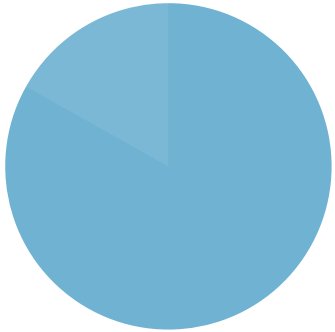


Trust & Safety in Federated Social Networks

A needs assessment of existing fediverse service providers and content moderators

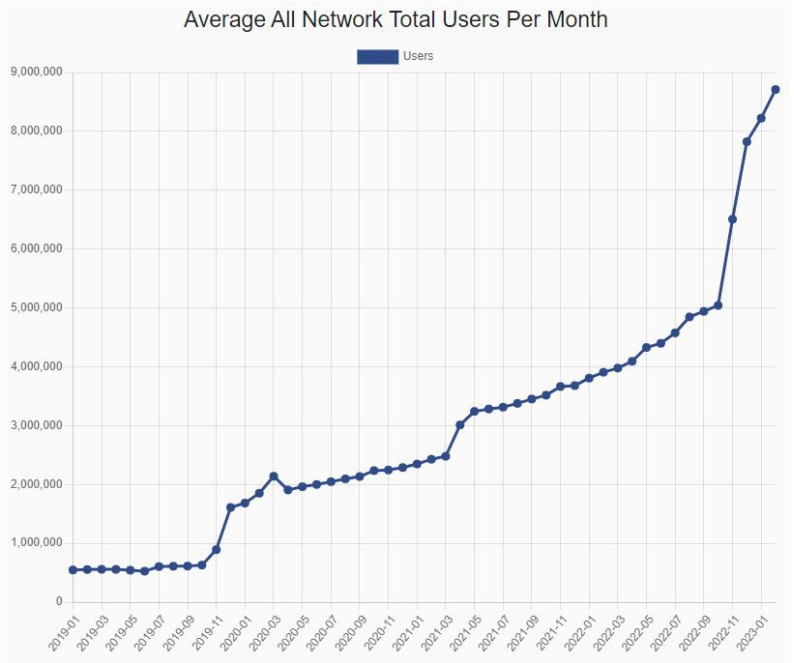
Scaling Trust & Safety in the Fediverse

Problem Statement

The Fediverse is growing rapidly, and there is concern the community cannot keep pace to adequately moderate the volume of user-generated content.

Moderator communities are probing for ways to cooperatively share burden, seek resources, but trust remains the highest issue

In addition, there is concern that service providers are uncertain of legal and regulatory requirements



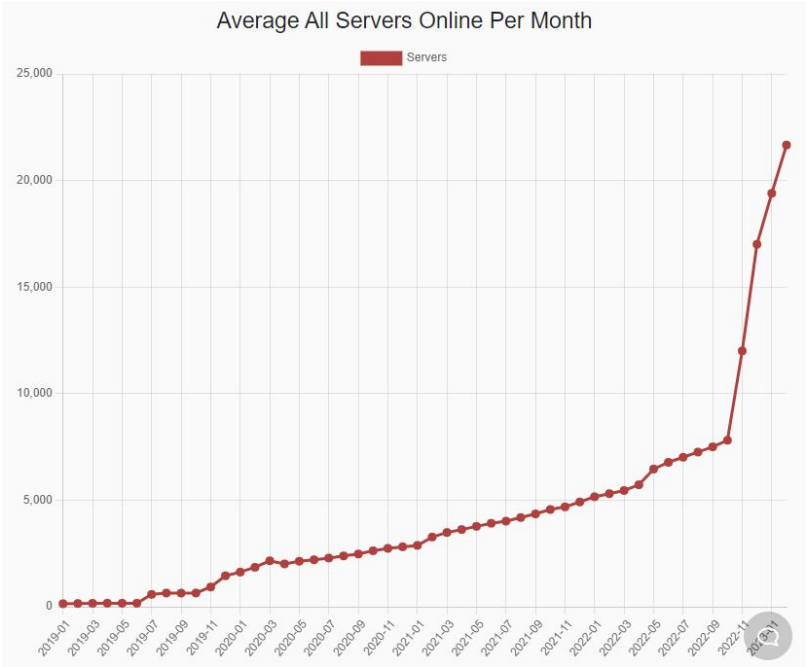
Current State

21,000+ activity pub services online, federating 800M posts / month

21,000+ individuals responsible for content moderation, many of whom may have at best a basic understanding of trust & safety

limited resources for server operators to obtain moderator support

no legal guidance, no compliance resources



Needs Assessment

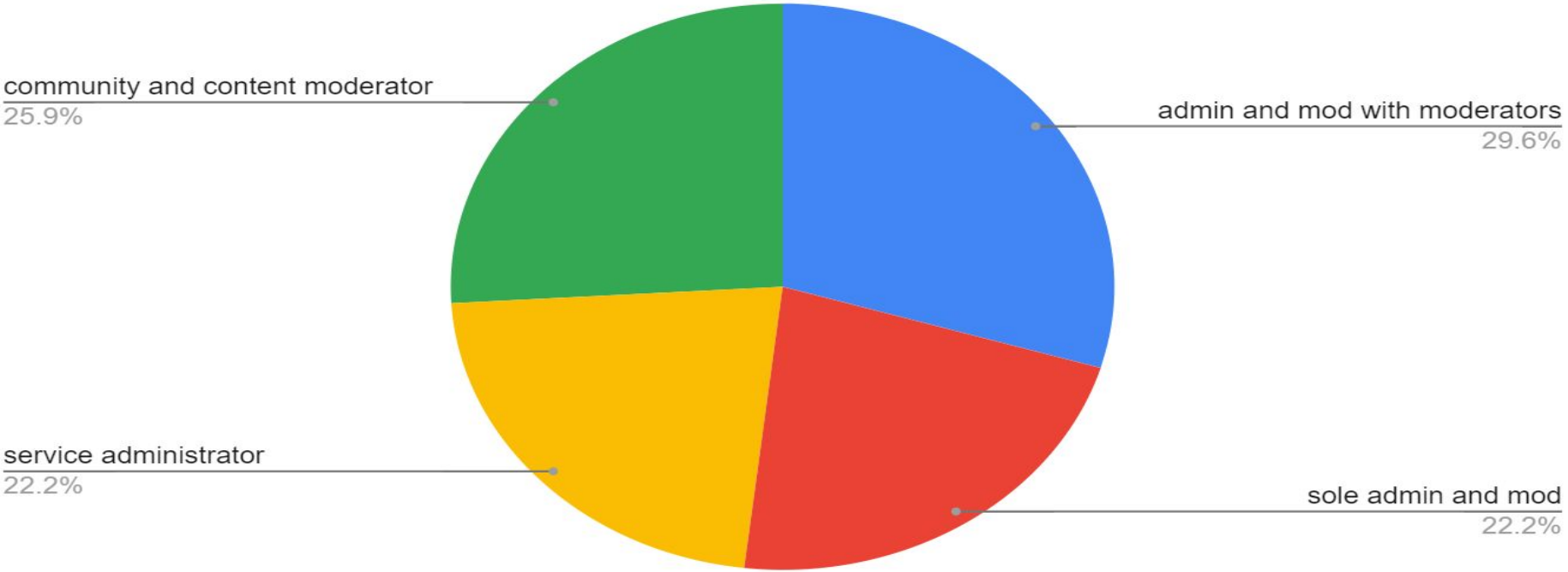
Survey sent to various Fediverse service provider chat room communities, hashtags

responses from service administrators and content moderators representing 475,000 registered members, 185,000 MAU (Feb 2023) - roughly 5% of Mastodon

