

Characterizing Bluesky Content Moderation Service: From Automation of Service to Landscape of Harms

Pushpdeep Singh^{1*}, Sayeh Jarollahi², Ayan Majumdar^{1,2}, Vabuk Pahari¹, Abhijnan Chakraborty³,
Krishna P. Gummadi¹, Ingmar Weber², Abhisek Dash¹

¹Max Planck Institute for Software Systems (MPI-SWS), Germany

²Saarland University, Germany

³Indian Institute of Technology Kharagpur, India

Abstract

Empirical research on content moderation is fundamentally constrained by the opaque deployment of moderation systems on major social media platforms. To this end, the recent emergence of decentralized platforms with transparent, public moderation logs presents an unprecedented opportunity for independent audits. In this work, we leverage this architectural transparency to conduct the first large-scale audit of the default moderation system on Bluesky, the *Bluesky Moderation Service* (BMS). Analyzing its 10.6M moderation labels from 2025, we investigate three foundational aspects: (i) its mechanism (the degree of automation versus human oversight), (ii) its efficacy (accuracy in detecting harms), and (iii) its purpose (the landscape of harms it identifies).

Our findings reveal a human-AI collaborative system where labels for sexual and graphic content are applied automatically in seconds, while nuanced and high stakes labels require more human oversight, taking hours or days. Through a manual annotation study, we find the BMS operates with high precision (0.837), but struggles with low recall (0.222), with our annotators identifying $4.5\times$ more harmful content than the moderation system in a random sample. Finally, unsupervised clustering of the most frequently applied labeled posts uncovers detected harms ranging from hostility in discourse toward protected groups to the spread of sexually explicit and other graphic content. Our work offers a look into the operational realities of a deployed moderation system, providing a concrete data-driven foundation for designing more effective and transparent moderation systems.¹

Code — <https://github.com/iampushpdeep/BMS>

Project — <https://bms-audit.github.io>

1 Introduction

Social media platforms deploy content moderation systems to govern the discourse on their forums. The stakes of such moderation are immense: while under-moderation may allow unchecked spread of unsafe content resulting in systemic harms, over-moderation risks stifling legitimate expression, leading to censorship (Gillespie 2018). Given this delicate balance, it is imperative to audit the functioning of deployed moderation systems on social media platforms.

However, independent audit of content moderation systems has been fundamentally constrained by the *non-transparent deployment* on established social media platforms (Gillespie 2018; Suzor et al. 2019; Juneja, Rama Subramanian, and Mitra 2020; Roberts 2016). Specifically, their moderation policies, decision-making algorithms, and enforcement data are proprietary: shielding them from independent scrutiny. In fact, such a lack of transparency has led policymakers to adopt regulations requiring greater transparency obligations for deployed content moderation systems. Concretely, the Digital Services Act (DSA) in the EU mandates platforms to disclose their moderation practices, including the degree of human oversight, the use of automated tools and the rationale behind moderation actions (Release 2022). However, even after these efforts, our understanding of prevalent systemic harms is limited due to *non-transparent reporting* by platforms (Trujillo, Fagni, and Cresci 2025; Shahi et al. 2025).

A new opportunity in this landscape arises from the emergence of decentralized platforms built on open protocols (Kleppmann et al. 2024; Raman et al. 2019; Balduf et al. 2024). By design, decentralized platforms offer greater transparency into their operations. For example, Bluesky (<https://bsky.app>), built on the Authenticated Transfer Protocol, publishes its moderation actions as a publicly accessible data stream (Kleppmann et al. 2024). Such architectural transparency transforms content moderation from an inscrutable ‘black box’ into an *auditable transparent public log*. This improved transparency enables some foundational inquiry into the *mechanism*, *efficacy* and *purpose* of the deployed content moderation systems.

Leveraging the publicly accessible moderation logs of Bluesky’s default moderation system (known as the *Bluesky Moderation Service* (BMS)²), in this study, we aim to answer the following important questions:

RQ1: *What is the degree of automation and human oversight in the Bluesky Moderation Service?*

RQ2: *How effective is the Bluesky Moderation Service in detecting harms on the platform?*

RQ3: *What is the landscape of harms detected by the Bluesky Moderation Service on the platform?*

Why are these questions important? Increasingly con-

*Corresponding author, psingh@mpi-sws.org

¹This work has been accepted for publication at ICWSM 2027.

²<https://bsky.app/profile/moderation.bsky.app>

tent moderation relies on *human-AI* collaborations (Bluesky 2025; Gillespie 2018; Halevy et al. 2022). To this end, the primary question regarding the moderation mechanism centers on the interplay between scalable automated detection and nuanced human oversight or interventions. While a number of platforms report such information through their transparency reports, due to the lack of procedural and outcome-level transparency, such details are rarely verifiable. RQ1 moves beyond platform assurances to empirically quantify the degree of automation and human oversight, providing a verifiable blueprint of the system’s underlying mechanism.

An understanding of the mechanism is insufficient without a rigorous evaluation of its performance, i.e., *how effective is the mechanism in practice?* While platforms may self-report high performance, such metrics are calculated on proprietary data and are impossible to independently validate. By assessing precision (what fraction of the detected content is actually harmful) and recall (what fraction of all harmful content has been detected), we aim to quantify the system’s real-world effectiveness.

Beyond the moderation mechanism and its effectiveness, another line of inquiry aims at the *purpose* of moderation systems by characterizing the landscape of harms it identifies. Such characterization is crucial for two reasons: (a) it reveals how closely stated principles, i.e., *moderation policies*, align with their *de facto* operationalization: *moderation practices*; and (b) it offers a ground-level view of the systemic harms present on the platform.

To answer these questions, we conduct the first large-scale audit of the default content moderation system deployed on Bluesky. It is applied by default to every user upon signing up on Bluesky, thereby establishing a universal safety standard for nearly 40M users on the entire platform (Bluesky 2026). We gathered and analyzed labels assigned by the BMS for 10.6M posts during 2025. Our analyses reveal the following key findings about BMS:

- **Findings for RQ1:** By using labeling delay as a proxy, our analyses show that the BMS labels sexual and graphic content quickly (e.g., median 7.1 seconds for *porn* label). Whereas for more nuanced labels, it relies more on human oversight, incurring longer delays (e.g., median 8.4 days for *rude* label). Our findings validate the disclosures made in Bluesky’s recent transparency report (Bluesky 2026).
- **Findings for RQ2:** By annotating a set of 1000 labeled posts and 1000 random posts from Bluesky’s firehose, we find that the BMS has high precision, but it suffers from low recall. While human annotators (coauthors) agree with 83.7% of all posts labeled by BMS in the set of 1000 labeled posts (precision 0.837), they flag $4.5\times$ more posts to be harmful in the 1000 random posts (recall 0.222).
- **Findings for RQ3:** By deploying unsupervised clustering on the labeled posts, we find several granular categories (within each label) of harmful content detected and moderated by BMS. While BMS detects more sexually explicit content, the most taken-down posts contain violent death wishes to political leaders. Our analyses reveal that many labels share overlap among the categories, which could lead to challenges for moderators and confusion for end-users.

2 Related Work

We position our work within three strands of prior work and provide a brief overview of each: (a) content moderation, (b) regulatory intervention, and (c) decentralized platforms.

2.1 Content Moderation

Content moderation has become a crucial component of social media platforms, which often mediate how consumers communicate and consume information online. The Digital Services Act (DSA) defines content moderation as *activities undertaken by online platforms aimed at detecting, identifying and addressing information incompatible with their terms and conditions* (Release 2022). Thus, it aims to strike a delicate balance between ensuring safety and empowering free speech (Gillespie 2018). Given its importance, a number of prior works have investigated content moderation on deployed online platforms. While some works study moderation mechanisms and their drawbacks (Halevy et al. 2022; Hartmann et al. 2025; Roberts 2019; Qiwei et al. 2024; Hartmann, Oueslati, and Stauer 2024; Ma and Kou 2022), some others delve into the transparency desiderata of moderation systems (Gillespie 2018; Suzor et al. 2019; Juneja, Rama Subramanian, and Mitra 2020; Roberts 2016). While valuable, the scope of such studies has often been constrained by the inherent opaqueness of traditional centralized platforms, making it difficult to comprehensively map the landscape of prevalent systemic harms and the corresponding moderation actions.

2.2 Regulatory Intervention

To improve transparency of centralized moderation systems, DSA requires all online platforms to submit statements of reasons for all undertaken moderation actions. This has resulted in the creation of the DSA Transparency Database (Comission 2025). However, several recent studies showcase the lack of compliance and inadequacies in the transparency database, ranging from limited reporting from platforms to a lack of information about content being moderated (Trujillo, Fagni, and Cresci 2025; Shahi et al. 2025). These inadequacies further limit the ability of researchers to understand the current state of systemic harms in digital platforms and the effectiveness of the detection strategies.

2.3 Decentralized Platforms

In this context, decentralized social media platforms (e.g., Mastodon, Bluesky) deployed on open protocols (e.g., Activity Pub, Authenticated Transfer Protocol) have emerged as more transparent alternatives, triggering new opportunities for public interest research. These platforms, deployed on open protocols, seem to be inherently more auditable than centralized platforms, which are increasingly becoming walled gardens by further restricting data access. Such data access has resulted in several large-scale studies on Mastodon (Raman et al. 2019; Zhang et al. 2024; Bono et al. 2024), and Bluesky (Kleppmann et al. 2024; Balduf et al. 2024, 2025) in recent years. Failla and Rossetti (2024) released a large-scale dataset of Bluesky user activity, covering over 4M accounts and 235M posts. Nogara

et al. (2025) conduct a longitudinal analysis of user activity around Bluesky’s public launch, examining misinformation dynamics, political polarization, and toxicity levels.

Baldur et al. (2024) provides a foundational contribution to the study of the Bluesky platform, and conducts the first large-scale architectural analysis of the platform and its sub-components: Labelers, Feed Generators, and Personal Data Servers, and provides an initial glimpse into the labeling ecosystem and third-party moderation providers. However, their focus remains infrastructural; the operation and evaluation of moderation decisions and the specific harms the labeler service addresses fall outside their scope of work.

