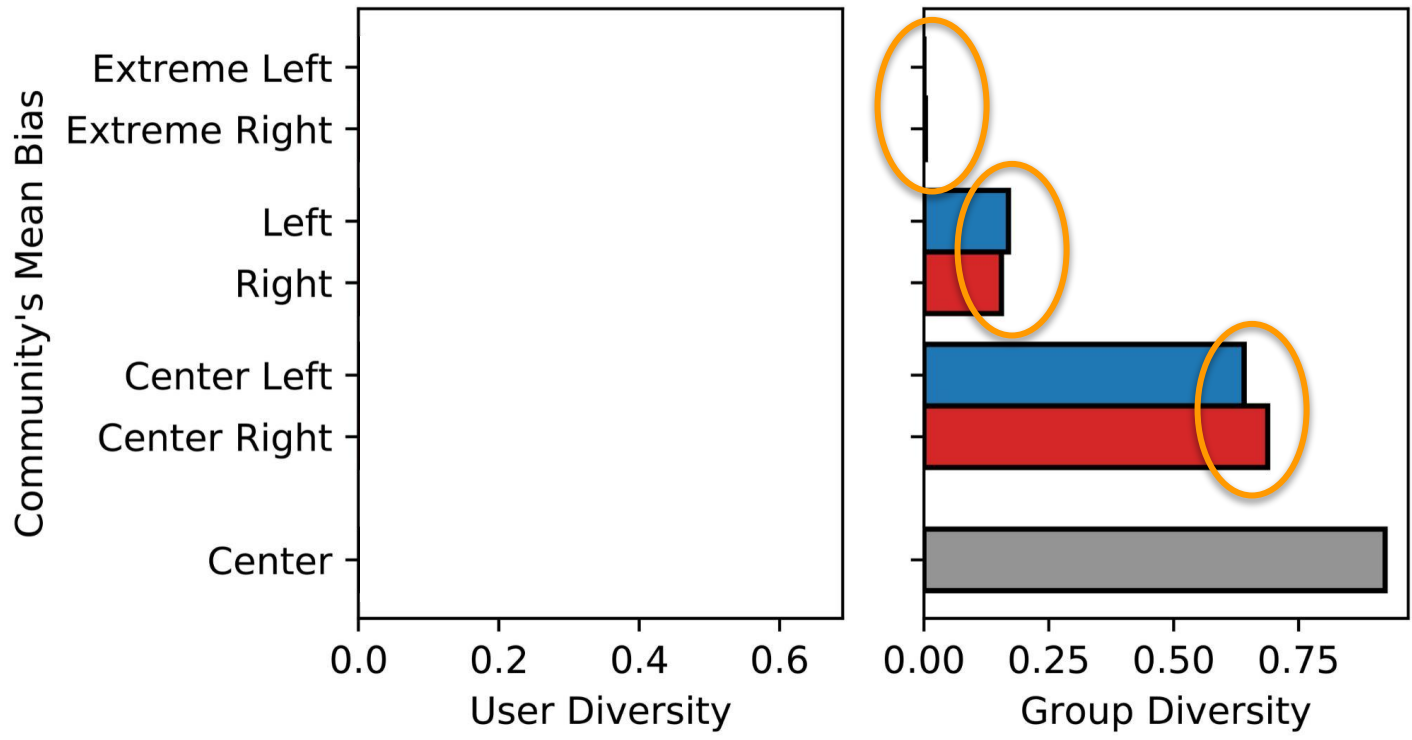
/r/news_is_cool



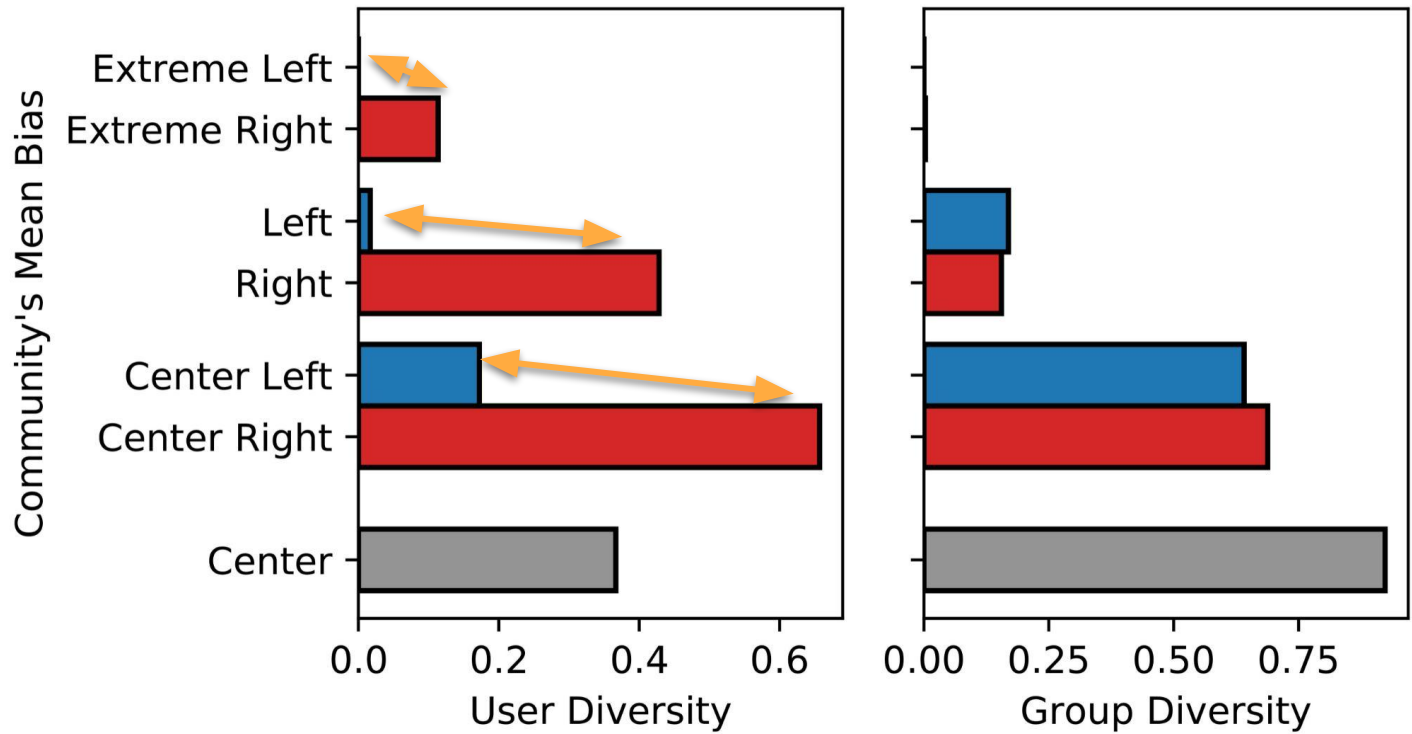
group diversity:
low

user diversity:
high





Right-leaning communities have more user diversity (and therefore more overall diversity of content)