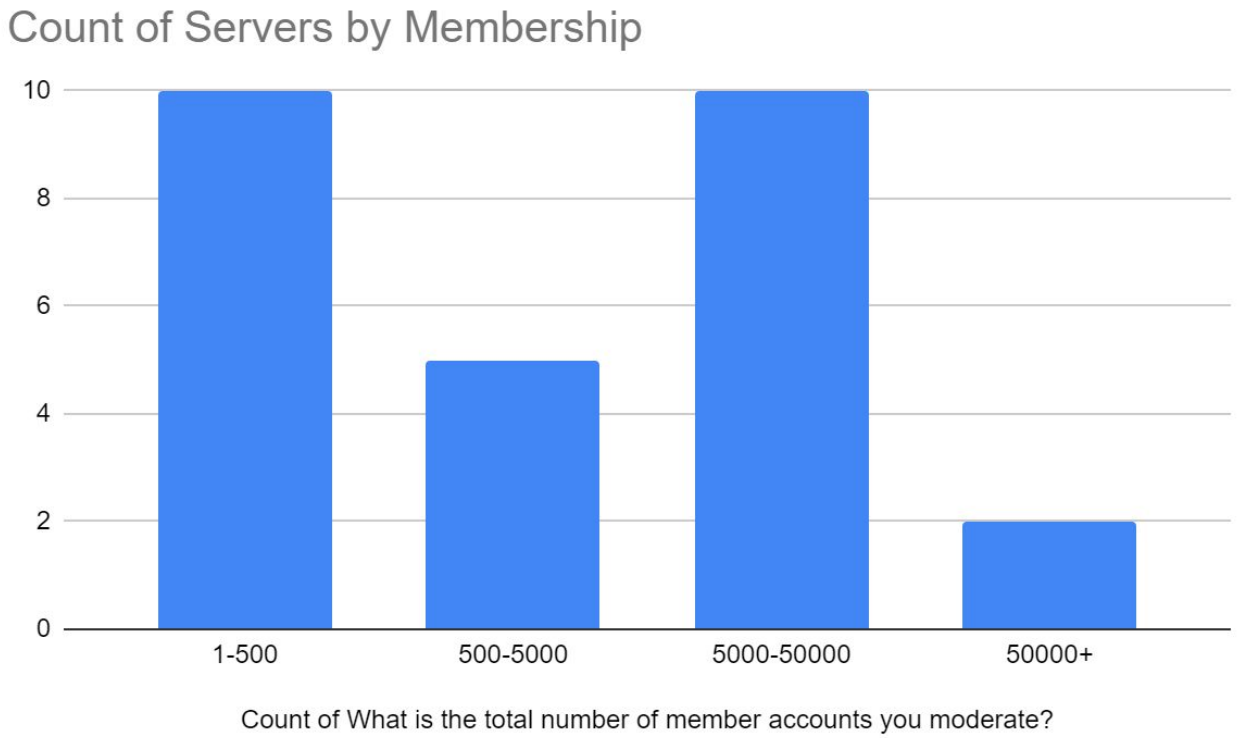
9 respondents have over 10,000 accounts

3 respondents account for roughly 50% of total members and MAU

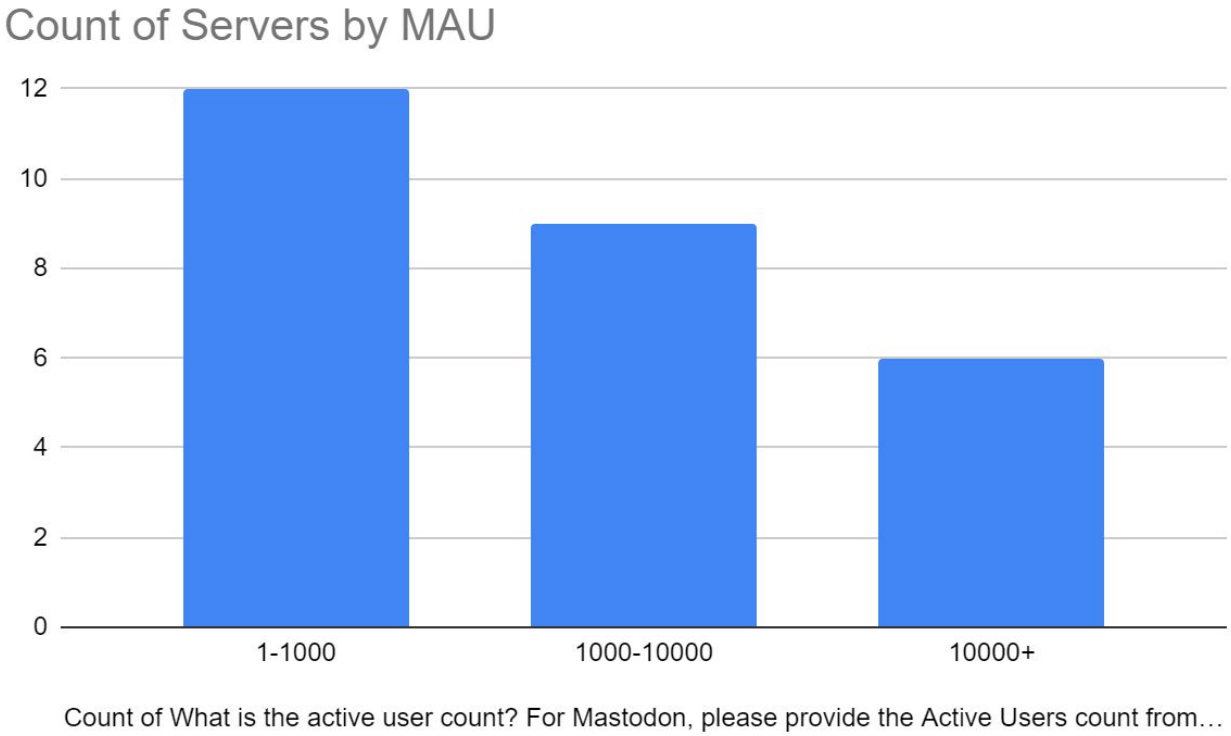
Are you a service administrator, a community and content moderator, or both?