To the best of our knowledge, this is the first study to perform a large-scale, empirical characterization of a transparent moderation system, analyzing its operational mechanisms, efficiency, and the specific harms it addresses.

3 Bluesky and its Content Moderation

Before we answer the key research questions, in this section we first provide a brief overview of the Bluesky ecosystem, its key architectural components, and its default moderation system, the Bluesky Moderation Service (BMS).

Identity and Userbase on Bluesky: Bluesky is a decentralized social media platform built on the Authenticated Transfer protocol (Kleppmann et al. 2024). When a user signs up on Bluesky, they are assigned a unique *Decentralized Identifier* (DID). Each DID can be resolved to a corresponding DID document maintained in the *Public Ledger of Credentials* (PLC) directory. To quantify Bluesky’s user base, we gathered the entire PLC directory using its ‘Export’ endpoint (<https://web.plc.directory/api/redoc#operation/Export>). We observe a total of 66.9M unique DID documents in the PLC directory. Such a large base of registered DIDs indicates the growing popularity of Bluesky and the need to study such evolving social media platforms. DID documents store information about where the corresponding user’s data repository resides, their user handle (human-readable identifier), etc. These DID documents are foundational to Bluesky’s architecture, acting as a *discovery mechanism* for some of its main components.

3.1 Key Architectural Components in Bluesky

In Bluesky, the different core functionalities are handled by separate architectural components (Kleppmann et al. 2024), namely: (a) Personal Data Servers, (b) Relay, (c) Labeler, (d) Feed generator, and (e) App View.

Personal Data Server (PDS): A personal data server authenticates users registered to it and hosts their data. Every Bluesky user must be registered with a single PDS. The service endpoint in a user’s DID document, specifically the `#atproto.pds` entry, points to the PDS that hosts their repository. This repository includes records of all posts, likes, comments, and other actions taken by a user.

Relay: Relay aggregates newly created user records across all active PDSes in the network to create a stream of records called *Firehose* that is publicly accessible.

Labeler: Labeler is the central content moderation engine

on Bluesky. Labelers function like a typical user account, and become a labeler by adding an `#atproto.labeler` service entry to their DID document. The labeler then publishes a `app.bsky.labeler.service` record in their repository to define and share their moderation policies. The labeler consumes the firehose as input and labels selected incoming records to generate a *label stream*—a publicly accessible log of actions taken by the labeler. It can apply labels to content, accounts, or profiles. The applied labels are leveraged to filter content and perform other moderation operations during content recommendation and dissemination.

Feed generator: Feed generator is the content recommendation arm of Bluesky. Any user can run their own feed generator that consumes the firehose and produces a relevant feed consisting of URIs pointing to filtered posts.

App View: The App View collates and indexes the data produced across the architecture by consuming the firehose and provides the results to users. In the context of content moderation, App View enforces the moderation labels on the user interface by taking the recommended actions.

3.2 Content Moderation on Bluesky

Primarily, labelers consume the firehose and label content that satisfies certain criteria, e.g., content violating terms of service or belonging to specific categories such as nsfw or disturbing imagery. Alternatively, Bluesky users can report posts and accounts to labelers, who can label them based on those reports. Thus, labelers are the core component responsible for content moderation on Bluesky.

Bluesky Moderation Service: To this end, labelers deployed by Bluesky PBC (BMS and country-specific moderation services) are applied to user accounts by default. While the open architecture of AT Protocol allows any third party to set up their labeler services, users are actively required to subscribe to them, which can naturally limit their reach. Hence, we leave the study of how third-party labelers complement or contradict the BMS for future work. While most of the analyses we propose here can be extended to any transparent moderation systems deployed on AT protocol, for this study, we primarily focus on the BMS.

BMS’ Content Moderation Policies and Labels: The policies associated with a labeler can be retrieved using the `app.bsky.labeler.service/self` record from the PDS using the `com.atproto.repo.getRecord` endpoint. For the BMS labeler, we collect these policies, which contain the list of labels they apply, their definitions, along with other attributes associated with each label, such as its *default action* initially determined by the labeler. Table 1 provides a summary of these label values, including their policy descriptions and default actions.

BMS’ labels include *intolerant*, *rude*, *porn*, *sexual*, *spam*, *!takedown*, etc. These labels can be broadly divided into two categories: (a) *harm labels*: refers to a certain category of harm (e.g., *intolerant*, *porn*, *self-harm*, etc.), (b) *action labels*: refers to an action taken by the Bluesky Moderation Service (e.g., *!takedown*, *!warn*, or *!hide*).

Crucially, the *default action* for each label, typically *hide*, *warn*, or *ignore*, determines how labeled content is presented

to a user who has not explicitly altered their preferences. *Hide* configuration hides posts with that label from the user’s feeds, *warn* provides a warning before the content is viewed, and *ignore* shows the label badge without any content filtering or visual obstruction. Based on the default action recommended by the labeler (or the modified setting of the user), the App View enforces the moderation labels to affect the Bluesky feed of the corresponding user.

Country-specific BMS: Furthermore, there are country-specific labelers for Germany, India, Turkey, Russia, and Brazil, which by default apply when a user connects to Bluesky from the mentioned countries. These labelers are also run by Bluesky PBC. However, unlike other moderation services, the labels applied by these labelers are not specified in their service record. Furthermore, they also do not provide users with any choice to change the moderation settings. Bluesky PBC deploys these country-specific moderation services to comply with legal or regulatory requirements imposed by national authorities.

Out of these five country-specific moderation services, only Germany, Turkey, and Brazil have issued labels. Turkey’s labeler issued a total of 31 labels, consisting of 6 *!hide* labels on *posts* and 25 *!hide* labels on *accounts*. The Germany-specific labeler issued 68 labels, almost all of which were applied to posts (66 *!hide* labels on *posts*) and 2 *!hide* labels on *accounts*. The Brazil-specific labeler issued 16 labels, including 13 *!hide* labels on *posts*, 1 *!takedown* label on a *post*, and 2 *!takedown* labels on *accounts*.

3.3 Data Collection for Moderated Content

Identifying BMS labeler: Labelers can be identified via their DID documents. DIDs with an `#atproto_labeler` service entry are labelers. DID document also contains the *serviceEndpoint* for the labeler, which can be used along with `com.atproto.label.subscribeLabels` API endpoint to retrieve the full label history available via the stream. We identified the *serviceEndpoint* for the BMS (with DID `did:plc:ar7c4by46qjdydhdevvrndac`) as `mod.bsky.app`.

Bluesky’s label stream data collection: As the label stream (`com.atproto.label.subscribeLabels`) on AT protocol (output of the labelers) is publicly accessible, to conduct an effective assessment of the BMS, we collected the label stream for the entire year of 2025. BMS applied 18,343,953 labels in total, out of which 15,744,081 were applied on *posts*, 2,418,534 on *accounts*, 84,647 on *profiles*, and 96,691 on other record types (e.g., *lists*, *feed generators*).

Additional metadata collection: We further gathered the details of the records labeled by Bluesky. The label stream only contains the Content Identifiers (CIDs) and AT URIs of labeled posts, not the complete post records (with metadata about when the post was created, embed information, post text, media content identifiers etc.). As an AT URI is of format `at://<did>/<collection>/<rkey>`, we use the DID to obtain the corresponding PDS endpoint via the PLC directory, and then retrieve the associated record using `com.atproto.repo.getRecord`.

For records with media objects, we fetch the corresponding blob (`image/thumb/video`) using `com.atproto.sync.getBlob` endpoint.

We were able to gather post records corresponding to 10,681,824 post labels.³ Table 2 presents a breakdown of the most frequently applied labels in 2025 along with their counts in our collected dataset. The most prevalent label is *porn* (7.3M posts), followed by *sexual* (2.2M) and *nudity* (271K). Based on these records, we also list distributions of the type of post (root vs. reply), embed-types and languages across the labeled posts in Table 5 (Appendix A).

Bluesky posts support five embed types: `app.bsky.embed.images` attaches up to four images per post; `app.bsky.embed.video` embeds an uploaded video; `app.bsky.embed.external` embeds an external URL as a thumbnail; and `app.bsky.embed.record` embeds a reference to another post, optionally combined with media via `app.bsky.embed.recordWithMedia`. Posts with none of these embed fields are counted as *text-only*.

Across all modalities, images dominate across many labels, including *porn*, *sexual*, *nudity* etc., whereas, labels like *rude* and *!takedown*, show a high proportion of *text-only* posts. Overall, 88% of labeled posts are root posts, though this varies markedly: labels such as *porn*, *sexual* and *nudity* are overwhelmingly root posts (91–92%), whereas posts labeled as *rude* (96%), *spam* (99%), and *threat* (79%) skew heavily toward replies. English is the dominant language across all label categories, ranging from 49% (*self-harm*) to 97% (*rude*) of all posts.

Our subsequent analyses focus on this set of 10,681,824 post labels applied by the Bluesky Moderation Service (including country-specific services) covering 13 label values. We provide the counts of labels on *accounts* and *profiles* in Appendix A. Similarly, we were able to collect 69 post records, all labeled as *!hide* by the country-specific BMS.

4 RQ1: Automation and Human Oversight in the Bluesky Moderation Service

For any given content moderation system, the Digital Services Act requires platforms to be transparent regarding the degree of human oversight and usage of automated tools in their decision making. Bluesky, in its terms of services and transparency reports, clarifies that it uses a human-AI collaborative approach for operationalizing its moderation policies (Bluesky 2025; Rodericks 2025; Bluesky 2026). However, to this date, apart from transparency reports from platforms, there is no other means to quantify and ascertain the claims made by platforms. Therefore, in this section, we analyze the degree of underlying Human-AI collaboration in Bluesky’s content moderation pipeline.

While the BMS has its label stream (i.e., outcome of its labeling process) transparent and publicly accessible, the exact methodology adopted by the BMS is largely non-transparent and proprietary.