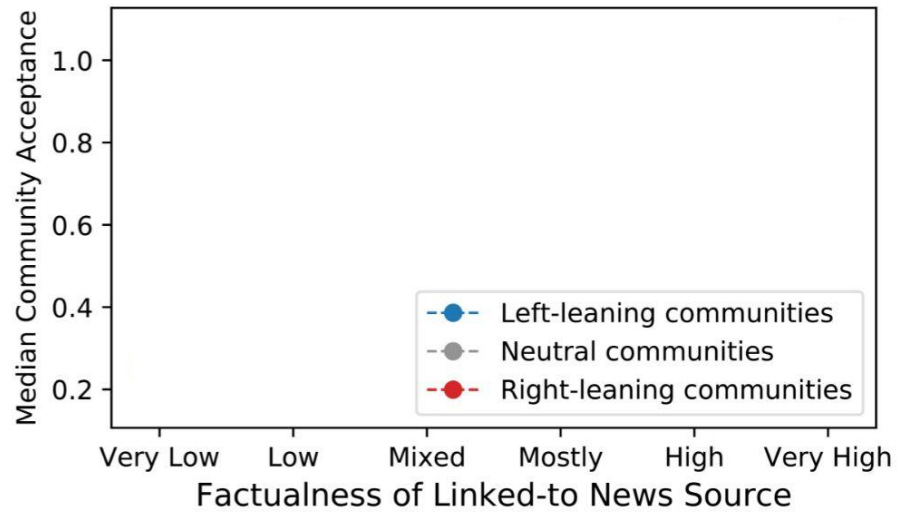
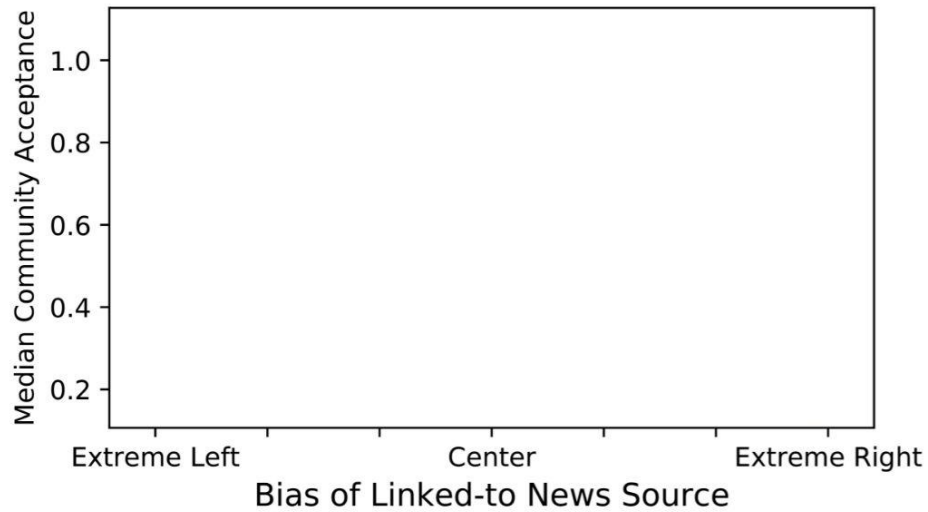


RQ **1**

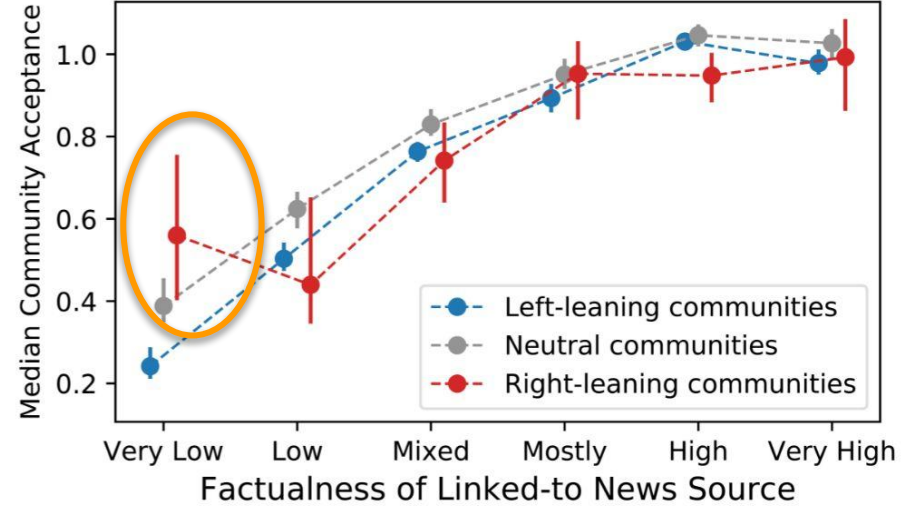
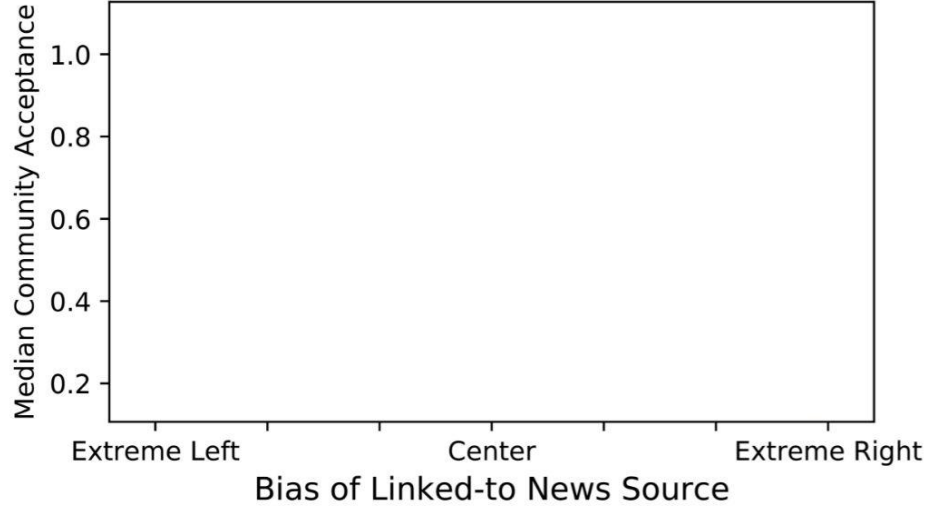
How is news shared differently across different communities?

RQ **2**

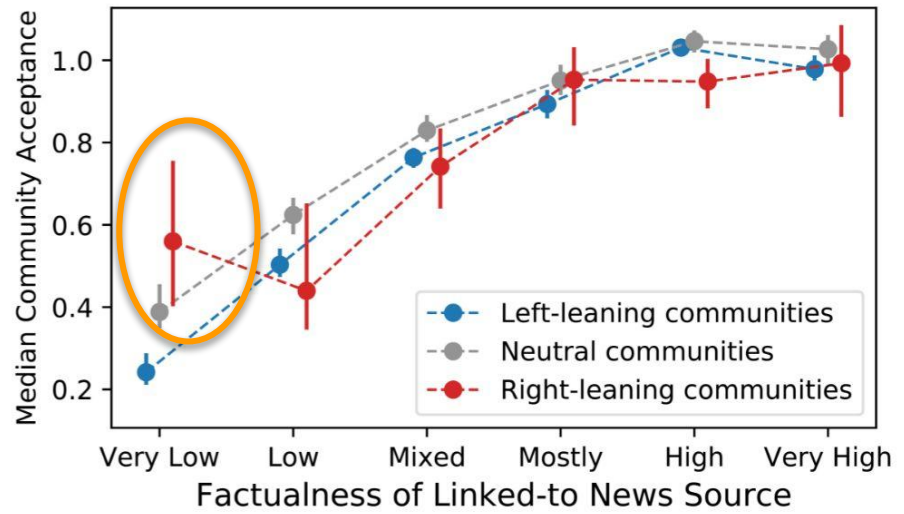
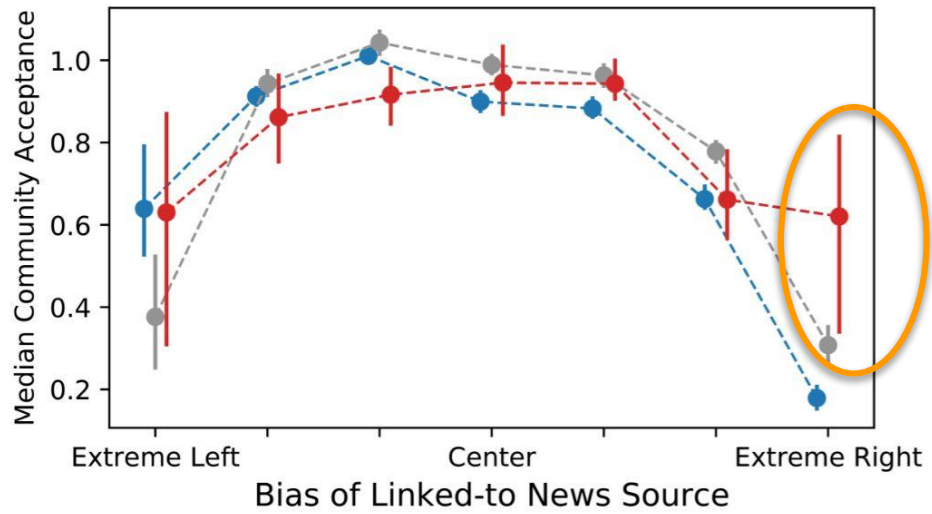
How effective are existing curation and amplification mechanisms?