Survey was sent to administrator and moderator communities

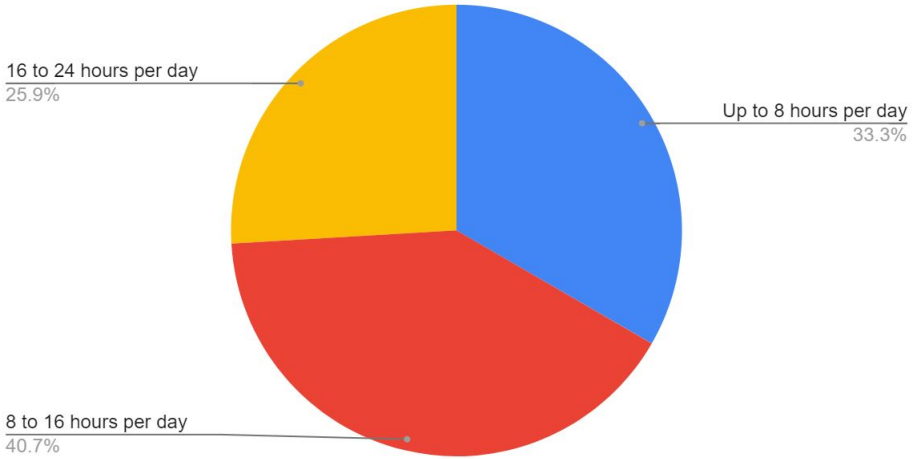


Respondents represent a range of service membership counts

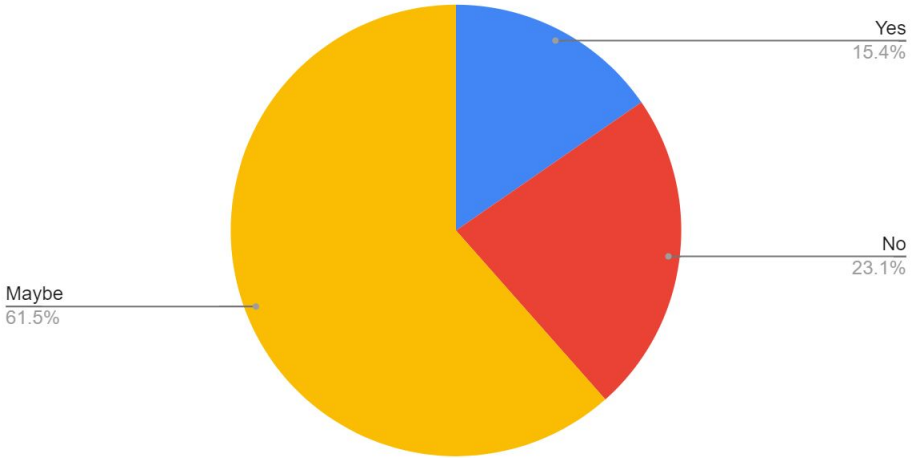


Respondents represent a range of service active membership usage

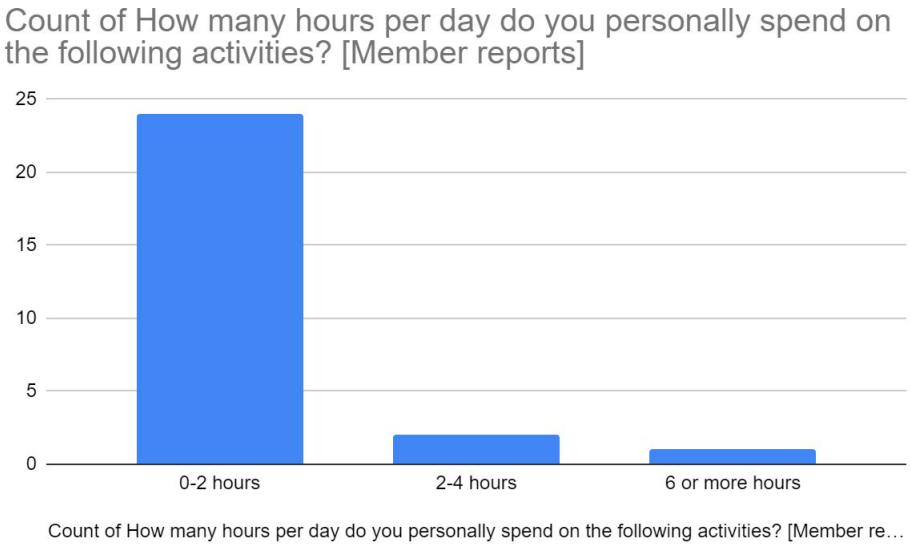
Count of What portion of the day does your service provide moderation coverage to respond to issues?

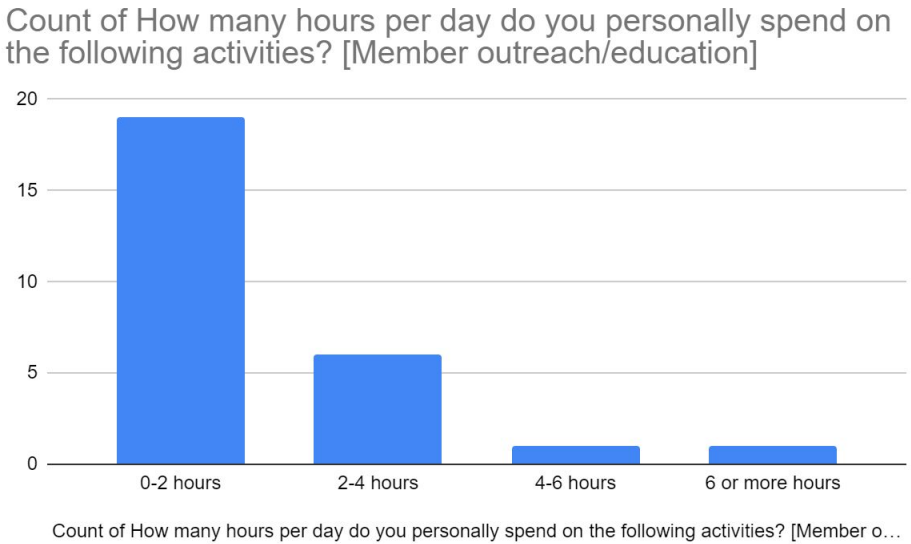


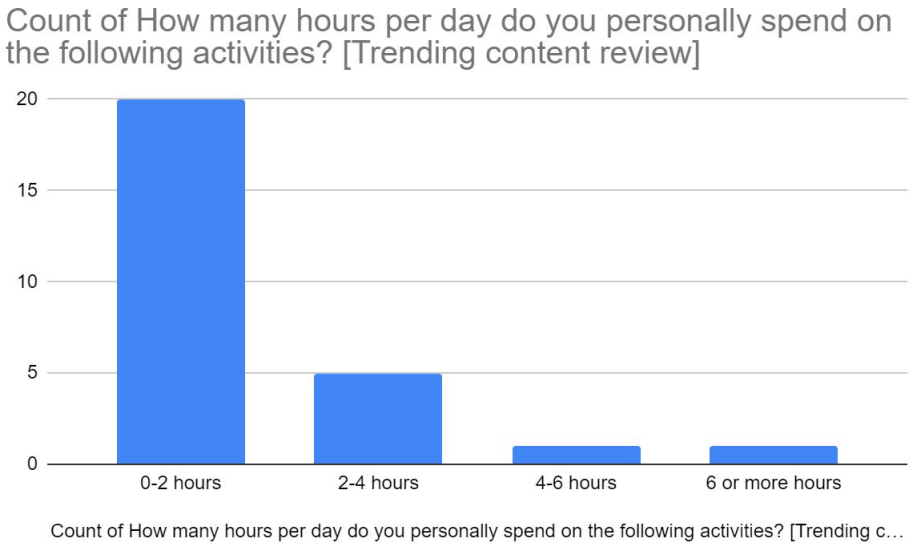
Count of Would you be interested in cooperatively sharing moderation privileges with another service provider? For exa…