³The remaining records could not be retrieved due to missing repositories, deleted records, or unavailable PDS endpoints.

Table 1: Labels used in our analysis, along with their names, descriptions, and their recommended default action.

Label	Policy Description	Default Action
intolerant	Discrimination against protected groups.	warn
rude	Rude or impolite, including crude language and disrespectful comments, without constructive purpose.	hide
threat	Promotes violence or harm towards others, including threats, incitement, or advocacy of harm.	hide
sexual-figurative	Art with explicit or suggestive sexual themes, including provocative imagery or partial nudity.	show
porn	Explicit sexual images.	hide
sexual	Does not include nudity.	warn
nudity	E.g., artistic nudes.	show
graphic-media	Explicit or potentially disturbing media.	warn
self-harm	Promotes self-harm, including graphic images, glorifying discussions, or triggering stories.	warn
spam	Unwanted, repeated, or unrelated actions that bother users.	hide
!hide	This content has been hidden by the moderators.	—
!warn	This content has received a general warning from moderators.	—
!takedown	This content has been removed according to policy.	—

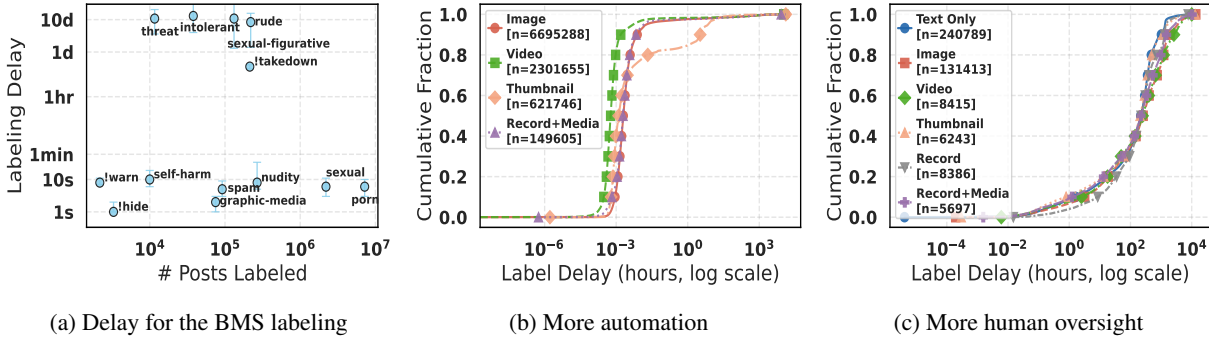


Figure 1: (a) Median labeling delay shown with the number of posts labeled. (b) Cumulative distribution of delays for labels with more automation. (c) Cumulative distribution of delays for labels with more human oversight. *self-harm*, *porn*, *!warn*, etc. seem to be more automated while *threat*, *rude*, *!takedown* etc. may involve more human oversight across major post types.

Table 2: BMS label distribution in our collected dataset.

Label	Count	Label	Count
<i>porn</i>	7,360,828	<i>spam</i>	78,077
<i>sexual</i>	2,243,593	<i>intolerant</i>	38,021
<i>nudity</i>	271,246	<i>threat</i>	11,640
<i>rude</i>	219,712	<i>self-harm</i>	10,187
<i>!takedown</i>	214,079	<i>!hide</i>	3,520
<i>sexual-figurative</i>	134,942	<i>!warn</i>	2,191
<i>graphic-media</i>	93,788		

Labeling delay as an empirical proxy: To better understand the degree of human oversight versus automation across different labels, we use labeling delay as an empirical proxy. For each labeled post, we define labeling delay as the difference between the time the post was created by the user (`createdAt`) and the time the labeler applied its label (`cts`). We hypothesize that:

Labels applied at scale within seconds are likely automated, whereas labels requiring hours or days indicate involvement of significant human oversight.

Posts were filtered out if they were backdated (created before the cutoff date, i.e., June 2024 or had negative labeling delays due to incorrect `createdAt`⁴ timestamps).

⁴Timestamp generally represents the time the record was “first seen” by an API server, though it might also be updated if the

Automated labels: For each label, we compute the median labeling delay with inter-quartile ranges. Figure 1a shows the labeling delay (on Y-axis) along with the number of labeled posts for each label (on X-axis). It shows a clear dichotomy in the timeliness of moderation. Labels such as *porn*, *sexual*, *nudity*, *self-harm*, and *graphic-media* are applied very quickly, with median delays typically a few seconds. For example, the median labeling delay for *porn* was 5.5 s and the upper quartile (Q3) delay was 9.8 s, already covering more than five million applied labels in our collected dataset. This pattern indicates that these labels are predominantly applied automatically. Figure 1b shows the cumulative distribution of labeling delay for these automated labels for content with different embed types. We observe that most of the posts across different embed types: images, video, thumbnails, and embedded records with media are labeled very quickly. However, posts with thumbnails exhibit longer delays for a significant fraction of cases as thumbnail embeds refer to external third-party URLs rather than media hosted directly on Bluesky’s CDN, introducing additional fetch and crawl latency before labels can be assigned.

While the above observations help us identify the potential degree of automation needed in some of the labels, there

record is edited. Using only this timestamp would mean that bulk-imported records would all get timestamps very close together around the import time, which may not be correct. <https://docs.bsky.app/docs/advanced-guides/timestamps/#createdat>

are some interesting insights if we closely look at the labeling delay distribution. Table 3 shows the median, 5th, 25th, 75th, and 95th percentiles of the labeling delay distribution of the different labels. For some of the potentially automated labels, e.g., *nudity* and *self-harm*, the 95th percentile of the labeling delay is of the order of a few days. This, coupled with the lower number of these labels, further indicates a greater degree of human oversight on some of these categories, albeit they are predominantly automated as indicated by the median of the distributions.

Need for automated moderation: Automated labeling is faster and scalable i.e., automated tools can detect harmful posts faster and in larger numbers. Faster labeling of harmful content not only reduces their dissemination and exposure to users, but also reduces the mental health and other societal issues for human moderators. BMS uses automated tools to reduce the spread of sexual and graphic content on its platform to keep it a safer digital place for its users.

Higher degree of human oversight: In contrast, labels such as *intolerant*, *rude*, *threat*, and *sexual-figurative* exhibit much longer median delays, ranging from hours to days, reflecting a significant likelihood of human intervention. Furthermore, these labels show larger interquartile ranges, indicating the variability in human moderation times. When it comes to variation across content with different multi-media attachments, Figure 1c shows that irrespective of the type of content, the labeling delay follows similar patterns.

Need for human oversight: While moderating harmful content at scale is one of the primary goals of any moderation system, preserving free-speech is also another important aspect of this activity. Given the vast array of issues being discussed on social media, some could potentially hurt the sentiments of an individual or a group of people. However, often a broader context and perspective are needed to disentangle absolute violations of community guidelines. For example, criticism of political parties or opinions should be allowed to preserve and empower free-speech. However, uncivil, impolite, and abusive posts, even in such contexts, need to be labeled (if not removed). At the same time, the delay we observe for some of the automated labels could be because of the inherent subjectivity of some of the situations. For example, although posts detailing war-time atrocities (Gillespie 2018) or breastfeeding (Sweney 2008; Arthur 2012) could be identified as *graphic-media* or *nudity* respectively, whether they *should* be labeled is often subjective. Our analyses bring to light the thoughtful effort taken up by the BMS for detecting harms prevalent on Bluesky.

Additional analyses on country-specific moderation services and Bluesky’s label redressal mechanism, both of which exhibit higher human oversight, are provided in Appendix D.

Validation of our approach: While labeling delay serves as a proxy for distinguishing automated from human-in-the-loop labels, AT Protocol offers no independent means to validate this. Nevertheless, Bluesky’s 2025 transparency report (Bluesky 2026) corroborates our findings: *rude*, *intolerant*, and *threat* are confirmed as manually applied, while *nudity*, *porn*, *sexual*, *graphic-media*, and *self-harm* have significantly lower human oversight. Their reported manual-

Table 3: Labeling delay for different labels.

Label (N)	5%ile	25%ile	Median	75%ile	95%ile
Potentially automated label					
porn (7,198,690)	1.3 s	2.8 s	5.5 s	9.8 s	55.0 s
sexual (2,200,976)	1.5 s	3.3 s	6.2 s	11.1 s	2.7 min
nudity (267,334)	2.1 s	4.6 s	8.1 s	33.9 s	32.0 d
graphic-media (91,832)	1.5 s	3.0 s	4.9 s	9.1 s	52.4 min
self-harm (9,972)	2.1 s	5.7 s	10.2 s	19.0 s	10.3 d
spam (74,984)	0.7 s	1.1 s	1.7 s	2.9 s	12.2 s
!warn (2,164)	5.8 s	7.0 s	8.1 s	10.8 s	26.8 s
!hide (3,294)	0.5 s	0.9 s	1.3 s	2.5 s	15.7 s
Labels with higher human oversight					
!takedown (213,677)	2.4 hr	7.3 hr	8.5 hr	9.7 hr	13.8 d
sexual-figurative (131,891)	20.4 min	1.3 d	10.8 d	43.1 d	150.7 d
intolerant (37,894)	2.5 hr	4.0 d	13.0 d	47.0 d	80.4 d
rude (219,557)	11.7 min	1.4 d	8.4 d	15.7 d	52.3 d
threat (11,601)	2.0 hr	3.6 d	10.8 d	20.5 d	58.7 d

application rates for *self-harm* and *nudity* (8% and 18%, respectively) align with our observation that delays for these labels extend to hours or days. For *sexual-figurative*, we observe high delays suggesting human oversight despite the 2025 report listing 0% manual application (Bluesky 2026). Their 2024 moderation report (Rodericks 2025) resolves this discrepancy, clarifying that this label was predominantly applied when users appealed automated misclassifications of figurative art, confirming that the observed delays do reflect human involvement.