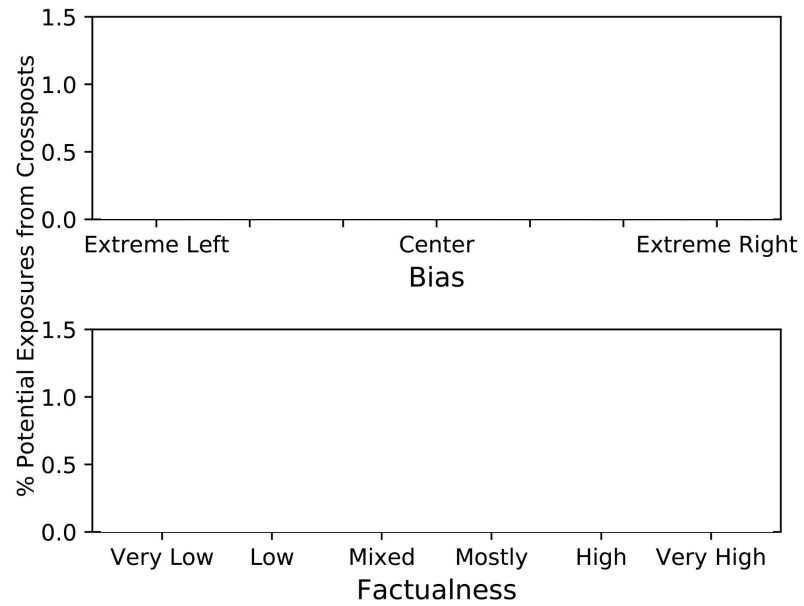
Right-leaning communities are more accepting of very low factual content



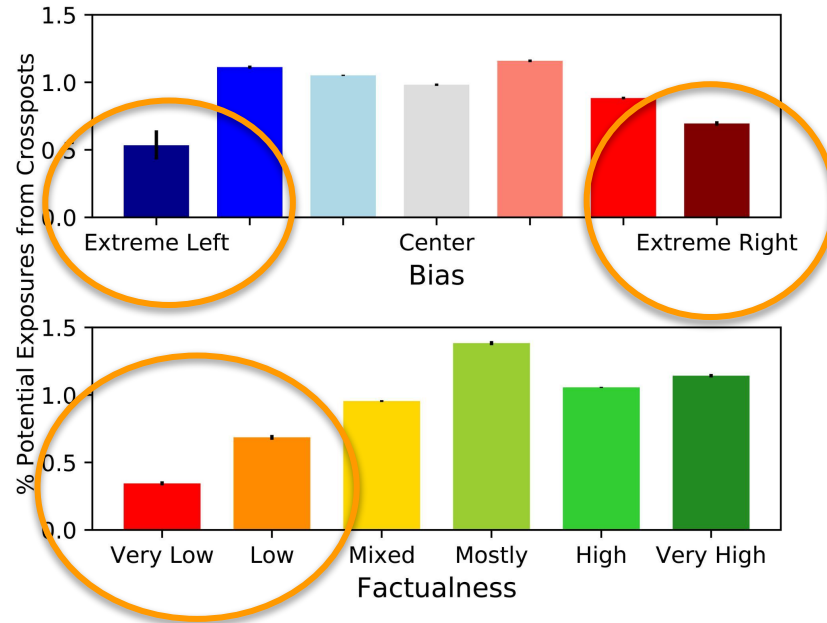
Right-leaning communities are more accepting of very low factual content and extreme right content



Current voting behavior generally reduces the relative visibility of biased and low factual content



Biased and low factual content receives fewer exposures from crossposts



Existing amplification behavior reduces the visibility of biased and low factual content

3. News Sharing Behavior

( ICWSM 2021)

- Existing curation and amplification mechanisms serve to reduce the visibility of biased and low factual news.
- Affordances like voting and crossposting are promising mechanisms to empower community members to improve the quality of content shared in their communities.

1. Community Values

(ICWSM 2022 & ICWSM 2024)

2. Perceptions of Mods

(arXiv 2024)

3. News Sharing Behavior

(ICWSM 2021)

4. Causal Inference Methods

(ICWSM 2022)

4. Causal Inference Methods

(ICWSM 2022)

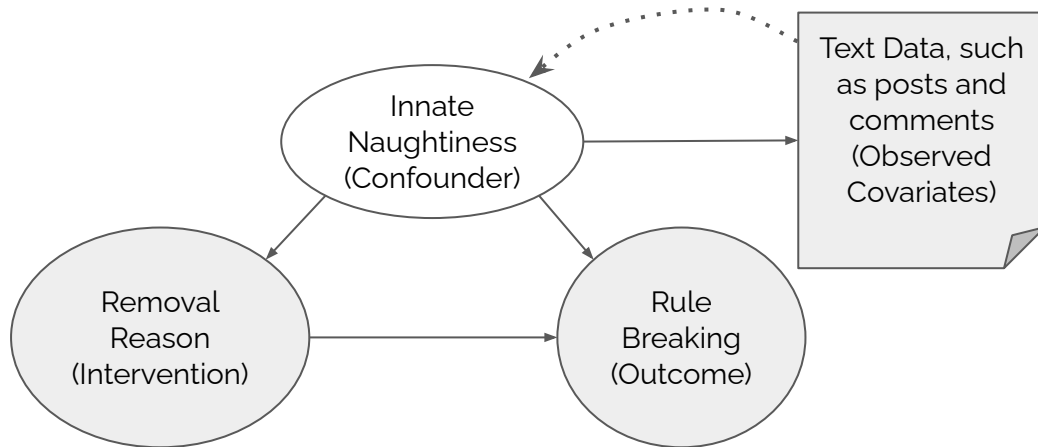
How can we use text data from online communities to produce **robust causal inferences?**

Causal Inference

Quantifying the impact of some **intervention** on some **outcome** (often we try and estimate the **Average Treatment Effect**, or ATE)

Observational Studies

Conducted without control over the intervention, so it's challenging to adjust for **confounding factors**.

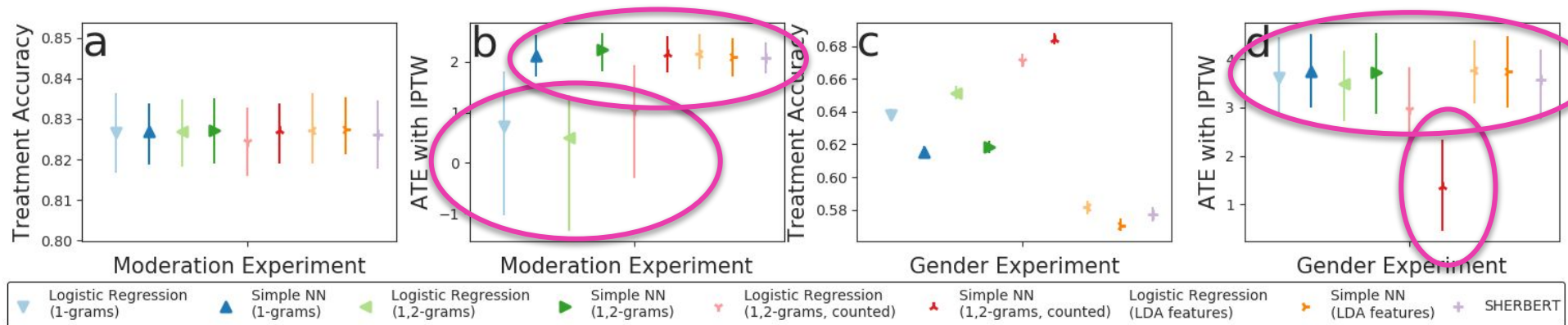


Problem

Causal inference methods are extremely difficult to evaluate because **there is almost never ground truth** - we don't know the right answer!

It's difficult to know which method is best!

In the real world, methods disagree!