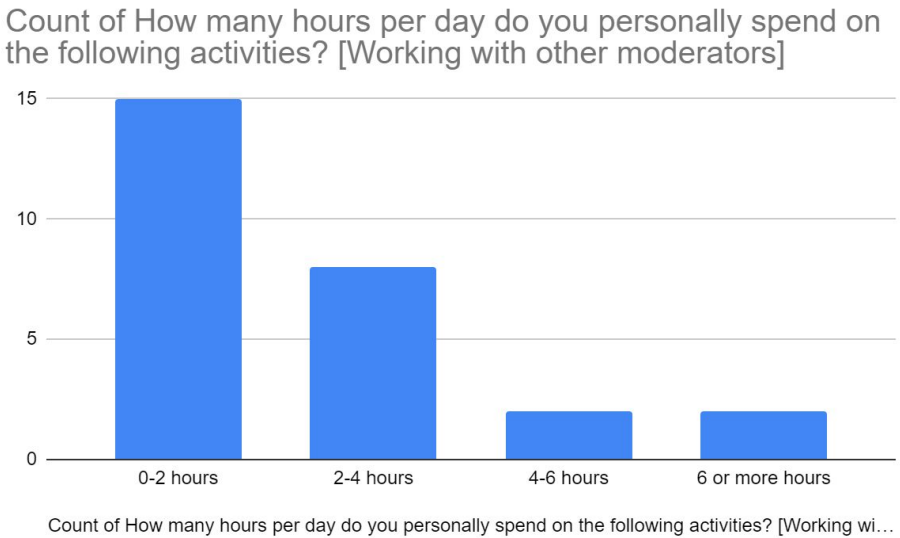


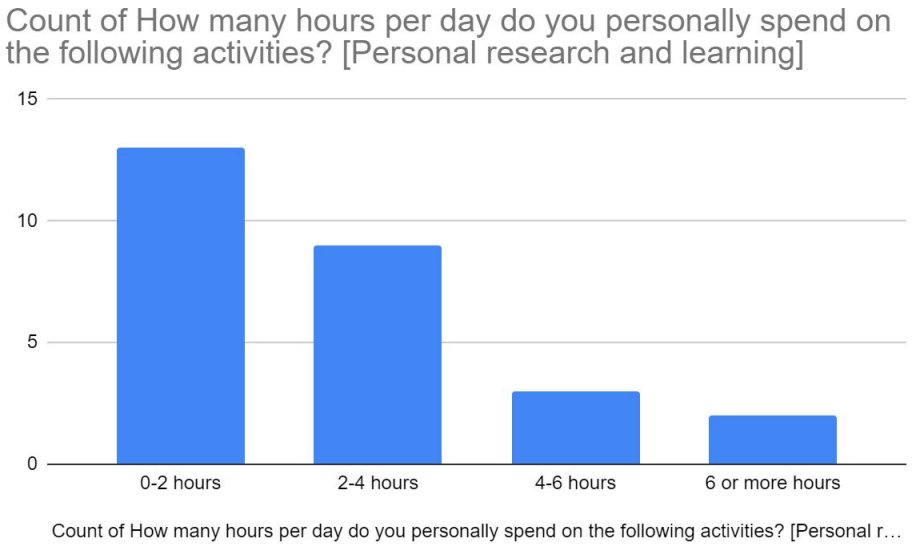
Most service providers do not offer 24/7 content moderation, interest in sharing through co-op / trusted second party

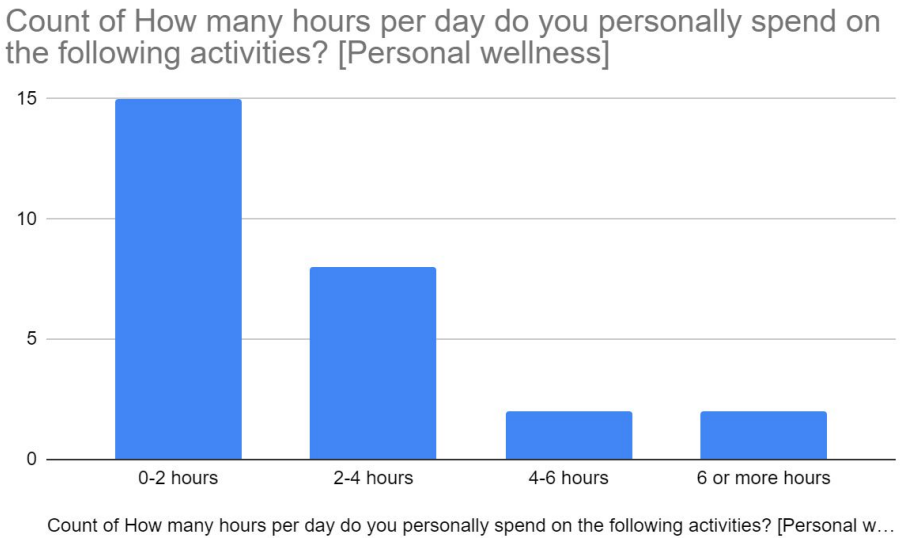

Count of How many hours per day do you personally spend on the following activities? [Member reports]
25
20
15
10
5
0
0-2 hours
2-4 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Member re…


Count of How many hours per day do you personally spend on the following activities? [Member outreach/education]
20
15
10
5
0
0-2 hours
2-4 hours
4-6 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Member o…


Count of How many hours per day do you personally spend on the following activities? [Trending content review]
20
15
10
5
0
0-2 hours
2-4 hours
4-6 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Trending c…


Count of How many hours per day do you personally spend on the following activities? [Working with other moderators]
15
10
5
0
0-2 hours
2-4 hours
4-6 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Working wi…


Count of How many hours per day do you personally spend on the following activities? [Personal research and learning]
15
10
5
0
0-2 hours
2-4 hours
4-6 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Personal r…


Count of How many hours per day do you personally spend on the following activities? [Personal wellness]
15
10
5
0
0-2 hours
2-4 hours
4-6 hours
6 or more hours
Count of How many hours per day do you personally spend on the following activities? [Personal w…

Total Moderation Hours Combined (All staff)

Sole admins reported least number of hours spent on moderation, one reporting 1 hour per day, the rest very minimal.

Community Moderators reported higher number of hours spent on moderation activities, with a bump in “working with other moderators”.

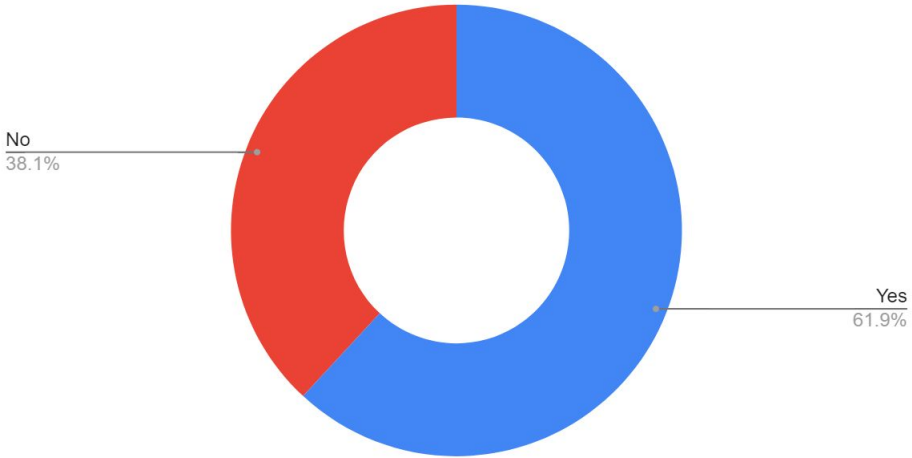
The remainder reported a wide range of total combined hours from all staff and volunteers, one reporting “1-2 hours per day”, another reporting “100+ hours per day”.

Also reported “I have no idea” and “very hard to guess”.