5 RQ2: Evaluation of Effectiveness of the Bluesky Moderation Service

The label delay analysis establishes that the BMS operates as two distinct pipelines: an automated system handling visual content and a human oversight system handling specific labels. A natural question follows: *how accurate (or effective) is each pipeline and the BMS labeling of harmful content overall?* To answer this, we evaluate the system along two dimensions – precision (of what the BMS flags, how much is actually harmful?) and recall (of all harmful content, how much does the BMS flag?)

5.1 Precision and Recall of the BMS

Evaluation sets and annotation procedure: Measuring precision requires a sample drawn exclusively from contents flagged by the BMS, so that human judgments can verify their accuracy. On the other hand, measuring recall requires a random sample drawn from the content stream i.e the firehose, so that humans can flag unsafe or harmful contents and compare the same with what the BMS might have flagged in the random sample. To evaluate precision, we randomly sample 1,000 labeled posts from our dataset (**labeled set**), stratified across all nine harm categories (*intolerant*, *rude*, *threat*, *self-harm*, *graphic-media*, *porn*, *sexual*, *nudity*, *sexual-figurative*) to ensure coverage of both automated and

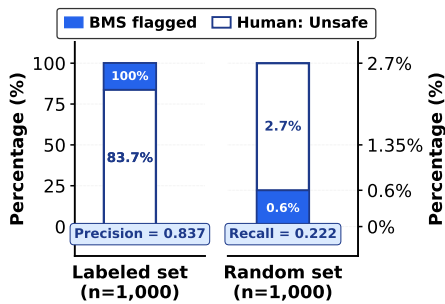


Figure 2: Unsafe content in the labeled set (left) and random firehose set (right). Blue = Flagged by the BMS, white = adjudicated unsafe by majority human annotators. Bluesky achieves high precision (0.837) on the labeled set, but low recall (0.222) on the random firehose.

human-oversight labels. Furthermore, to evaluate recall, we randomly sample 1,000 posts from the Bluesky firehose (input of the labeler) data in 2025 (**random set**).

Annotation procedure: Given the BMS policies, two coauthors marked the posts in both the evaluation sets as safe and unsafe. If they mark a post to be unsafe, they were additionally asked which harm label should be assigned to the given post and why. In case of disagreement, a third co-author performed the same task acting as a tiebreaker. Across the full evaluation set (n=2,000), the primary annotators achieved an inter-annotator agreement, cohen’s kappa $\kappa = 0.802$ in the binary task of determining whether a post is safe or unsafe. They agreed on 1,827 (i.e., 91.35%) of the posts and the tiebreaker annotator annotated the remaining posts.

Precision of the BMS: Table 4 shows the detailed results on the evaluation set that contained 1000 posts labeled by the BMS. Note that we report the agreement only for the binary setting i.e., agreement between our majority voting for a post (safe vs. unsafe) and if the BMS has labeled it (irrespective of the label). Across all label categories, annotators agreed with the BMS’ assessment in 83.7% posts, indicating a significant percentage of posts flagged by the BMS to be genuinely unsafe i.e., the BMS has high precision (0.837) in detecting harms on Bluesky. But, this aggregate conceals substantial variation across labels and post structures.

In the automated labels, our binary annotation (safe vs unsafe) and the BMS’ flagging achieve an agreement of 90.3% (i.e., BMS’ precision of 0.903). In the *porn* label, the BMS achieves a perfect precision (1.000, n=111): every post the BMS flagged was annotated as unsafe by majority of the annotators. Similarly, in *self-harm* (0.919), *nudity* (0.937), and *sexual* (0.856), the BMS achieves high precision in detecting harmful content that is inherently sexual or self-harm related. However, its precision drops to 0.802 for *graphic-media*. On the other hand, the BMS’ precision among manual labels drops to 0.755. Among manual labels, the BMS achieves high precision in *sexual-figurative* (0.955), and *threat* (0.892). However, its precision drops significantly for *intolerant* (0.667) and *rude* (0.509).

Root posts vs. Replies: In further analyses, we observed that a large percentage of the posts labeled by us in *intolerant* (79%), *rude* (97%), and *threat* (85%) are in fact replies to some other posts. This indicates a potential lack of addi-

Table 4: BMS’ precision per label broken down by post type.

Label	Total		Root		Reply	
	n	Precision	n	Precision	n	Precision
Automated Labels						
<i>porn</i>	111	1.000	102	1.000	9	1.000
<i>sexual</i>	111	0.856	95	0.884	16	0.688
<i>nudity</i>	111	0.937	100	0.940	11	0.909
<i>self-harm</i>	111	0.919	99	0.939	12	0.750
<i>graphic-media</i>	111	0.802	86	0.814	25	0.760
Total	555	0.903	482	0.919	73	0.795
Manual Labels						
<i>intolerant</i>	111	0.667	23	0.609	88	0.682
<i>rude</i>	112	0.509	3	0.667	109	0.505
<i>threat</i>	111	0.892	16	0.812	95	0.905
<i>sexual-figurative</i>	111	0.955	104	0.952	7	1.000
Total	445	0.755	146	0.877	299	0.696
Overall	1000	0.837	628	0.909	372	0.715

tional context which Bluesky’s moderators might have had (e.g., the root post and/or authors to which these replies were directed) which was unfortunately not provided to our annotators as they annotated each individual post independently treating them as an isolated post. In fact, in many such cases, we observe that annotators have clearly marked that they need more context to adjudicate such posts as unsafe. Overall, while our annotations agree with the BMS’s flagging in 71.5% cases for posts which were replies, the agreement increases to 90.9% for root posts.

Disagreement analysis: To characterize the false positives qualitatively, we examined a random sample of 50 posts judged *safe* by majority vote despite carrying a Bluesky label and observe some noticeable patterns. Firstly, as discussed earlier *subjectivity and lack of context*— posts where annotators evaluated reply text without access to the parent thread, making it impossible to assess tone, intent, or conversational register — content that could or could not be harmful depending on the conversational context in which it is produced. Secondly, *ambiguous visual content*— images that may resemble harmful content, but carry benign intent in context e.g., body art mimicking self-harm imagery, film and media stills flagged as graphic, pop album covers without sexual intent etc. for a notable share of automated label false positives. Overall, these cases also reflect inherent subjectivity of the labeling task: annotators disagreed substantially on what constitutes harmful expression for labels such as *rude* and *intolerant*, underscoring that such judgments are sensitive to individual thresholds and cultural background.

Recall of the BMS: In the **random set**, we label a total of 27 posts to be unsafe i.e., 2.7% of all posts in the random set of 1000 posts. Out of the 27, only 6 had been labeled by the BMS, yielding a recall of 0.222. On the positive side, the BMS achieved a 100% precision in the random set where all posts labeled by the BMS are adjudicated as unsafe by human annotators. Notably, all 6 posts (flagged by the BMS) carried automated labels. Therefore, its recall for manually applied labels for this random set is 0. This is consistent with

the operational reality of human-oversight labels: moderators usually only review reported or surfaced content and cannot read every post at a platform scale, so content of this type can go unlabeled unless explicitly reported. Of the 10 posts human annotators identified as automated-label harm, Bluesky caught 6 and missed 4, resulting in a recall of 0.600. Such low recall for automated labels raises a question: *why does Bluesky’s automated pipeline fail to catch these posts?*

5.2 Blindspots of Automated Moderation

To diagnose the potential reasons for low recall of the BMS’ automated labeling, we first describe the technical architecture underlying Bluesky’s automated moderation, then design a setup to study the potential blindspot of the pipeline.

Bluesky’s automated moderation pipeline is powered by two components: (i) **Hive AI**⁵, a commercial API providing multi-head vision classifiers for moderation. When a post containing an image (individual frames in case of a video) is submitted, Bluesky calls Hive API to get a set of class-score pairs for the image, with each class representing a specific visual concept (e.g., `yes_sexual_activity`, `yes_self_harm`, `very_bloody`) with a confidence score between 0 and 1. Note that Hive does not return a single label; it returns a probability distribution across a hundred visual categories simultaneously. (ii) **Automod**⁶, Bluesky’s open-source rule engine that translates Hive’s raw scores into Bluesky’s label vocabulary via hard-coded rules⁷. For example, `yes_self_harm` ≥ 0.96 triggers the *self-harm* label, while sexual content follows a priority cascade: *porn* is checked (using multiple head scores) before *sexual*, which is checked before *nudity*. Critically, Automod reads only a subset of Hive’s output heads; scores on all others are ignored, regardless of magnitude. Details on Hive classes and Automod are in Appendix C.

Experimental setup: To investigate the low recall for automated labels applied by the BMS, we start with our labeled set of 1,000 posts. We filtered to posts carrying one of the five automated labels and containing a single image, yielding a set of 336 labeled posts distributed across *porn* (78), *sexual* (67), *nudity* (72), *self-harm* (57), and *graphic-media* (62). For each labeled post, we retrieved the most semantically similar image post from a firehose set carrying no Bluesky label, yielding a second set of 336 unlabeled posts. This firehose set is created by randomly sampling 40M posts between March and June 2025. Similarity is computed using cosine similarity between embeddings of labeled post and firehose posts (indexed using faiss) generated using Qwen3-VL-Embedding-2B. Details are provided in Appendix B. These unlabeled posts were selected for their visual and semantic proximity to known harmful content, making them very likely candidates for missed moderation,

thus helping us investigate recall failures. Both sets were passed through Bluesky’s Automod pipeline via the Hive API⁸, obtaining raw classifier scores across all heads.

Validating Automod as a proxy: Before analyzing the unlabeled set, we verified that Automod faithfully reproduces Bluesky’s labeling decisions. Of the 336 labeled posts, Automod independently flagged 311 (92.6%), confirming high agreement with Bluesky’s deployed moderation system. This slight mismatch could be due to API versions: developers can only access Hive’s V3 API, whereas enterprises have access to V2. Nonetheless, our findings show that we can treat Automod as a reliable proxy for Bluesky’s automated pipeline in subsequent analyses.