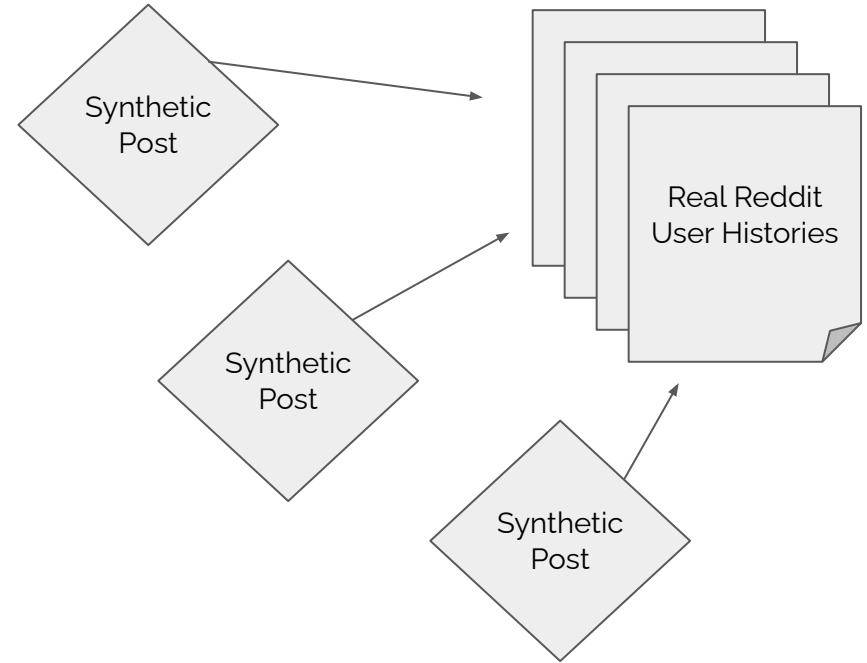


Our Contribution: a **Semi-Synthetic Evaluation Framework**

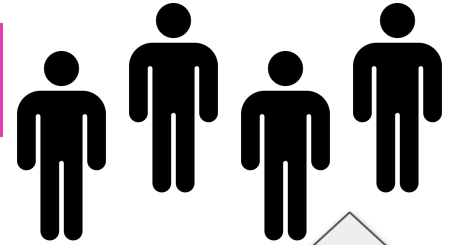
Our framework consists of **five tasks**.

Tasks are created by inserting **synthetic posts** into **real reddit user histories**.

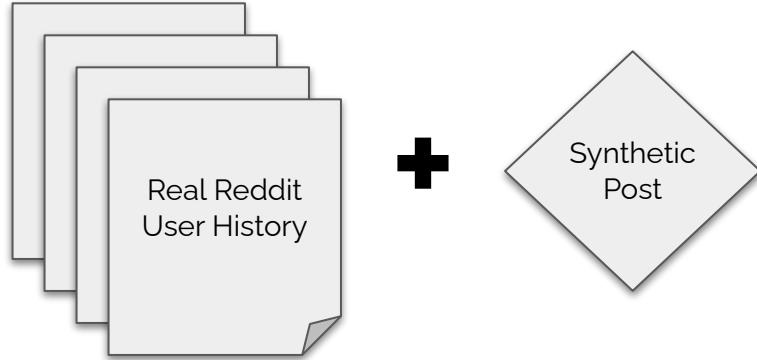
Using a **known probability distribution** lets us have a known ground truth ATE.



An Example: **Signal Intensity Task**



Class 1



Append 1 Synthetic Post

Class 2



Append 10 Synthetic Posts

Why Semi-Synthetic Evaluation?

Best of Both Worlds

Balances realism of real world data with the known ground truth of synthetic tasks.

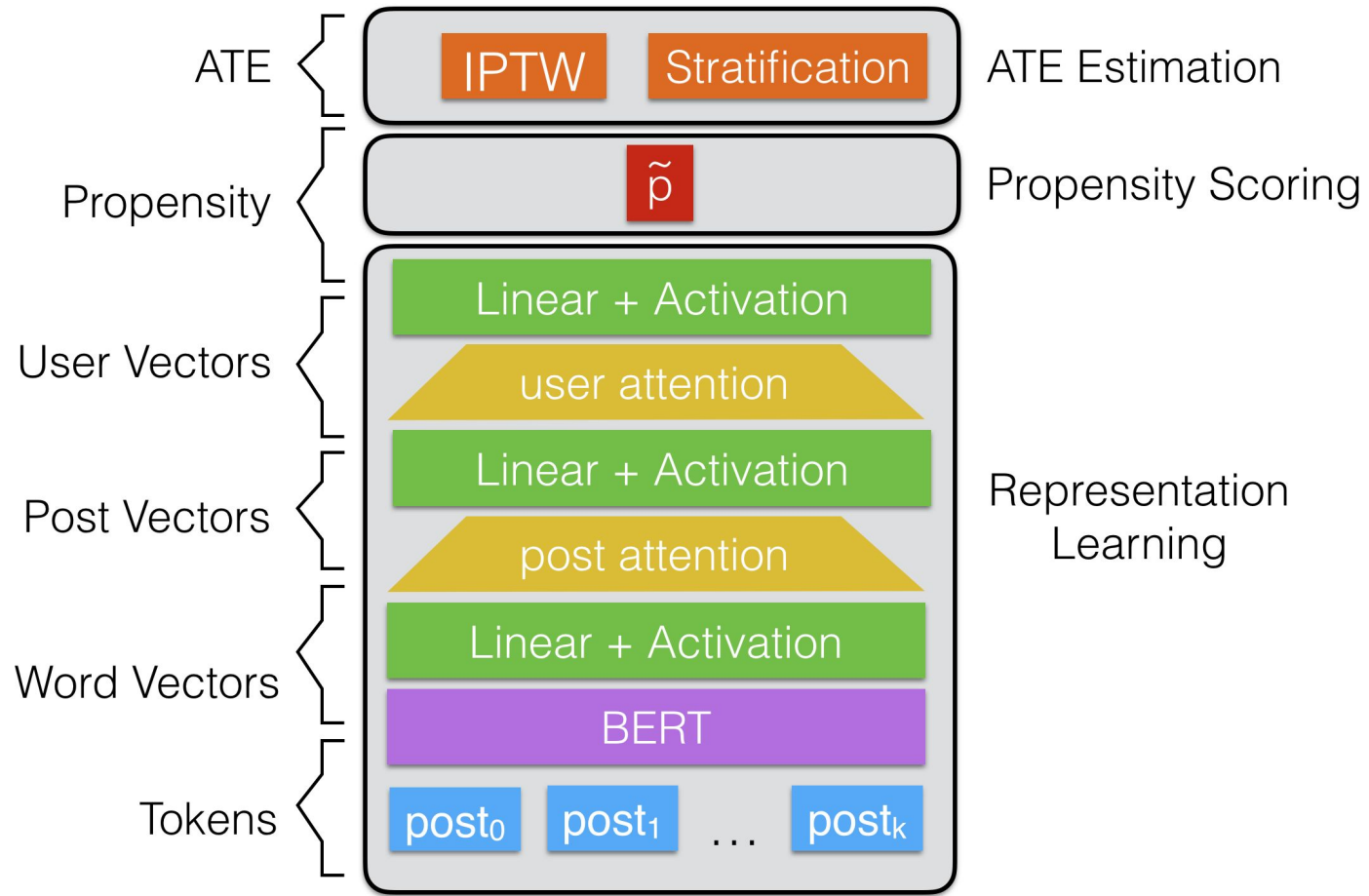
Widely Used

Semi-synthetic evaluation frameworks are widely used in other domains.