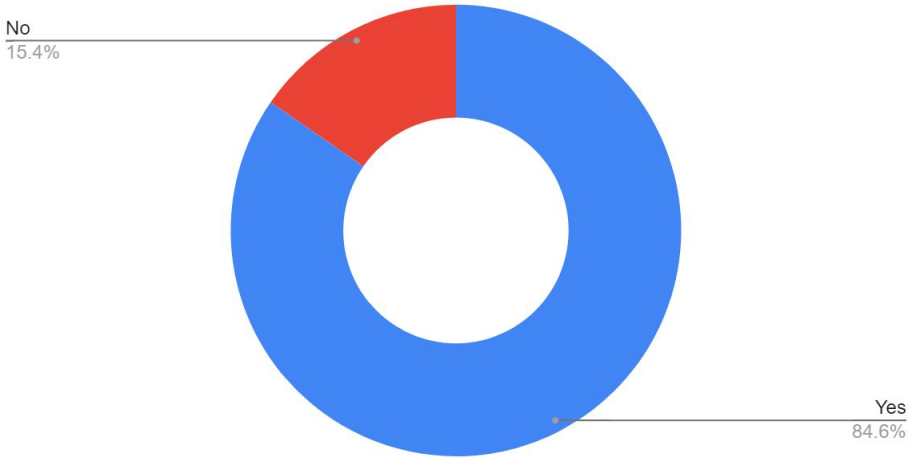
Quick Hits

- Relatively under moderated by portion of day, one third of services moderators active less than 8 hours per day
- Size does not matter, moderation team sizes vary at each service size, some larger servers have smaller teams than some smaller services.
- The medium-sized services tend to have the largest moderation teams

Count of Do you have a formal agreement or code of conduct for your moderators?



Count of Would you be interested in a standard moderator agreement template?



Services that have additional moderator staff do not always have a formal agreement with their moderator support, but most are interested in a standard document

In what legal jurisdictions are you required to demonstrate compliance?

21 respondents stated their legal jurisdictions, including 1 “worldwide”. Of these several are based on “where the server is”; others reference one country; 10 reference two or more jurisdictions

4 other responses include “not sure”, “no idea”, god knows”, “you tell me”

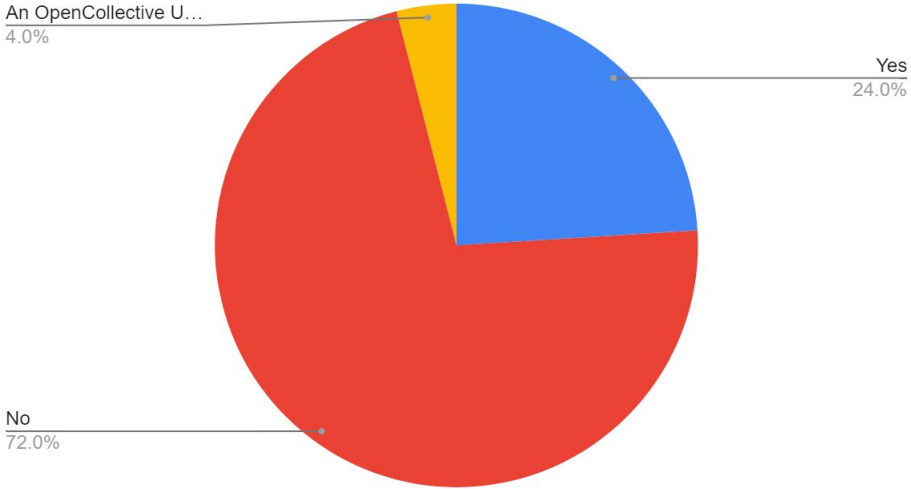
Only 8 included the EU in their response, with some referencing GDPR

8 included the US in their reply, some with additional states

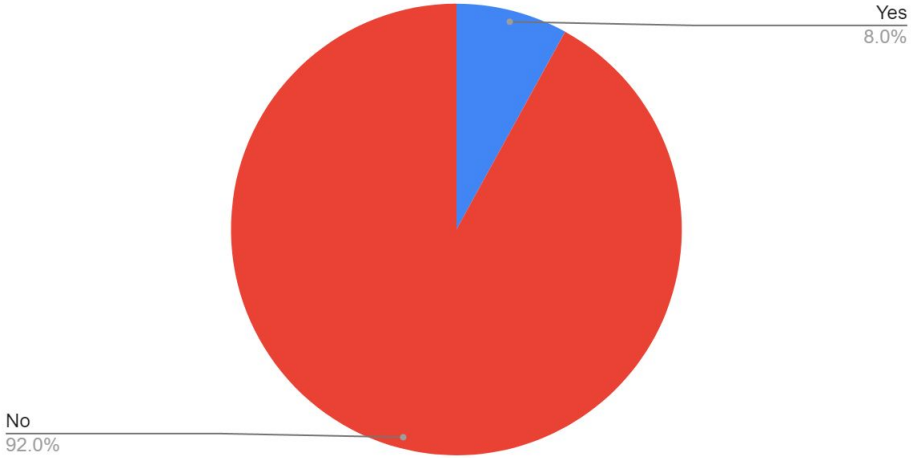
Several respondents stated a US state but not US, or a EU country but not EU

Clear need for clarifying guidance, documentation

Count of Is your service provided by a business entity?

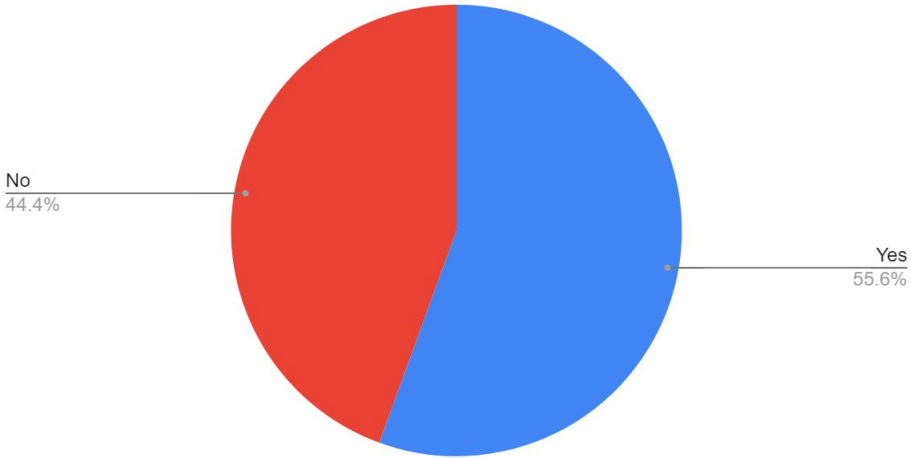


Count of Does your service provide insurance for your actions as an administrator or moderator?

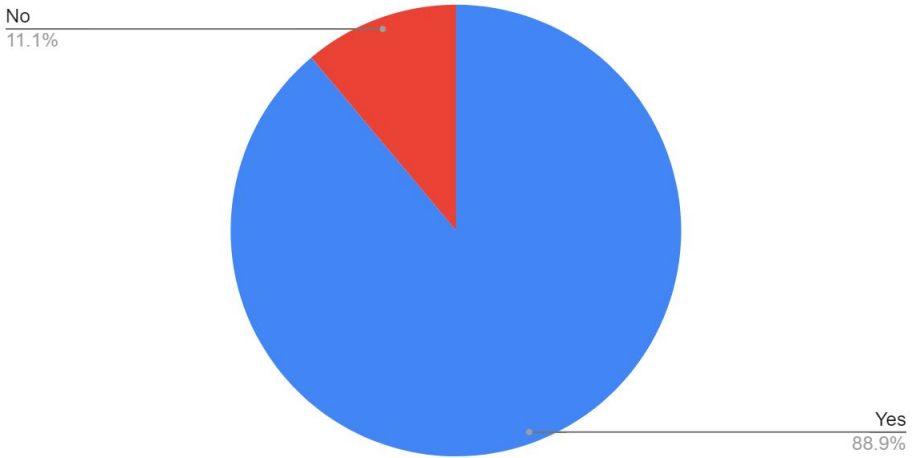


Majority of services are non-commercial and non-insured

Count of Have you ever or do you anticipate requiring legal advice with regard to your activity as an administrator or mod…