Identifying Missed Content: Of the 336 most similar unlabeled posts, Automod flagged only 47 (14.0%), leaving 289 posts flagged by neither BMS nor Automod. We manually annotated these 289 posts: two annotators independently judged 36 of them as unsafe content ($\kappa=0.63$, substantial) Bluesky’s system had missed entirely. This corresponds to 10.7%⁹ of all 336 similar unlabeled posts, establishing a lower bound on recall loss for automated labels.

Blindspots: We examined the 36 missed posts against Hive’s classifier outputs and identified two failure modes:

H1 – Close threshold misses (22 posts, 61.1%): Hive’s visual concept classes produce confidence scores in the near-miss zone [0.55, threshold) – the classifier detects the content as potentially harmful on heads Automod uses, but falls short of threshold i.e it is excluded due to a conservative threshold. Figure 3 (left) shows the score distributions for each near-miss head, sorted by median gap to threshold. Heads like `yes_sexual_intent` and `yes_self_harm` have median gaps of 0.28 and 0.21, while `yes_female_underwear` (median gap=0.08) and `yes_male_underwear` (median gap=0.07) are the closest – consistently below Automod’s threshold(0.98).

H2 – Rule set gap (14 posts, 38.9%): No visual concept class from Hive (considered by the Automod) approaches the threshold, but classes outside Automod’s rule set score ≥ 0.80 . Figure 3 (right) shows these soft classes sorted by median score. `yes_cleavage` (n=7, median=1.000) and `yes_male_shirtless` (n=2, median=1.000) achieve perfect or near-perfect scores, while `general_suggestive` dominates by volume (n=14, median=0.979). The content passes through Automod entirely because no rule acts on the presence of these visual categories.

Brittleness of rule-based moderation: These findings expose the brittleness of a rule-based system sitting on top of classifiers. Automod’s rules are hard-coded: with a fixed (and potentially arbitrary) threshold and fixed set of heads. This rigidity could be by design: having explicit rules is auditable, explainable, and deployable at scale. However, a post scoring 0.89 on `yes_sexual_activity` is treated

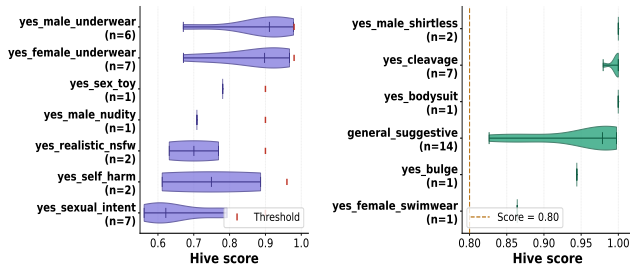
⁵<https://docs.thehive.ai/docs/visual-content-moderation>

⁶https://github.com/bluesky-social/indigo/blob/main/automod/visual/hiveai_client.go

⁷The Automod code was open-sourced in April 2024. It is difficult to establish whether the BMS still uses the exact rules or not. However, this is the best available approximation for the automatic moderation rules provided by Bluesky itself.

⁸We investigate on a smaller set since Hive AI API has a rate limit of 100 requests/day for developers.

⁹The relatively low proportion of unsafe posts, even among the closest semantic matches to known harmful content is expected: semantic similarity is computed over a general embedding space where proximity reflects similarity across all topical and visual features jointly, and not exclusively along the harm subspace.



(a) H1: close threshold misses (b) H2: Rule set gap

Figure 3: (a) Hive scores on Automod-mapped heads sorted by median gap to hard-coded thresholds (red markers); some harmful posts narrowly miss. (b) Hive scores on heads outside Automod’s rule set sorted by median score; Automod does not act despite near-perfect scores.

identically to one scoring 0.10, despite being qualitatively far closer to a flagged post (H1). A post with strong signal on, say, `yes_bulge` or `general_suggestive` is missed by such a system not because the content is ambiguous, but because there is no rule written for those heads (H2). In both cases, the failures are due to the translation layer between continuous scores and binary moderation decisions. In this regard, Vision-Language models (VLMs) offer a more general approach to moderation: by conditioning on the policy text itself, they can reason toward its intent rather than being confined to a fixed set of thresholds and heads (Majumdar et al. 2026).

6 RQ3: Landscape of Harms Detected by the Bluesky Moderation Service

The BMS provides each label along with its policy description (see Table 1). As described in the previous section, these descriptions are often understandably abstract or insufficiently detailed, raising the difficulty of understanding any additional intended nuances. However, due to the availability of the label stream, the outcome of the labeling process is transparent. Such transparency not only enables better understanding of how the mentioned policy descriptions are interpreted and operationalized by moderators, but also empowers us to have a better understanding of the kind of harms prevalent on Bluesky. In this section, we aim to analyze the posts labeled by the BMS to characterize its detected harms.

6.1 Clustering of Labeled Posts

To analyze patterns in posts labeled by the BMS, we perform a multi-modal, unsupervised semantic analysis of the labeled content using clustering. Our goal is to examine the dominant content themes within each label to understand the prevalent online harms and compare these themes with the label’s policy description to understand the potential interpretations being operationalized.

Data preparation: For each label, we retain posts that carry text, images, or videos. We exclude¹⁰ posts where records

¹⁰We observed a lot of duplicates for `!takedown` which contained 210,516 posts but only 31,129 unique records after deduplication.

were available, but failed downloading blobs or with the embed type `app.bsky.embed.record` (i.e., pure quote-posts or record embeds with no standalone media), which have incomplete information. The final post counts used for clustering per label are reported in Table 9 in the Appendix.

Embeddings and clustering: We generate a 2048-dimensional embedding for each post using Qwen3-VL-Embedding-2B (Li et al. 2026), a vision-language model that jointly encodes text and media (images, video frames (R3-6)). These embeddings are projected into a low-dimensional manifold using UMAP (McInnes, Healy, and Melville 2020) with cosine distance, then clustered using HDBSCAN (Campello, Moulavi, and Sander 2013). Hyperparameters (*n_neighbors*, *min_cluster_size*, *min_samples*) are jointly tuned per label via Bayesian optimization (Optuna TPE sampler (Akiba et al. 2019)), evaluated using the Density-Based Clustering Validation (DBCV) score (Moulavi et al. 2014), an intrinsic measure rewarding internally dense, well-separated clusters. DBCV scores range from -1 (poor) to $+1$ (ideal), and the configuration maximizing this score per label is selected. For labels exceeding one million posts (*porn* and *sexual*), we uniformly sample 300,000 posts and perform all steps on them, ensuring computational efficiency while preserving semantic diversity for meaningful clusters.

Full details of the hyperparameter tuning and selected hyperparameters are in Appendix E. Posts not assigned to any cluster by HDBSCAN are designated as noise; these represent posts that do not belong to any sufficiently dense region of the embedding space and are excluded from our analysis. Noise fractions vary substantially across labels (Table 9), reflecting heterogeneity in some label categories.

Cluster interpretation. After obtaining the resulting clusters, we analyze the posts within each cluster. Due to the sensitivity of labeled content, particularly label values such as *graphic-media* and *self-harm*, we adopt an automated pipeline to avoid exposing human annotators to potentially harmful material (Arsht and Etcovitch 2018; Spence et al. 2023) and validate it on a subset of clusters. Posts are sampled per cluster using the Cochran (Cochran 1977) finite-population-corrected formula (Eq. 1 in the Appendix E), weighted by HDBSCAN soft membership probabilities, and passed to Qwen3-VL-32B-Instruct using a map-reduce summarization strategy to generate a cluster description that describes the theme of posts in that cluster which is then used to get a name for the cluster. The full pipeline including the sampling, prompts used, LLM generation parameters is described in Appendix E.

Human validation. For human validation of the assigned cluster names, we randomly selected 31 clusters across 7 labels and asked three annotators to independently assess the appropriateness of the cluster names. Each annotator was provided with 10 sample posts from the cluster along with the cluster name and was asked to rate the cluster names on a five-point Likert scale ranging from *very inappropriate* to *very appropriate*. Cluster names received a mean rating of $\mu = 4.58$ (SD = 0.63), with 92.5% of ratings at *appropriate* or *very appropriate* (≥ 4) with substantial inter-annotator agreement (Krippendorff’s $\alpha = 0.686$) (Landis and Koch

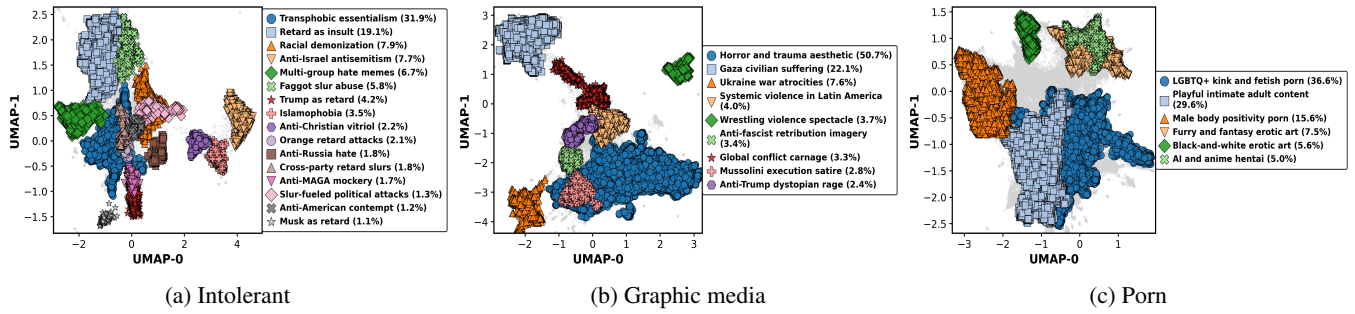


Figure 4: Visualizations of prominent clusters for representative labels that detect (a) Harms to protected groups, individuals, and civic discourse (b) Harms to physical and mental well-being, (c) Harms to minors from sexual content.

1977) indicating our pipeline produces cluster names largely faithful to underlying post content. Full details are in Appendix E.

6.2 Detected Harmful Contents by Bluesky

Figure 4 shows the top clusters observed in *intolerant*, *graphic-media* and *porn* labels, respectively. (Please refer to Figure 7 in Appendix E for other labels). Next, we describe the prominent topics within each moderation label.