Public Framework

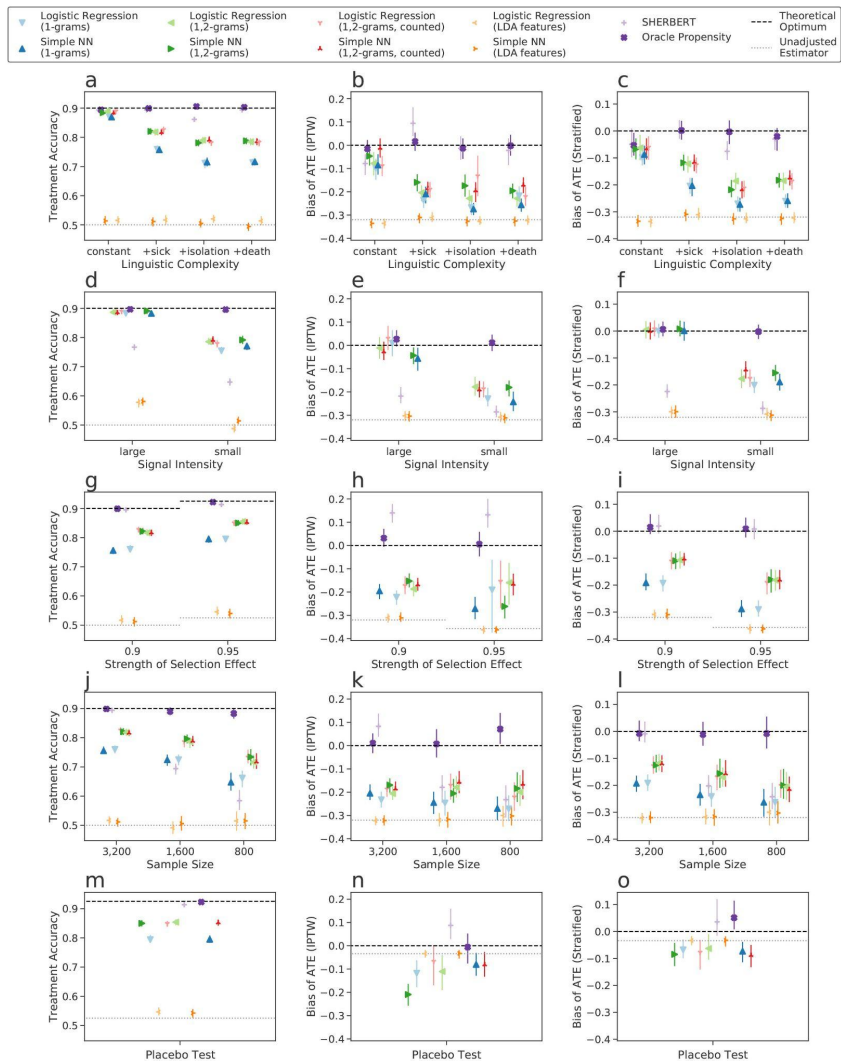
We make our framework public, so that we can all use it to evaluate and improve causal inference methods!

SHERBERT Method



Results

We evaluate every commonly used method, conducting over 600 experiments.



4. Causal Inference Methods

(ICWSM 2022)

- Transformers generally perform best, *however*
- Transformers struggle with counting and limited amounts of data.
- High accuracy often reflects strong selection effects, not low ATE bias.
- Many methods fail a placebo test!
- SHERBERT leverages recent NLP developments and performs well.
- **Principled evaluation of causal inference methods is really important!**
- **We make our framework and code public!**

1. Community Values

(ICWSM 2022 & ICWSM 2024)

2. Perceptions of Mods

(arXiv 2024)

3. News Sharing Behavior

(ICWSM 2021)

4. Causal Inference Methods

(ICWSM 2022)

5. Ongoing & Proposed Work

(Ongoing)

5. Ongoing & Proposed Work

(Ongoing)

Some personal and 2-body considerations but hoping to apply to faculty positions this Fall (2024, ~6 months from now).

Deploy interventions to demonstrate real impact

Mentor students and build teaching skills

Use and expand upon causal inference methods

Publish at venues other than ICWSM

Develop and demonstrate experience with LLMs

Honest feedback very much appreciated!

**Spring
2024**

**Summer
2024**

**Autumn
2024**

**Winter
2025**

Spring 2025

General
Exam

Photocritique Coaching

How can we improve discussion quality?

Rules Deep Dive (with REU Leon)

What rules and enforcement is best?

Community Dashboards

Do better-informed mods make for better communities?

Constructed Observational Study

Evaluate causal inference methods in a real world setting.

Defense

We are here!

Ongoing Projects



Photocritique Coaching

How can we improve discussion quality?

Rules Deep Dive (with REU Leon)

What rules and enforcement is best?

Community Dashboards

Do better-informed mods make for better communities?

Constructed Observational Study

Evaluate causal inference methods in a real world setting.

Photocritique Coaching

(Ongoing)

Problem

Providing high quality discussion and feedback online is hard!

Solution

We are building a system to measure high quality comments, and coach community members to improve their discussion.

Photocritique Coaching

(Ongoing)

RQ **1**

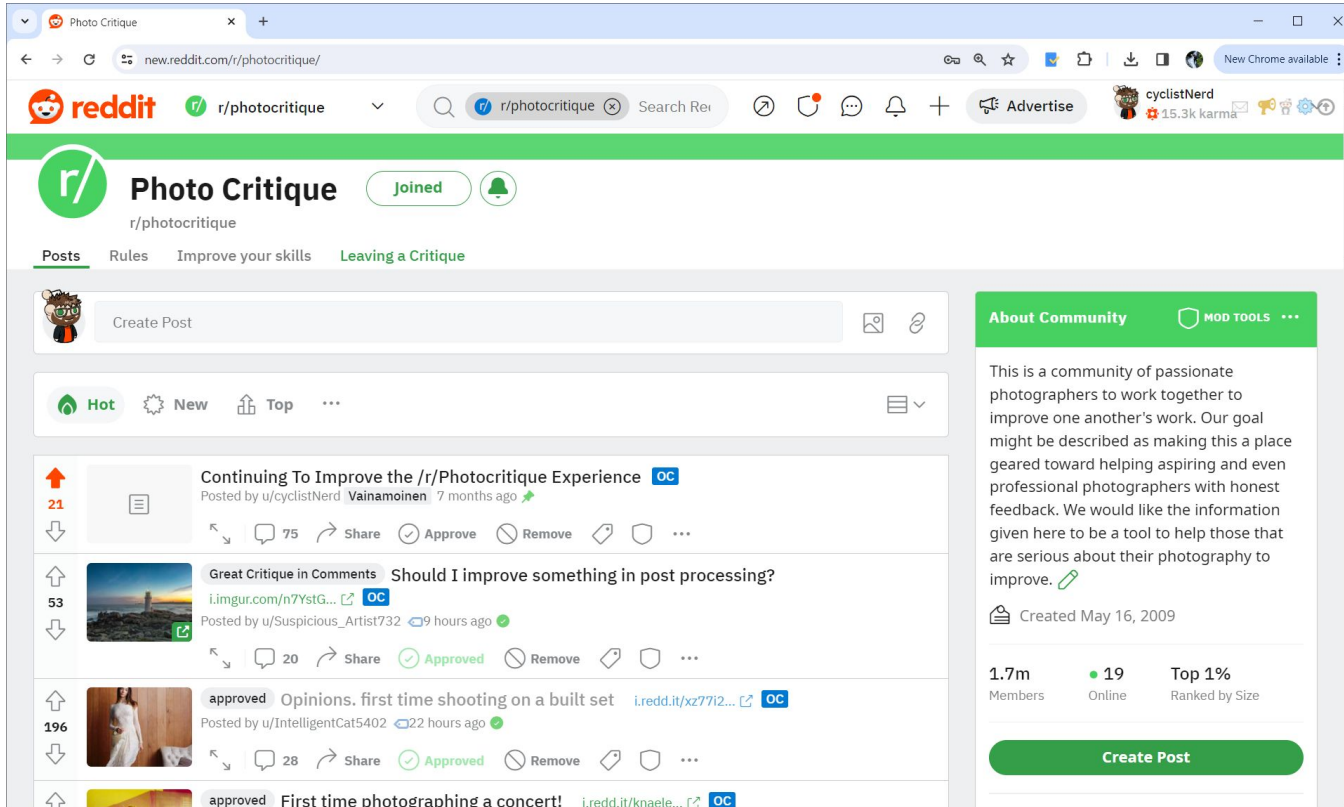
How can we scalably and robustly measure the quality of content in a community-specific manner?

RQ **2**

How can we use LLMs to provide an adaptable system to help improve discussion?

Photocritique Coaching

(Ongoing)



This is
/r/photocritique.

I've moderated it for
8+ years.

In that time, it's
grown from 37K to
1.7M subscribers



r/photocritique • 2 days ago



Thoughts?

Great Critique in Comments

Looking for some critiques on this shot. I'm not really a big landscape guy because I feel like they are often boring. Not sure how to improve this. I'm not interested in stacking exposures.



11



12

Share



• 1d ago

Phone shot essentially

0



Reply

Share



1 more reply



r/photocritique • 2 days ago



Thoughts?

Great Critique in Comments

Looking for some critiques on this shot. I'm not really a big landscape guy because I feel like they are often boring. Not sure how to improve this. I'm not interested in stacking exposures.