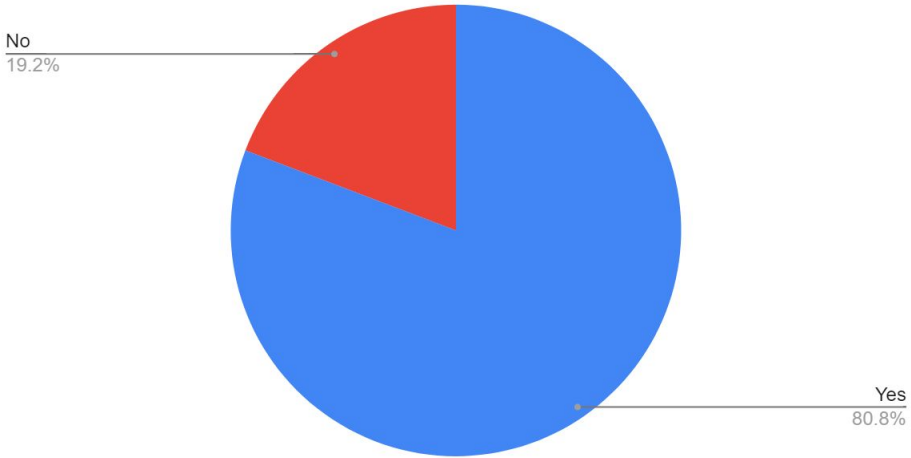


Count of Would you be interested in access to a collective fund for advice from Fediverse-knowledgeable legal services if/wh…

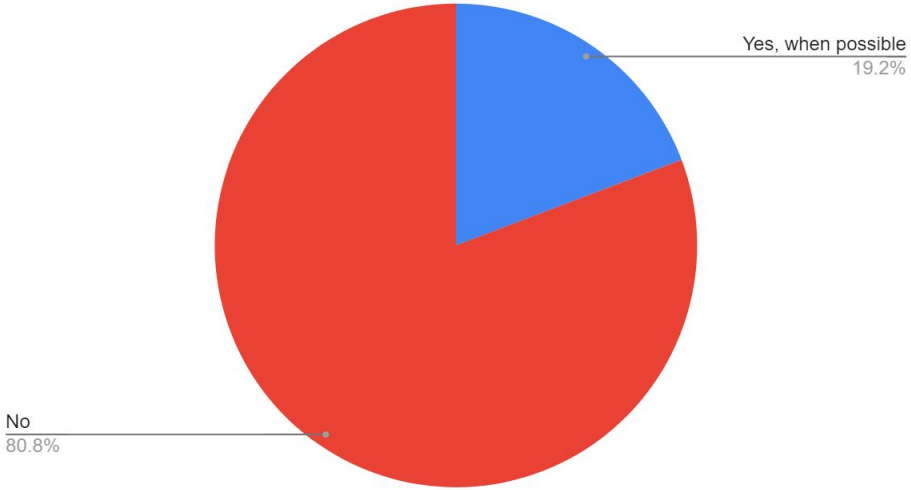


Split on legal requirements, but positive support for access to legal support when needed

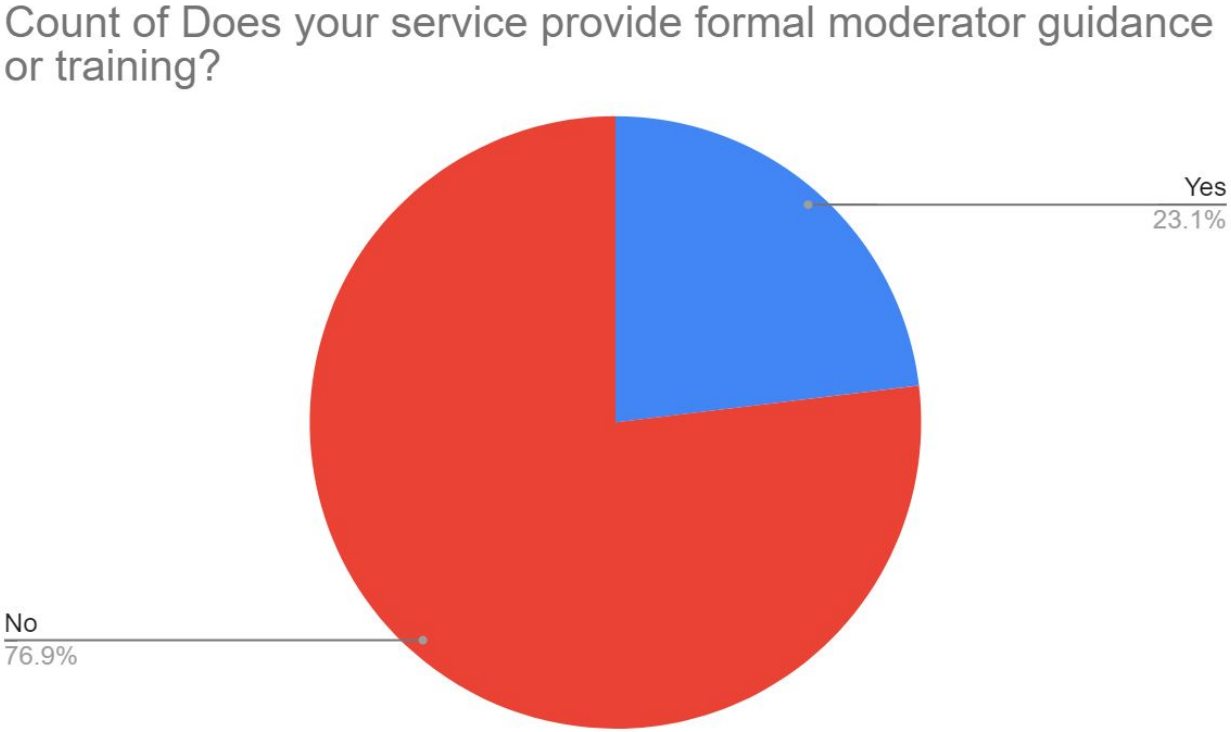
Count of Does your service collect subscriptions, donations, or other monetary support?



Count of Does any of your funding compensate moderators?