Harms to protected groups, individuals, and civic discourse: Figure 4a shows the prominent clusters found among posts labeled as *intolerant*. The BMS defines intolerance as ‘discrimination against protected groups’. To this end, our clusters show potential interpretations and operationalization of this abstract nuance. For example, we observe clusters of posts targeting people based on their gender identities (cluster: Transphobic essentialism) e.g., LGBTQ+ identities (cluster: faggot slur abuse), religious identity (cluster: islamophobia, anti-christian vitriol) e.g., anti-semitic, anti-muslim content, racial identity (cluster: racial demonization) e.g., toward black, white and immigrant communities, political preference targeting political figures and their supporters, with posts centered on mocking Trump (clusters: Trump as retard, Orange retard attacks) and MAGA supporters (cluster: Anti-MAGA mockery), alongside broader cross-partisan attacks (clusters: Cross-party retard slurs, Slur-fueled political attacks).

rude and *threat* labels are an interesting case study, which could be thought of as a stronger degree of incivility than intolerance. Clusters in the *rude* label include posts characterized by personal attacks, vulgar insults, often reflecting deep political polarization. These manifest across several rhetorical patterns: targeting individuals with slurs and commands (clusters: Eat s**t insults, Go f*** yourself, F*** off demands), *identity-based attacks* deploying misogynistic slurs or accusations of pedophilia and nazism to dehumanize opponents (clusters: Misogynistic slurs, Pedophilia accusations, Nazi/fascist slurs), and *politically charged hostility* directed at figures and movements across the spectrum (clusters: Anti-MAGA vitriol, Political tribal rage). Other clusters reflect outrage over ongoing geopolitical conflicts (clusters: Israel-Palestine hate, Genocide accusations), generalized contempt expressed through mockery (clusters: Idiot accusations, Stupidity mockery, Loser attacks).

Clusters within *threat* label include similar topics, with

an even higher degree of incivility wherein authors express extreme hostility and violent ideas, including calls for death and/or violence, celebrating the pain of others, directed at political figures (clusters: Anti-Trump death wishes, Broad political murder fantasies, Cross-party political violence), geopolitical adversaries (clusters: Anti-Putin nuclear rage, Anti-Israel genocidal rhetoric), institutions (cluster: Anti-ICE assassination calls), and public figures such as billionaire elites (cluster: Anti-Musk vigilante memes).

Harms to physical and mental well-being: For *graphic-media* label (shown in Figure 4b), clusters group posts containing explicitly violent, gory imagery, often documenting real-world atrocities or leveraging graphic visuals to express extreme political sentiment. Some clusters revolve around historical execution photography repurposed for contemporary political hostility; others are about the war-related civilian deaths in Gaza or Ukraine, where graphic visuals of injured or deceased people (including children) are shown. Several clusters also include graphic stills from horror films or wrestling, showing blood or stylized gore, mixing violence with dark humor. We also observe self-harm-related posts getting labeled as *graphic-media*, especially posts with images of self-inflicted wounds.

Clusters within the related *self-harm* label include posts explicitly depicting or discussing self-inflicted injury, especially cutting, often including hashtags such as #sh or #shtwt. Other categories of posts include dark-humor memes involving suicidal themes, such as repeated edits of a man holding a gun to his head juxtaposed with cartoon characters in meme format. In contrast, one cluster includes posts sharing the historical 1863 photograph of an enslaved man’s scarred back as political commentary. These posts are grouped under *self-harm* despite using the image to protest the erasure of slavery history (cluster: Slavery scars vs. Trump erasure). Since *self-harm* is a largely automated label, this case suggests that automated moderation systems can misclassify the *type of harm* depicted in content based on visual appearance alone, without accounting for the broader context or intent behind the imagery.

Harms to minors from sexual content: The BMS uses four different labels to identify content that may be sexually explicit. Figure 4c shows the prominent clusters of *porn* label. Contents within these clusters include highly graphic content spanning kink and fetish pornography, ex-

PLICIT male body-focused content, AI-generated and anime-style hentai, and furry or fantasy erotic art with exaggerated anatomy—frequently accompanied by monetization links to platforms such as OnlyFans and Fansly. The *porn* label detects overtly explicit and commercial adult material on Bluesky and is the top applied label. On the other hand, the *sexual* label detects content which could be deemed inappropriate but is often significantly less explicit. These clusters contain labeled posts encompassing erotic body-focused self-presentation, monochrome sensual art, anime fan art, and queer NSFW art communities, with both real-user imagery and AI-generated content, which may pose difficult challenges to content moderation systems.

Similarly, *nudity* label detects a broad spectrum of non-explicit nude content, from fine art nude photography, life drawing, and naturism to casual adult selfies. Finally, *sexual-figurative* label primarily contains anime, furry, or digitally rendered characters with exaggerated sexual features. These clusters often reflect niche online subcultures built around franchises like Pokémon, Marvel, and Final Fantasy, and include a strong commission-based creator economy (Patreon, YCH auctions). Compared to the other labels, sexual-figurative consists more of fictional or stylized bodies.

Similarly, Bluesky also issues action labels such as *!hide*, *!warn* and *!takedown*, which enforce moderation decisions for all users. This also includes labels issued by country-specific moderation services. Most of the clusters in these action labels are centered around extreme political hostility and posts promoting conspiracies and extreme violence. For brevity and better readability, we place the detailed discussion on action labels and *spam* in Appendix F.

6.3 Implications of the Observations

Across labels, our analyses point to potential correspondences between the observed landscape of harms and the systemic risks defined in DSA Article 34(1) (Release 2022).

The clusters within *intolerant*, *rude*, and *threat* – esp., those involving hate speech – may map to risks concerning *illegal content* (some forms of hate speech) (Art. 34(1)(a)) and negative effects on *fundamental rights* (Art. 34(1)(b)), particularly the rights to *non-discrimination* and *human dignity* and implicate risk to *civic discourse* (Art. 34(1)(c)).

Similarly, the *graphic-media* label, by identifying violent and gory imagery, hints at potential threats to *public security* (Art. 34(1)(c)) and users’ *mental well-being* (Art. 34(1)(d)). Furthermore, the *self-harm* label suggests possible risks to *physical and mental well-being* (Art. 34(1)(d)) through the potential normalization of self-injury.

Sexual content labels such as *sexual*, *porn* could indicate risks related to the *protection of minors* and *gender-based violence* (Art. 34(1)(d)), particularly where content might involve exploitative or non-consensual imagery. While a granular post-level analysis to map specific content to DSA risk categories is outside the scope of the current study, our characterization of labeled content broadly hints at the existence of these potential systemic harms.

7 Ethical considerations

Our research is grounded in a commitment to ethical and responsible data analysis. The study exclusively uses publicly available data, analyzing social media posts that include text and images, and does not involve any non-public accounts related data, or direct user interactions. As such, this work does not constitute human-subjects research. Moreover, to protect the integrity and security of the data, all storage and processing were conducted on secure institutional infrastructure. Furthermore, our analytical approach was designed to minimize risk; we used in-house, open-source tools for our analyses. We did not interact with platform users, deploy automated agents, or manipulate platform behavior in any way. All results are presented using aggregated statistics. This careful methodology ensures that the positive impacts of the work substantially exceed any associated risks.

8 Concluding Discussion

By leveraging the architectural transparency of the Bluesky decentralized platform, our work conducts the first systematic audit of a live moderation system (Bluesky Moderation Service), examining its mechanism, efficacy, and purpose. Our findings paint a detailed portrait of a human-AI collaborative system grappling with the classic trade-off between precision and recall—one that acts swiftly on clear-cut violations but relies on slower, nuanced human judgment for more complex harms. Finally, through unsupervised clustering, we reverse-engineered the de facto meaning of abstract policy labels, finding that their application in practice is often much broader than their written definitions suggest. This characterization also provides a detailed account of the types of harm detected by the Bluesky Moderation Service. We contend that this study is itself evidence of the value of openness: the audit was only possible due to the architectural openness of the platform, which renders both the input data stream and moderation actions auditable, and our findings are in turn validated by the platform’s own transparency reporting.

Limitations and future work: This study has several limitations that open avenues for future research. Our analysis was confined to Bluesky’s official moderation service; a crucial next step is to extend this audit to the broader federated ecosystem of third-party labelers to understand how they complement or contradict the default system. Furthermore, some of the methodologies—e.g., reliance on labeling delay for understanding human-AI collaboration—are proxy-dependent. Post-hoc data collection, particularly of media content, may also introduce some data loss: posts that are taken down between the moderation action and our collection window are not captured, which could skew estimates of the relative prevalence of harms. While our findings corroborate Bluesky’s official transparency report, future work can delve into developing more robust methodologies. Finally, our work focused on the application of labels, but not their downstream effects on user behavior or content prevalence. Longitudinal studies are needed to measure these impacts and any temporal variations in moderation actions.

Takeaways for future moderation systems: Even with

these limitations, our findings offer meaningful guidance to the development of better content moderation systems. First, our cluster-based analyses provide a data-driven roadmap to improve the current labeling paradigm through increased label granularity. Second, although identifying unsafe content is a challenge all digital platforms grapple with, they largely do so independently. Moreover, even with regulatory transparency requirements such as those under the DSA, moderation practices remain opaque: the most common reason reported in the DSA Transparency Database is the uninformative “Other violation of provider’s terms and services.” To this end, the research community could draw inspiration from the collaborative efforts in spam analytics to develop joint initiatives for moderation analytics by adopting open standards like those on Bluesky. We believe our study is an initial step in this direction.

Acknowledgments

Ingmar Weber is supported by funding from the Alexander von Humboldt Foundation and its founder, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung). Abhijnan Chakraborty acknowledges the funding support from the Max Planck Society.