11



12

Share



• 1d ago

What an amazing place! To try and be constructive, I think the frame feels a bit unbalanced. All the visual interest is over to the left hand side of the frame. I like that the river acts as a leading line, perhaps a longer focal length to help make the formations a bit more prominent or a vertical shot, crop or slight pan to the left might help direct the eye to the formations and balance it out a bit?

Also might be my eye but the scene looks a little wonky, like you've made the far horizon horizontal but maybe it naturally drops from left to right so now the rock formations look a bit off?



Vote



Reply



Share





r/photocritique • 2 days ago



Thoughts?

Great Critique in Comments

Looking for some critiques on this shot. I'm not really a big landscape guy because I feel like they are often boring. Not sure how to improve this. I'm not interested in stacking exposures.



11



12



Share



• 1d ago

What an amazing place! To try and be constructive, I think the frame feels a bit unbalanced. All the visual interest is over to the left hand side of the frame. I like that the river acts as a leading line, perhaps a longer focal length to help make the formations a bit more prominent or a vertical shot, crop or slight pan to the left might help direct the eye to the formations and balance it out a bit?

Also might be my eye but the scene looks a little wonky, like you've made the far horizon horizontal but maybe it naturally drops from left to right so now the rock formations look a bit off?



Vote



Reply



Share



OP • 22h ago

Thanks for the feedback. I'll try another crop. I'm not sure about the horizon line you are talking about, I haven't done anything there. Maybe cropping if the right side would help with that also.



Vote



Reply



Share



Thoughts?

Great Critique in Comments

Looking for some critiques on this shot. I'm not really a big landscape guy because I feel like they are often boring. Not sure how to improve this. I'm not interested in stacking exposures.



██████████ • 1d ago

What an amazing place! To try and be constructive, I think the frame feels a bit unbalanced. All the visual interest is over to the left hand side of the frame. I like that the river acts as a leading line, perhaps a longer focal length to help make the formations a bit more prominent or a vertical shot, crop or slight pan to the left might help direct the eye to the formations and balance it out a bit?

Also might be my eye but the scene looks a little wonky, like you've made the far horizon horizontal but maybe it naturally drops from left to right so now the rock formations look a bit off?

↑ Vote ↓ Reply ↑ Share ...



OP • 22h ago

Thanks for the feedback. I'll try another crop. I'm not sure about the horizon line you are talking about, I haven't done anything there. Maybe cropping if the right side would help with that also.

(+) ↑ Vote ↓ Reply ↑ Share ...



OP • 22h ago

!CritiquePoint

↑ Vote ↓ Reply ↑ Share ...

CritiquePointBot MOD • 22h ago

Confirmed: 1 helpfulness point awarded to u/Economy-Wash5007 by /u/tsi.

See [here](#) for more details on Critique Points.

↑ Vote ↓ Reply ↑ Share ...

Critique Points

- System deployed two years ago
- Since then, collected 4,732 (and counting) examples of high quality feedback
- Used for an empirical study of what makes for good feedback

Good Feedback is

- Polite!
- Specific and actionable
- Answers the question(s) asked by the OP
- Goes beyond what the OP asked for

Prototype Demo!

(credit to Oscar)

Photocritique Coaching

(Ongoing)

We plan on deploying this tool in an RCT with participants from /r/photocritique to evaluate how it improves discussion quality.

Ongoing Projects



Photocritique Coaching

How can we improve discussion quality?

Rules Deep Dive (with REU Leon)

What rules and enforcement is best?

Community Dashboards

Do better-informed mods make for better communities?

Constructed Observational Study

Evaluate causal inference methods in a real world setting.

Rules Deep Dive (with REU Leon)

(Ongoing)

The image shows a screenshot of a web browser with two tabs open: "reddit for mountain bikers" and "mountainbiking". A large purple rectangular overlay covers the main content area. The overlay contains the following text:

Problem
It's hard for mods to know what rules to set and how to enforce them.

Solution
Measure the impact of rules and their enforcement at a massive scale.

Below the text, a list of rules is partially visible, with some items highlighted by pink boxes:

- 1 Be cool to each other!
- 1 Be NICE
- 2 Need help choosing a bike?
- 2 No Self-Promotion Spam

On the left side of the overlay, there are numbered markers from 1 to 7, with pink brackets indicating the vertical extent of the text blocks.

Intuition

Subreddits change their rules over time.

We have reconstructed rules timelines using the Wayback Machine.

72.6K
rules

7.3K
subreddits

7
years

Rules Deep Dive (with REU Leon)

(Ongoing)

RQ **1**

How does adding **different types of rules** affect communities' attitudes towards their moderators?

RQ **2**

How do different **rule enforcement strategies** affect communities' attitudes towards their moderators?

Rules Deep Dive (with REU Leon)

(Ongoing)

Moderator Perceptions Dataset

/r/memes



/u/Reddit_4_Life

This subreddit is a cesspool of garbage, but the moderators do their best...

Reddit Rules! Taxonomy

Advertising & Commercialization	Images	Prescriptive
Consequences/ Enforcement	Links & Outside Content	Reddiquette/ Sitewide
Content/Behavior	Low-Quality Content	Reposting
Copyright/Piracy	NSFW	Restrictive
Doxxing	Off-topic	Spam
Format	Personal Army	Spoilers
Harassment	Personality	Trolling
Hate Speech	Politics	Voting

Hand-labeled training data for classifier.

Content Removals

Content removals can be counted for all subreddits. Some subreddits provide additional details.

**Proposed
Projects**

Photocritique Coaching

How can we improve discussion quality?

Rules Deep Dive (with REU Leon)

What rules and enforcement is best?

Community Dashboards

Do better-informed mods make for better communities?

Constructed Observational Study

Evaluate causal inference methods in a real world setting.

Community Dashboards

(Proposed)

Problem

It's hard for mods to make informed, data-driven decisions.



Community Dashboards

(Proposed)

Problem

It's hard for mods to make informed, data-driven decisions.

Solution

Build and deploy community dashboards to consolidate metrics into a unified tool.



Community Dashboards

(Proposed)

RQ **1**

What information needs do community moderators have to inform moderation decisions, and how can we design a system to provide this information?

RQ **2**

How does this system improve how well community moderators are able to do their jobs, and as a result, measurably improve community outcomes?

/r/photocritique Community Dashboard

Discussion of Mods ⓘ

65% positive 

(up 2 points since last week)

Recent Posts Mentioning Mods

modmail



/u/photographer123



/u/cyclistNerd is a great moderator, he was quick to respond to my modmail



/u/camera_boy76



No one responded to my modmail message for over 2 days!

Rude comments

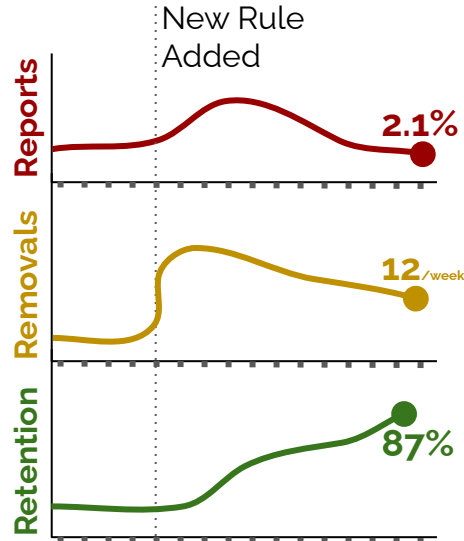


/u/catpics4me



I wish the mods would remove

Trends ⓘ



Discussion ⓘ

I'm optimistic that the new rule is helping [-User12](#)

Me too! Things seem better already... [-Mod1](#)

/r/photocritique Community Dashboard

Discussion of Mods (i)

65% positive 
(up 2 points since last week)

Recent Posts Mentioning Mods

modmail

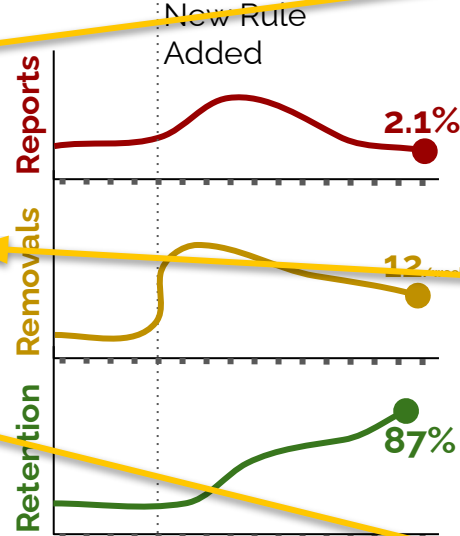
/u/photographer123
 /u/cyclistNerd is a great moderator. he was quick to respond to my modmail

/u/camera_boy76
 No one responded to my modmail message for over 2 days!