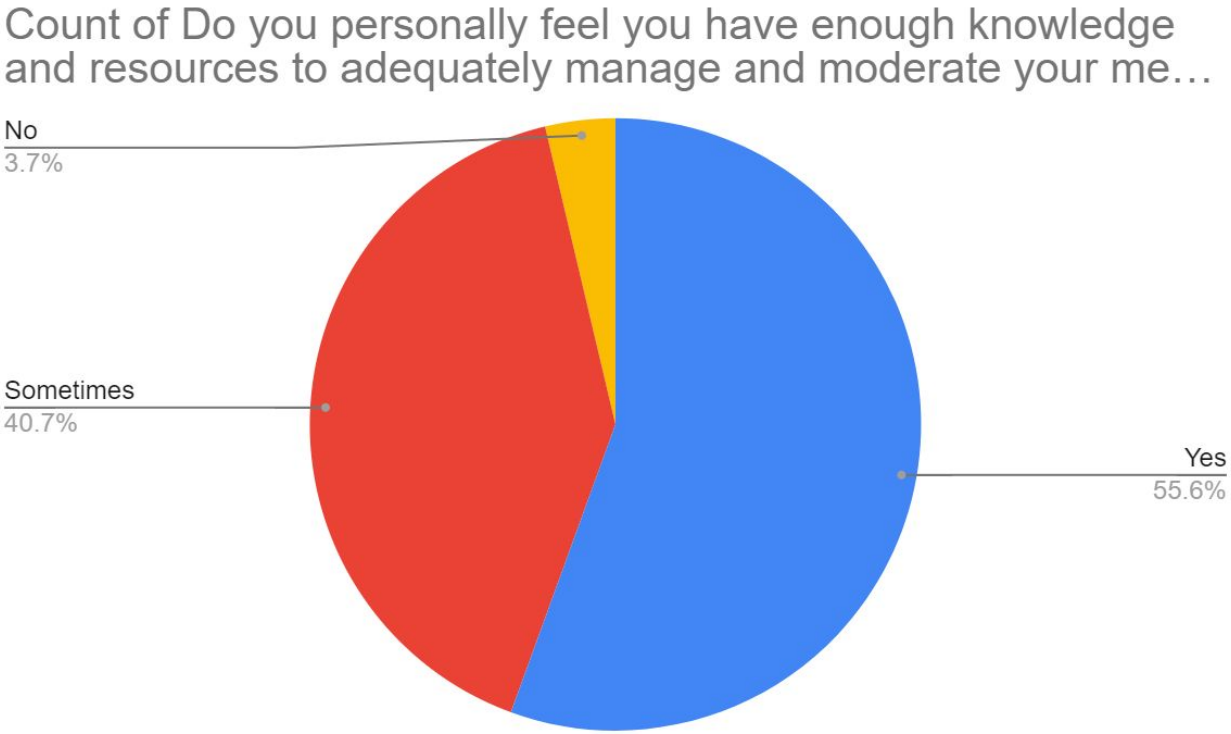
Most services collect some funding, but an inverse proportion compensate moderators (“Yes” was also an option, 0 respondents)



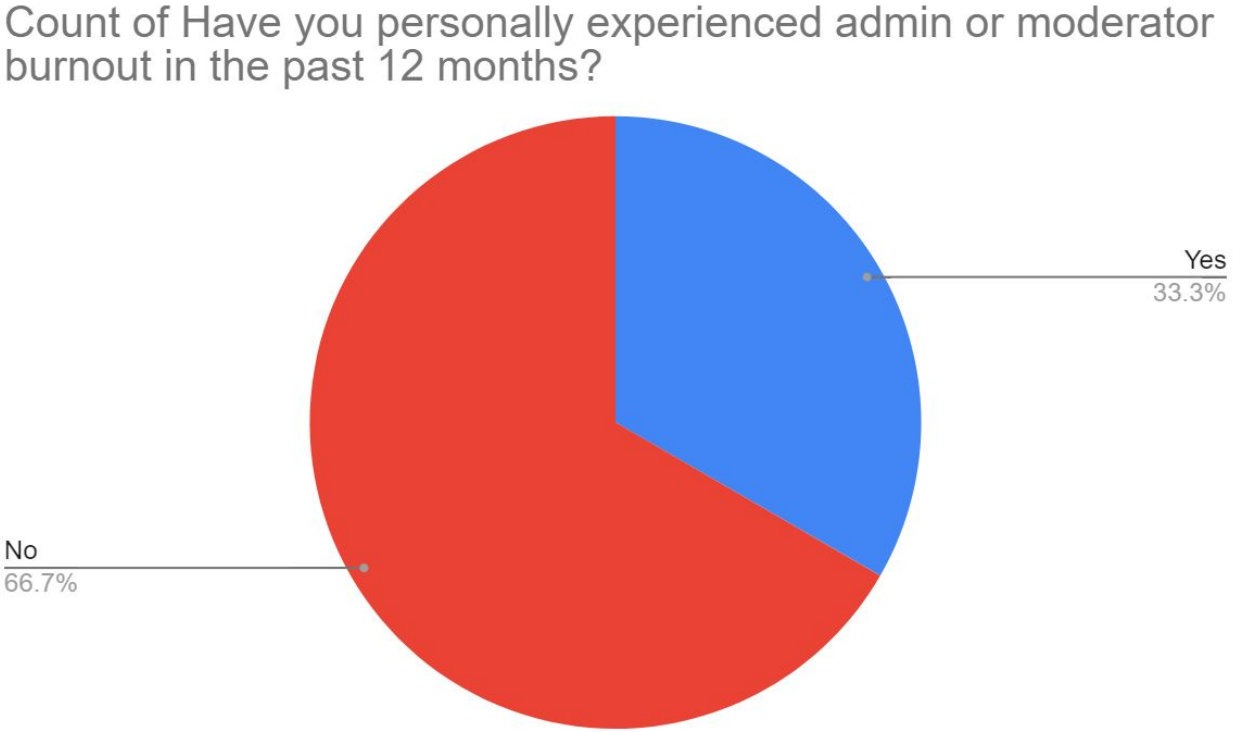
under-resourced for moderator guidance

Quick Hits

- Guidance on compliance, legal issues needed
- Small amount of respondents anticipate legal issues in the future, but heavy support for access to legal services
- Very little compensation for moderation activity
- Lack of moderator training, guidance
- Potential opportunity for group liability insurance, legal fund, guidance on referrals to fediverse-aware legal services

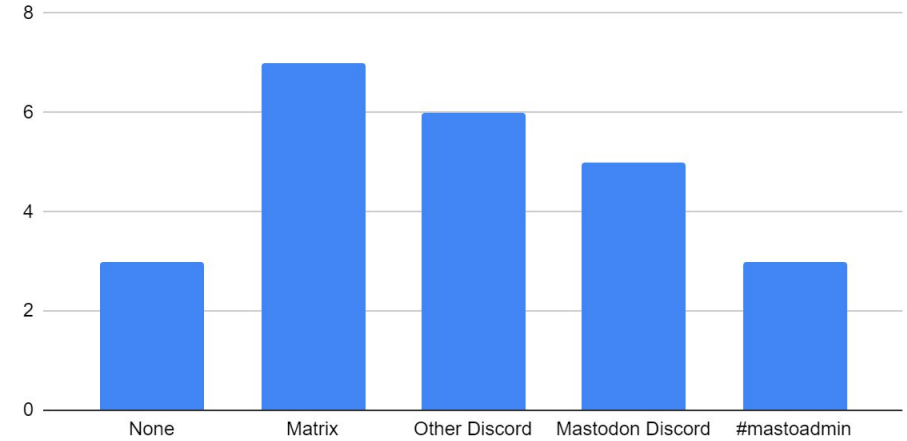


A healthy 40% “sometimes” - people know what they don’t know



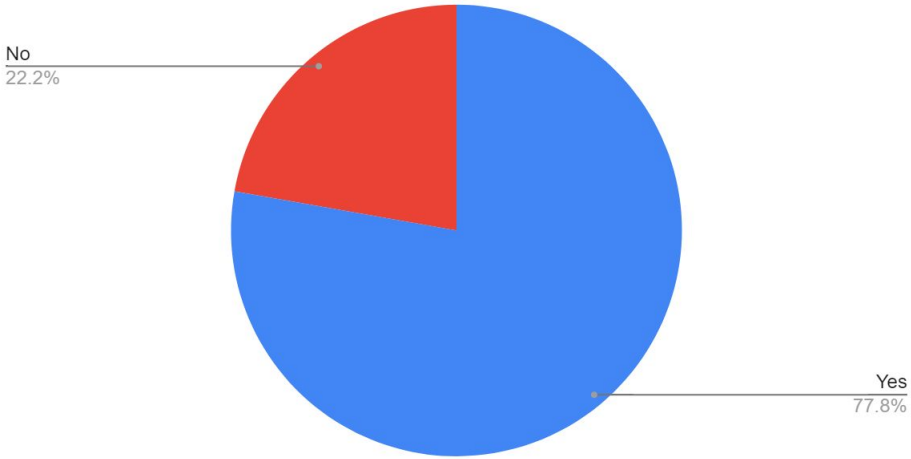
High incidence of burnout

Count of Which moderator communities do you participate in?