References

- Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *ACM SIGKDD*.
- Arshat, A.; and Etcovitch, D. 2018. The human cost of online content moderation. *Harvard JOLT*.
- Arthur, C. 2012. Facebook’s nudity and violence guidelines are laid bare. <https://bit.ly/3ReefVh>.
- Balduf, L.; Sokoto, S.; Ascigil, O.; Tyson, G.; Scheuermann, B.; Korczynski, M.; Castro, I.; and Król, M. 2024. Looking AT the Blue Skies of Bluesky. *ACM IMC*.
- Balduf, L.; Sokoto, S.; Baronchelli, A.; Castro, I.; Król, M.; Tyson, G.; Pavlou, G.; Scheuermann, B.; and Ascigil, O. 2025. Bootstrapping Social Networks: Lessons from Bluesky Starter Packs. In *AAAI ICWSM*.
- Bluesky. 2025. Terms of Service. <https://bsky.social/about/support/tos>.
- Bluesky. 2026. Bluesky 2025 Transparency Report. <https://bit.ly/4nwAUZ5>.
- Bono, C. A.; La Cava, L.; Luceri, L.; and Pierri, F. 2024. An exploration of decentralized moderation on Mastodon. In *ACM WebSci*.
- Campello, R. J.; Moulavi, D.; and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In *PAKDD*.
- Cochran, W. 1977. *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780471162407.
- Comission, E. 2025. DSA Transparency Database — transparency.dsa.ec.europa.eu. <https://transparency.dsa.ec.europa.eu/>. [Accessed 27-11-2025].
- Failla, A.; and Rossetti, G. 2024. “I’m in the Bluesky Tonight”: Insights from a year worth of social data. *PLOS ONE*, 19(11): e0310330.
- Ghali, M.-K.; Farrag, A.; Lam, S.; and Won, D. 2025. BEYONDWORDS is All You Need: Agentic Generative AI based Social Media Themes Extractor. *arXiv:2503.01880*.
- Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Halevy, A.; Canton-Ferrer, C.; Ma, H.; Ozertem, U.; Pantel, P.; Saeidi, M.; Silvestri, F.; and Stoyanov, V. 2022. Preserving integrity in online social networks. *CACM*.
- Hartmann, D.; Oueslati, A.; and Stauffer, D. 2024. Watching the Watchers: a Comparative Fairness Audit of Cloud-based Content Moderation Services. *ArXiv*, abs/2406.14154.
- Hartmann, D.; Oueslati, A.; Stauffer, D.; Pohlmann, L.; Munzert, S.; and Heuer, H. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. *ACM CHI*.
- Juneja, P.; Rama Subramanian, D.; and Mitra, T. 2020. Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices. *Proc. ACM Hum.-Comput. Interact.*, 4(GROUP).
- Kleppmann, M.; Frazee, P.; Gold, J.; Graber, J.; Holmgren, D.; Ivy, D.; Johnson, J.; Newbold, B.; and Volpert, J. 2024. Bluesky and the at protocol: Usable decentralized social media. In *DIN Workshop (ACM Conext)*.
- Krippendorff, K. 2011. Computing Krippendorff’s Alpha-Reliability. Technical report, University of Pennsylvania.
- Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*.
- Li, M.; Zhang, Y.; Long, D.; Chen, K.; Song, S.; Bai, S.; Yang, Z.; Xie, P.; Yang, A.; Liu, D.; et al. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*.
- Ma, R.; and Kou, Y. 2022. “I’m not sure what difference is between their content and mine, other than the person itself”: A Study of Fairness Perception of Content Moderation on YouTube. *PACM HCI (CSCW2)*, 6.
- Majumdar, A.; Paul, S.; Singh, P.; Abdelaziz, I.; Jarollahi, S.; Lee, S.; Gummadi, K. P.; Weber, I.; and Dash, A. 2026. Can Foundation Models Moderate Online Content? Evaluating Instruction- vs. Example-Driven Policy Operationalization. *arXiv:2609.10410*.
- McInnes, L.; Healy, J.; and Melville, J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*.
- Moulavi, D.; Jaskowiak, P. A.; Campello, R. J.; Zimek, A.; and Sander, J. 2014. Density-based clustering validation. In *SIAM SDM*.
- Nogara, G.; Sahneh, E. S.; DeVerna, M. R.; Liu, N.; Luceri, L.; Menczer, F.; Pierri, F.; and Giordano, S. 2025. A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch. *arXiv:2505.02317*.
- Qiwei, L.; Zhang, S.; Kasper, A. T.; Ashkinaze, J.; Eaton, A. A.; Schoenebeck, S.; and Gilbert, E. 2024. Reporting

Non-Consensual Intimate Media: An Audit Study of Deep-fakes. arXiv:2409.12138.

Raman, A.; Joglekar, S.; Cristofaro, E. D.; Sastry, N.; and Tyson, G. 2019. Challenges in the decentralised web: The mastodon case. In *ACM IMC*.

Release, E. P. 2022. Digital Services Act: Council and European Parliament provisional agreement for making the internet a safer space for European citizens. <https://bit.ly/4nxWb4B>.

Roberts, S. T. 2016. Commercial Content Moderation: Digital Laborers’ Dirty Work. In *The Intersectional Internet: Race, Sex, Class, and Culture Online*, 147–160. New Haven, CT: Peter Lang.

Roberts, S. T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Rodericks, A. 2025. Bluesky 2024 Moderation Report. <https://bit.ly/4nx82jj>.

Shahi, G. K.; Tessa, B.; Trujillo, A.; and Cresci, S. 2025. A Year of the DSA Transparency Database: What it (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election. *arXiv preprint arXiv:2504.06976*.

Spence, R.; Bifulco, A.; Bradbury, P.; Martellozzo, E.; and DeMarco, J. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*.

Suzor, N. P.; West, S. M.; Quodling, A.; and York, J. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13: 18.

Sweney, M. 2008. Mums furious as Facebook removes breastfeeding photos. <https://bit.ly/43885Zs>.

The Hive. 2026. Visual Moderation — Overview. <https://docs.thehive.ai/docs/visual-content-moderation>.

Trujillo, A.; Fagni, T.; and Cresci, S. 2025. The DSA Transparency Database: Auditing self-reported moderation actions by social media. *PACM HCI (CSCW)*.

Zhang, Z.; Zhao, J.; Wang, G.; Johnston, S.-K.; Chalhoub, G.; Ross, T.; Liu, D.; Tinsman, C.; Zhao, R.; Van Kleek, M.; et al. 2024. Trouble in Paradise? Understanding Mastodon Admin’s Motivations, Experiences, and Challenges Running Decentralised Social Media. *PACM HCI (CSCW2)*, 8.

A Bluesky Moderation Service: Statistics

Table 5 gives statistics about labeled records collected and their distribution across label values, post embed types and languages. Table 6 shows the top labels applied to accounts and profiles by BMS. For account labels, *!takedown* is the most frequently applied label, whereas *porn* is the most frequent label on profiles. When a label is applied to an account, it will show warnings on all of its content. On the other hand, a label on a profile is usually used to blur its avatar without showing any warnings on its posts.

B Additional Experimental Details for RQ2: Firehose Collection and Indexing

Firehose Data Collection: Using the `com.atproto.sync.subscribeRepos` endpoint, we collected firehose events between March and December 2025, yielding 1.14B post records (events of type `app.bsky.feed.post`), of which 11.9M (~1%) posts were labeled by the BMS. Each post record contains fields such as the post text, creation timestamp, and optional embedded media content identifiers (CIDs). From this collection, we derive two datasets used in our experiments. For the recall evaluation (Section 5), we randomly sample 1,000 posts. For experiment in (Section 5.2), we curate a subset of 40M posts drawn randomly from the March–June 2025 window (383.2M posts total), restricted to text and single-image posts.

Embedding, Indexing and Nearest-Neighbour Retrieval. For each post in the 40M subset, we compute multimodal embeddings using `Qwen3-VL-Embedding-2B` (Li et al. 2026)¹¹. Each post is encoded as a multimodal instruction, combining available text and a single image into a 2048 dimensional vector. All embeddings are normalised and inserted sequentially into a `faiss.IndexFlatIP` index of same dimension. Embeddings for the 336 labeled posts (described in Section 5.2) are searched against the firehose index using `k=1000` candidates per query. For each labeled post, we walk the ranked neighbor list and select the highest-ranked neighbor that (a) carries no BMS label and (b) contains exactly one image, yielding one semantically similar but unmoderated post per labeled query i.e set of 336 unlabeled posts used in Section 5.2.

C Automod and Hive AI Moderation

Overview. BMS’s automated content moderation is powered by two components operating in sequence: *Hive AI*, a commercial multi-head vision classifier, and *Automod*, BMS’s open-source rule engine that converts Hive’s raw scores into actionable labels. Figure 5 illustrates this pipeline.

Hive AI. When a post containing an image (individual frames in case of video) is submitted, BMS submits it to the Hive API. Hive AI’s visual moderation API returns a flat list of 128 class-score pairs organised into 54 model heads spanning five content domains: sexual content (26 heads, 59 classes), violence and gore (10 heads, 29 classes), drugs and vices (6 heads, 15 classes), hate imagery (5 heads, 10 classes), and miscellaneous image attributes (7 heads, 15 classes), as documented in the Hive Visual Moderation API (The Hive 2026). Within each head, classes are mutually exclusive, and their scores sum to 1. For example, the *Sexual Activity* head returns `yes_sexual_activity` and `no_sexual_activity`; the *Blood* head returns `very_bloody`, `a_little_bloody`, `other_blood`, and `no_blood`. Table 7 lists the heads and classes relevant to BMS’s five automated labels. Table 8 lists a few heads outside Automod’s rules.

¹¹<https://huggingface.co/Qwen/Qwen3-VL-Embedding-2B>

Table 5: Label statistics: available records, embed-type distribution, root post vs. reply breakdown, and top-3 languages per label. Here, thumb=external link thumbnail, record=quoted post, record+media=quoted post with media. Language tags normalised to BCP-47 base; posts with no tag counted as *no-lang*.