Rude comments

/u/catpics4me
 I wish the mods would remove

Trends (i)



Discussion (i)

I'm optimistic that the new rule is helping -User12
Me too! Things seem better already... -Mod1

Discussion of Mods Panel leverages Perceptions of Mods pipeline from earlier work.

This section **highlights specific comments** from around the subreddit, or from dedicated discussion threads (similar to Polis).

Topic modeling can be used to **group comments and identify frequently mentioned topics**.

/r/photocritique Community Dashboard

Discussion of Mods ⓘ

65% positive  65%
(up 2 points since last week)

Recent Posts Mentioning Mods

modmail

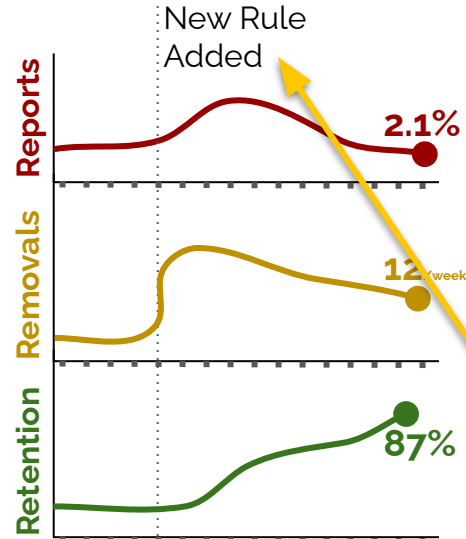
 /u/photographer123
 /u/cyclistNerd is a great moderator, he was quick to respond to my modmail

 /u/camera_boy76
 No one responded to my modmail message for over 2 days!

Rude comments

 /u/catpics4me
 I wish the mods would remove

Trends ⓘ



Discussion ⓘ

I'm optimistic that the new rule is helping -User12
Me too! Things seem better already... -Mod1

The **Trends Panel** shows multiple metrics (reports, removals, user churn, and more) in a unified plot over time.

Mods can add markers for specific events, to track resulting changes. (Inspired by CivilServant)

CivilServant: Community-Led Experiments in Platform Governance.
J. Nathan Matias, Merry Mou. CHI 2018.

/r/photocritique Community Dashboard



Discussion of Mods ⓘ

65% positive 

(up 2 points since last week)

Recent Posts Mentioning Mods

modmail

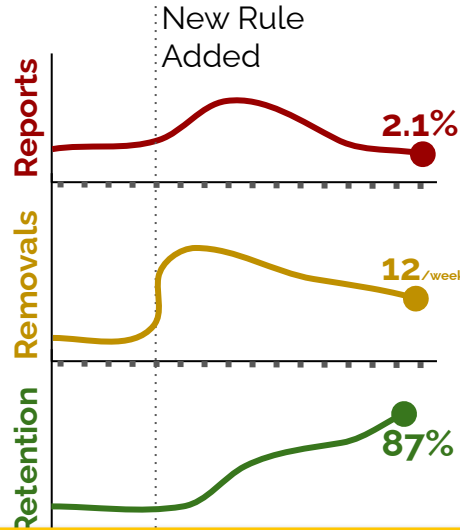
 /u/photographer123
 /u/cyclistNerd is a great moderator, he was quick to respond to my modmail

 /u/camera_boy76
 No one responded to my modmail message for over 2 days!

Rude comments

 /u/catpics4me
 I wish the mods would remove

Trends ⓘ



Discussion ⓘ

I'm optimistic that the new rule is helping [-User12](#)

Me too! Things seem better already... [-Mod1](#)

The **Discussion Panel** allows community members to discuss the community with moderators and one another, anonymously (if desired).

Community Dashboards

(Proposed)

By partnering with several communities, we plan on deploying our dashboards and measuring our moderators and community outcomes change and improve as a result.

**Proposed
Projects**

Photocritique Coaching

How can we improve discussion quality?

Rules Deep Dive (with REU Leon)

What rules and enforcement is best?

Community Dashboards

Do better-informed mods make for better communities?

Constructed Observational Study

Evaluate causal inference methods in a real world setting.

Constructed Observational Study

(Proposed)

Problem

It's challenging to evaluate causal inference methods for observational studies in a real world environment.

Solution

Conduct a constructed observational study with a parallel RCT to evaluate causal inference methods.

Constructed Observational Study

(Proposed)

RQ **1**

How can observational causal inference methods be deployed in a robust and real-world settings to produce meaningful insights?

RQ **2**

How can we produce a realistic and useful evaluation benchmark for causal inference methods that use text data?

Constructed Observational Study

(Proposed)

Constructed observational studies use a parallel RCT to provide 'ground truth' for an observational study.

Common in other domains, but rare with social media data (it's challenging to run large scale RCTs unless you work at reddit, Facebook, *etc.*)

Constructed Observational Study

(Proposed)

What intervention?



cyclistNerd MOD  · 3 days ago · [Stickied comment](#)

Vainamoinen

In the future, please post one photo per submission - no collages.



1



Reply

Share



Constructed Observational Study

(Proposed)

What intervention?



cyclistNerd **MOD**  · 3 days ago · [Stickied comment](#)

Vainamoinen

In the future, please post one photo per submission - no collages.



 **OP** · 3 days ago



Sorry about that. I'll make sure to do that next time :)



Constructed Observational Study

(Proposed)

Removal Reasons

- Lots of natural variance in my subreddit (and many others) in how reasons are (or not) used
 - Variance in the intervention is needed for observational studies
- Easy to experiment with when collaborating with mod teams
- Unit is a user - possible to collect many users

Still clearly many challenges with this idea.

There's a reason constructed observational studies are rare!

Constructed Observational Study

(Proposed)

Collaborate, ideally with reddit, or a small set of subreddits, to run an RCT and conduct a parallel observational study.

The RCT provides ground truth, allowing us to evaluate many observational methods for adjusting for confounding with text.

Could include an LLM-based propensity score model.

Research Activities

Defining

Assessing

Deploying

Community Values (ICWSM '22 & '24)

No "one size fits all" set of values!

Perceptions of Mods (arXiv '24)

Happiest when mods are engaged and not overworked!

News Sharing Behavior (🏆 ICWSM '21)

Voting and crossposting can reduce problematic content!

Causal Inference Methods (ICWSM '22)

Methods are hard to deploy irl., evaluation is critical!

Ongoing & Proposed Work

Coaching, Rules, Dashboards, Constructed Studies

A huge thanks to my coauthors, labmates, and advisors!

Coauthors

Tim Althoff, UW
Amy X. Zhang, UW
Leon Leibmann, UW
Peter West, UW
Maria Glenski, PNNL
David Arbour, Adobe
Ryan Rossi, Adobe

bdata

Tim Althoff
Bret Nestor
Chris Rytting
Vainayak Gupta
Xinyi Zhou
Ashish Sharma
Inna Lin
Ken Gu
Margaret Li
Mike Merrill
Oscar Liu
Leon Leibmann

Social Futures Lab

Amy X. Zhang
Jim Chen
Xinyi Zhou
Nick Vincent
Ruotong Wang
Kevin Feng
Jina Yoon
Leijie Wang
Alicia Guo

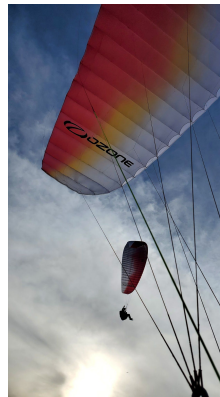


behavioral
data science



Social
Futures
Lab

...and friends & family for their support!





Questions? and get in touch!