Count of Which moderator communities do you participate in? Include ad hoc chat rooms, professi…

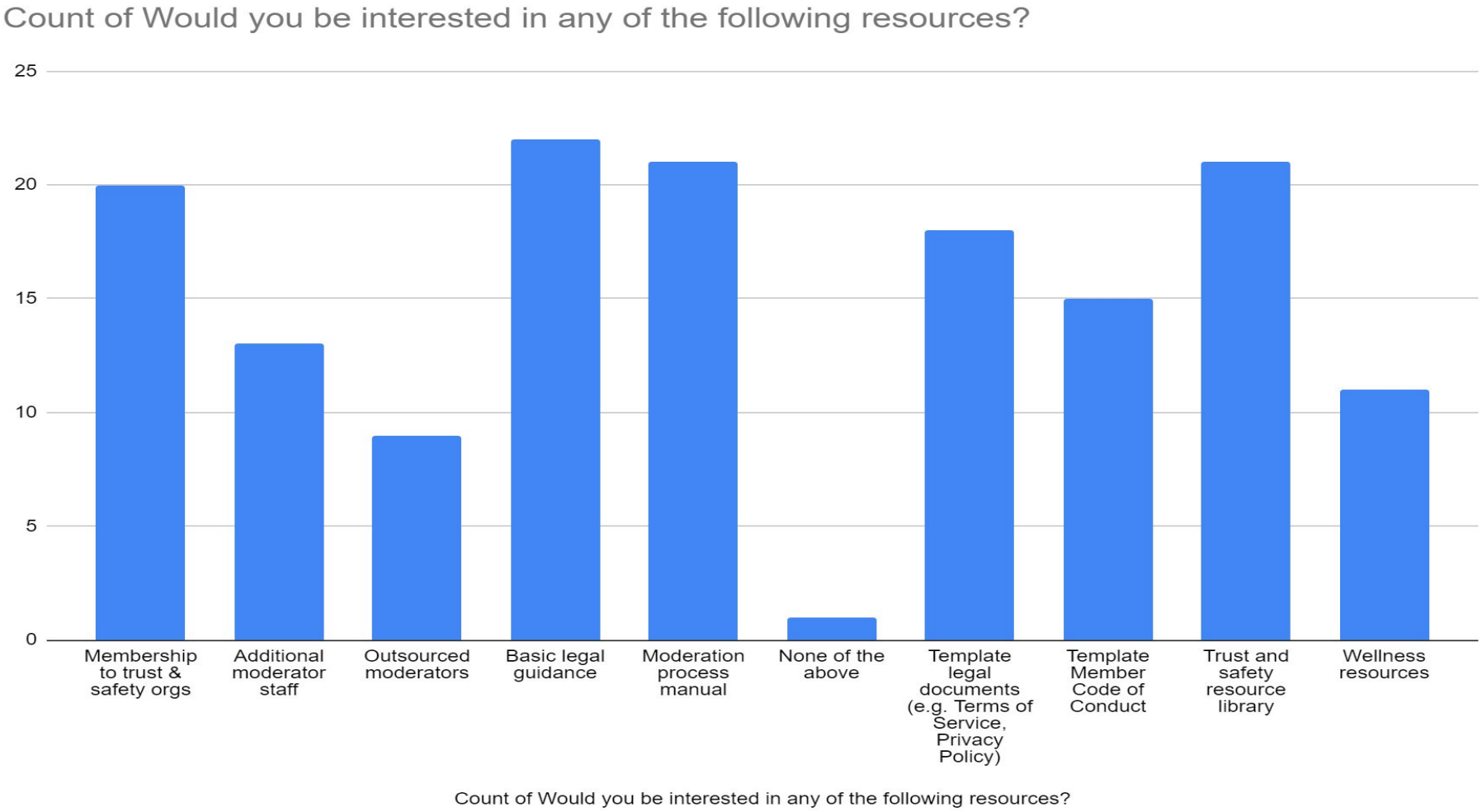
Count of Would you be interested in learning more about existing Moderator communities?



22% have no interest, perception is they have adequate community interaction

Open Comments

- Thanks for your efforts!
- looking for better / open-source / self-hosted platforms for discussion / project-management / peer collaboration
- thanks for setting this up, i'm curious :)
- Scalable tools that make use of an api to act would make a massive impact on moderation tools and analysis
- I would like non-politically aligned moderation.
- interested in some kind of inter-instance admin discussion space
- I'd like to see moderators give more guidance to those offenders who are willing to shed the toxic words
- Banning people without educating them on why just sends them to another community unchanged.
- The biggest point of interest for me is reporting illegal activity outside of server moderation.
- Having a formal moderation funnel for reporting illegal activity like this would be really meaningful.
- Development of report and moderation aggregation and sharing tools that allow servers to form moderation collectives for shared moderation duties across multiple smaller servers sharing the same rules


Count of Would you be interested in any of the following resources?
25
20
15
10
5
0
Membership to trust & safety orgs
Additional moderator staff
Outsourced moderators
Basic legal guidance
Moderation process manual
None of the above
Template legal documents (e.g. Terms of Service, Privacy Policy)
Template Member Code of Conduct
Trust and safety resource library
Wellness resources
Count of Would you be interested in any of the following resources?

Quick Hits

- 23 of 27 are interested in additional moderator staff, either in-house or outsourced
- High interest in TSPA or similar membership; Legal support; Content moderation process manual; Trust and Safety resource library
- Overall a clear picture of folks doing well but knowing they need help and resources
- High number noting burnout, need to address

Next Steps

We have secured initial funding to obtain membership to TSPA. We are waiting on a draft contract to begin offering membership in phases, likely a handful to start with, then 20

We are talking with some legal resources to create core template documents and offer consultations by jurisdiction, in the meantime we have compiled some resources at https://fedifence.social/about

We will soon be creating a non-profit entity to accept donations to begin exploring ways to compensate moderators, and/or to offer moderation services. This entity will also seek to provide group access to regulatory and compliance documents and resources, legal services, and insurance.

Next Steps

As we build out the non-profit entity we will be seeking input from service providers and content moderators like you.

If you'd like to be part of guiding the formation and mission of the non-profit, please let us know!

The mission will be to support content and account moderation to help ensure moderation can scale with the anticipated influx of new members

Some ideas of possible services and resources are on the next page:

Idea Pad

- Facilitating co-op contracts to share/swap moderators
- Offering existing moderators paid hours on other servers
  - Could be simply account approval, trending content moderation, illegal content removal, or
  - Moderating content on “like” servers, or
  - Moderating a server in accordance with its own CoC
- Hiring and training content moderation staff to perform moderation on member servers
- Central funnel for illegal content/mandatory reporting requirements; insurance; TSPA and similar group memberships
- Directory of legal resources
- Central moderator wiki, forum, and chat group
- What else?