 /u/cyclistNerd

 @galenweld

 gweld@cs.washington.edu

 www.galenweld.com

Prediction Task Details

- 27 tasks (3 dimensions x 9 values)
- Distinguish between top and bottom quartiles (small differences less important)
- L2-regularized Logistic Regression and cross-validation
- Average ROC AUC of .667 across all tasks.
 - Current state is easiest to predict
- Perception of size 0.936 ROC AUC

Value	Dimension	ROC AUC
Democracy	Current State	0.622
	Desired Change	0.684
	Importance	0.541
Diversity	Current State	0.634
	Desired Change	0.800
	Importance	0.716
Engagement	Current State	0.635
	Desired Change	0.532
	Importance	0.642
Inclusion	Current State	0.730
	Desired Change	0.708
	Importance	0.555
Quality	Current State	0.725
	Desired Change	0.624
	Importance	0.677
Safety	Current State	0.441
	Desired Change	0.714
	Importance	0.391
Size	Current State	0.936
	Desired Change	0.655
	Importance	0.661
Trust	Current State	0.709
	Desired Change	0.688
	Importance	0.922
Variety	Current State	0.625
	Desired Change	0.589
	Importance	0.838

Prediction Task Details

<code>sub_num_posts</code>	The number of posts in the subreddit.
<code>sub_num_removed_posts</code>	The number of posts removed by a moderator in the subreddit.
<code>sub_num_deleted_posts</code>	The number of posts deleted by their author in the subreddit.
<code>sub_num_selfposts</code>	The number of selfposts (text-posts) in the subreddit.
<code>sub_num_linkposts</code>	The number of posts which link to external websites.
<code>sub_num_comments</code>	The number of comments in the subreddit.
<code>sub_num_removed_comments</code>	The number of comments removed by a moderator.
<code>sub_num_deleted_comments</code>	The number of comments deleted by their author.
<code>sub_distinct_users</code>	The number of distinct contributors to the subreddit.
<code>sub_num_subscribers</code>	The number of users who ‘subscribe’ to the subreddit.
<code>sub_age</code>	The number of days since the subreddit was founded.
<code>sub_topic_specificity</code>	The manually-categorized specificity of the topic of the subreddit, on an 3-point scale.
<code>sub_topic_category</code>	The manually-categorized (see §3.4) topic of the subreddit.

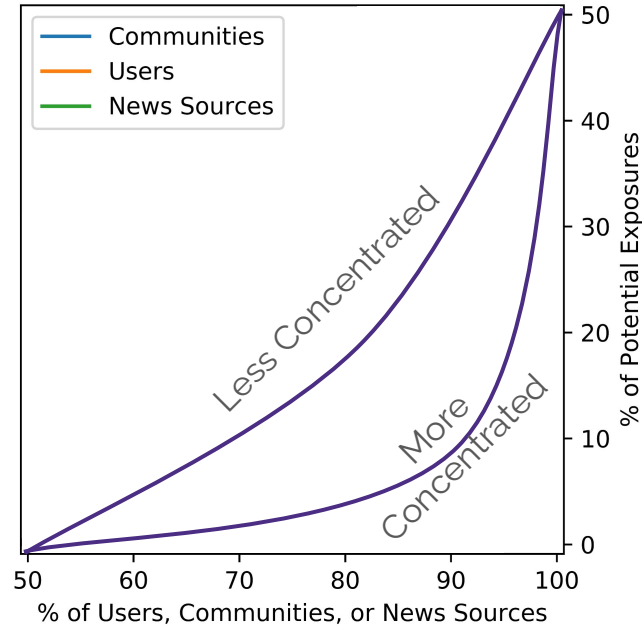
Table 4: Descriptions of features used in the prediction tasks (§7).

Variance Decomposition (Law of Total Variance)

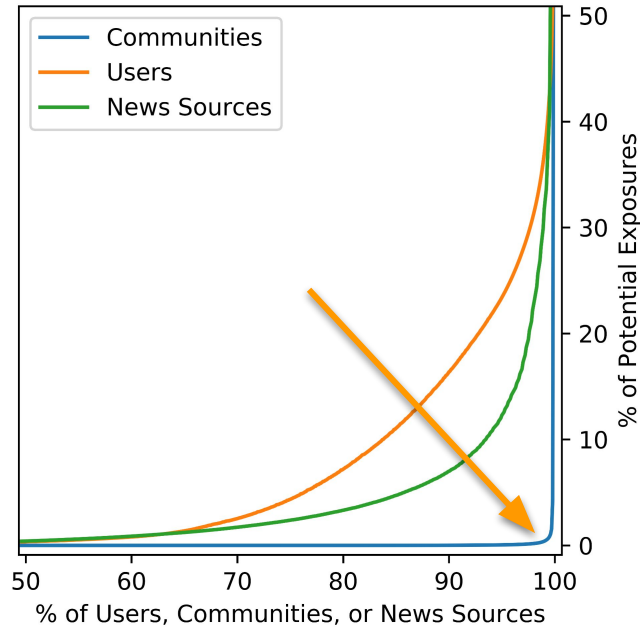
$$\text{Var}(Y) = E[\text{Var}(Y | X)] + \text{Var}(E[Y | X])$$

Where, in our context, Y represents the political bias of a link posted in a community, and X represents the member who posts it.

Biased and low factual content is concentrated in a small number of subreddits



Biased and low factual content is concentrated in a small number of subreddits



Chandrasekharan, et al. 2020. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *TOCHI*.

Ribeiro et al. 2021. Does Platform Migration Compromise Content Moderation? Evidence from r/The Donald and r/Incels. *CSCW*.

This concentration gives credence to reddit's recent decisions to quarantine or ban toxic communities

Details on Real World Tasks

In the **Moderation Experiment**, we test if having a post removed by a moderator impacts the amount a user later posts to the same community.

For this experiment, we use 13,786 public Reddit histories (all of which contain more than 500 tokens) from users in /r/science from 2015-2017 who had a not had a post removed prior to 2018.

Our treated users are those who have had a post removed in 2018. Our untreated users are those who have not had a post removed in 2018 (nor before). The outcome of interest is the number of posts they made in 2019.

In the **Gender Experiment**, we use the dataset made public by Veitch, Sridhar, and Blei (2019), which consists of single posts from three subreddits: /r/okcupid, /r/childfree, and /r/keto.

Each post is annotated with the gender (male or female) of the poster, which is considered the treatment.

The outcome is the score of the post (number of 'upvotes' minus number of 'downvotes').

Five Tasks in our Evaluation Framework

1. Linguistic Complexity
2. Signal Intensity
3. Strength of Selection Effect
4. Number of Users
5. Absence of Treatment Effect

1. Linguistic Complexity

Ideally, models should be able to recognize the mutual importance of different phrases

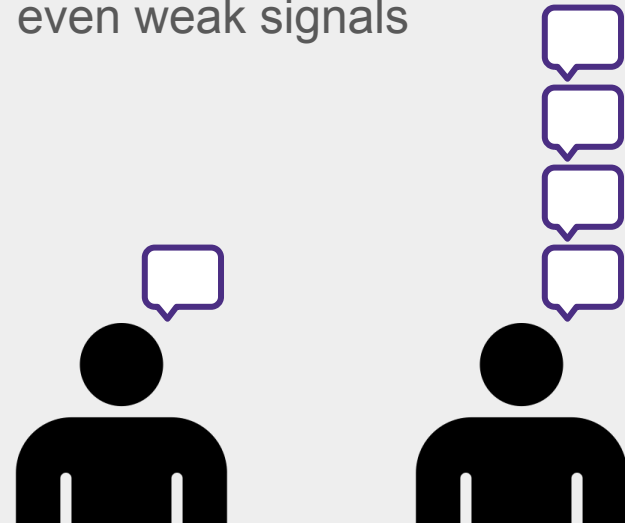
Indicative of depression

I feel
depressed

I am isolated
from my peers

2. Signal Intensity

Models should be able to detect even weak signals



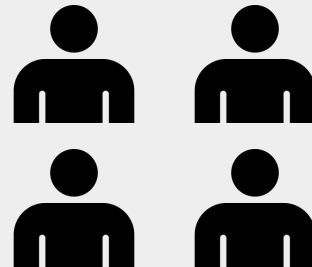
3. Strength of Selection Effect

Models should be able to adjust for confounding, even when there is reduced overlap in the distribution of language between treated and untreated users

4. Number of Users

Ideally, models would be able to perform well even with limited training data

Study A



Study B



5. Absence of Treatment Effect

Models should not predict causality when none is present