Label	Records	Embed type						Root / Reply		Top langs
		thumb	image(s)	record	record+media	video	text-only	root	reply	
porn	7,360,828	415,280	4,907,245	171	100,400	1,937,717	15	6,704,119 (91%)	656,709 (9%)	en (69%), no-lang (11%), es (4%)
sexual	2,243,593	170,884	1,632,310	33	41,808	398,541	17	2,061,542 (92%)	182,051 (8%)	en (65%), no-lang (14%), ja (4%)
nudity	271,246	17,515	226,590	21	5,445	21,669	6	248,942 (92%)	22,304 (8%)	en (68%), no-lang (12%), es (4%)
rude	219,712	4,007	8,860	5,700	627	284	200,234	8,502 (4%)	211,210 (96%)	en (97%), fr (1%), de (1%)
!takedown	214,079	1,409	8,931	1,946	420	1,691	199,682	195,252 (91%)	18,827 (9%)	no-lang (84%), en (14%), pt (0%)
sexual-figurative	134,942	961	121,184	23	4,861	7,904	9	126,611 (94%)	8,331 (6%)	en (81%), ja (7%), no-lang (4%)
graphic-media	93,788	28,229	39,179	8	4,117	22,252	3	72,853 (78%)	20,935 (22%)	en (64%), no-lang (11%), es (7%)
spam	78,077	24,143	4,775	3,543	113	212	45,291	711 (1%)	77,366 (99%)	en (71%), ja (5%), pt (4%)
intolerant	38,021	997	3,723	1,988	180	185	30,948	10,550 (28%)	27,471 (72%)	en (90%), ja (2%), es (1%)
threat	11,640	296	677	682	58	44	9,883	2,408 (21%)	9,232 (79%)	en (93%), fr (1%), pt (1%)
self-harm	10,187	388	8,108	7	373	1,088	223	8,944 (88%)	1,243 (12%)	en (49%), pt (39%), ja (3%)
!hide	3,520	298	483	236	32	11	2,460	1,520 (43%)	2,000 (57%)	en (92%), no-lang (5%), es (2%)
!warn	2,191	68	526	184	22	6	1,385	1,090 (50%)	1,101 (50%)	en (96%), de (1%), no-lang (0%)
Total	10,681,824	664,475	6,962,591	14,542	158,456	2,391,604	490,156	9,443,044 (88%)	1,238,780 (12%)	

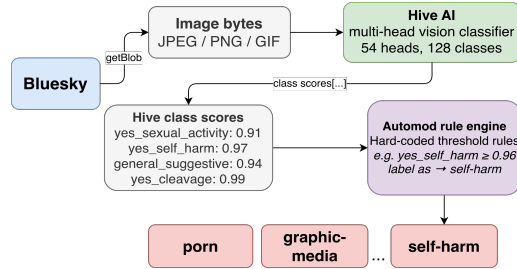


Figure 5: Pipeline of Automod. A post image is submitted to the Hive AI classifier, which returns scores across 128 classes and 54 heads. Automod applies hard-coded threshold rules over 16 of these classes to produce BMS labels.

Automod rule engine. Automod applies hard-coded threshold rules over Hive’s output scores across the different classes to produce BMS labels. Each rule checks a specific Hive class against a fixed threshold: for example, $\text{yes_self_harm} \geq 0.96$ triggers *self-harm*, and $\text{yes_sexual_activity} \geq 0.90$ triggers *porn*. Sexual content labels follow a priority cascade (*porn* before *sexual* before *nudity*). In total, Automod acts on 16 classes across 10 heads out of 128 classes and 54 heads total; the remaining 112 classes across 44 heads are not used. This design gives rise to both recall failure modes identified in Section 5.2.

Table 6: Top labels applied on accounts and profiles by the BMS in 2025.

Account Labels	Count	Profile Labels	Count
!takedown	2,040,503	porn	63,969
needs-review	213,927	nudity	13,859
spam	113,576	sexual	3,827
!hide	43,417	sexual-figurative	1,398
impersonation	3,185	graphic-media	859
rude	1,179	!takedown	347
intolerant	828	self-harm	295
!warn	656	spam	28

Table 7: Hive classes read by Automod rules.

Hive class	Description
yes_sexual_activity	Sex acts or genital stimulation
yes_realistic_nsfw	Live or photo-realistic nudity/sex acts
general_nsfw	Genitalia, sexual activity, nudity, buttocks, sex toys
yes_sexual_intent	Occluded, blurred, or hidden sexual activity
yes_sex_toy	Dildos or certain lingerie
yes_male_nudity	Male genitalia
yes_female_nudity	Breasts or female genitalia
yes_undressed	Nude even if genitals occluded by pose or overlay
yes_male_underwear	Male underwear
yes_female_underwear	Female underwear
yes_self_harm	Self-cutting, burning, or suicide methods
very_bloody	Gore or visible bleeding
human_corpse	Human dead body present
hanging	Human hanging by noose

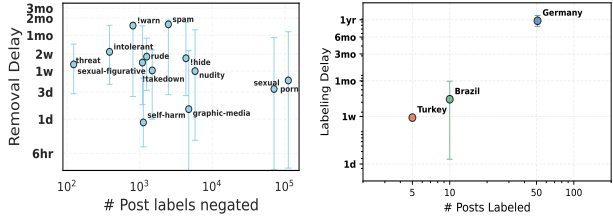
Table 8: Hive classes outside Automod’s rule set — H2 blindspot heads.

Hive class	Description
general_suggestive	Shirtless men, underwear/swimwear, suggestive poses without genitalia
yes_cleavage	Identifiable female cleavage
yes_male_shirtless	Shirtless below mid-chest
yes_bulge	Penis visible underneath clothing
yes_female_swimwear	Bikinis, one-pieces
yes_bodysuit	Bodysuits not covering the thigh

D Human Oversight in Redressal and Country Specific Moderation

D.1 Country-specific Moderation Services:

As shown in Figure 6b, country-specific moderation services operate on a fundamentally different timescale and scale compared to the BMS. The labeling delays (*!hide* label) are often exceptionally high, ranging from days to weeks (Turkey, Brazil) to nearly a year (Germany) for a few posts. This contrast from the default moderation service suggests that country-specific services do not function as a proactive safety system but rather as a reactive compliance mechanism.



(a) Redressal delays (b) Delay for country specific labeler

Figure 6: (a) Redressal Latency: The median delay between post creation and the negation (removal) of a label. (b) Labeling delays for country-specific moderation (Germany, Brazil, Turkey).

Table 9: Optimal hyperparameters for clustering each moderation label and the corresponding DBCV scores.

Label	<i>n_neighbors</i>	<i>min_cluster_size</i>	<i>min_samples</i>	DBCV	Total Posts	Noise Posts	Noise %
sexual	37	4179	540	0.626	300,000	88,783	29.6%
porn	32	4815	85	0.407	300,000	159,110	53.0%
nudity	38	1106	190	0.371	268,376	122,229	45.5%
sexual-figurative	35	1065	48	0.390	134,855	71,345	52.9%
graphic-media	43	1164	160	0.751	93,313	17,228	18.5%
intolerant	21	299	27	0.696	36,023	8,233	22.9%
rude	27	2706	46	0.387	214,001	105,267	49.2%
self-harm	16	139	23	0.566	10,169	3,557	35.0%
threat	16	222	24	0.693	10,958	2,246	20.5%
spam	38	1095	25	0.541	74,521	18,922	25.4%
!warn	29	107	1	0.910	2,006	0	0.0%
!hide	25	50	4	0.764	3,278	54	1.6%
!takedown	25	674	3	0.885	31,129	860	2.8%

D.2 Negation labels & Redressal mechanism:

BMS also often negates or reverses its labels. A negation label functions as an explicit override, i.e., it cancels or reverses a previously assigned label. We conduct a similar analysis on negation labels across 13 label categories, covering 206,724 negated post labels in total. The median delay between post creation and label negation varies substantially across label types, as shown in Figure 6a. Action labels (*!warn*, *!hide*) and labels such as *spam*, *intolerant* and *rude* show the longest removal delays, with median delays of 44–47 days for *spam* and *!warn*, and 12–15 days for *rude* and *intolerant*. In contrast, content labels such as *self-harm* (21h), *graphic-media* (1.5d), and *sexual* (3.4d) are negated considerably faster. Most of the negated labels are potentially automated, e.g., *porn*, *sexual*, *nudity*, etc., indicating these labels might have been automatically (but erroneously) applied to the posts and the authors of the posts may have reported such mislabeling through Bluesky’s redressal mechanism.

E Clustering labeled posts

The goal of this analysis is to obtain robust clusters for each moderation label as mentioned in Section 6.

E.1 UMAP dimensions and HDBSCAN clustering:

We performed hyperparameter tuning for HDBSCAN clustering to ensure coherent and meaningful groupings of posts. High-dimensional multimodal embeddings (2048-d) were obtained using Qwen3-VL-Embedding-2B (Li et al. 2026). This model allows encoding text, multiple images,

Table 10: Prompt used for the chunk-level summarisation with Qwen3-VL-32B-Instruct (map stage).

User Prompt:

Summarize these social media posts in 2–3 sentences. Focus on shared themes, sentiment, and any important visual context.

```
[POST_START 1]
<text of post 1>
<image(s) of post 1, if any>
<video frames of post 1, if any>
[POST_END]
[POST_START 2]
<text of post 2>
<image(s) of post 2, if any>
<video frames of post 2, if any>
[POST_END]
⋮
```

Table 11: Prompt used for the reduce stage of map-reduce summarisation (combining chunk-level summaries into a final cluster description).

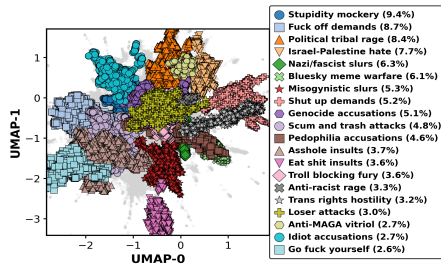
User Prompt:

Combine these summaries into ONE final output of 1–2 sentences. Capture overall themes, sentiment, and major visual trends.

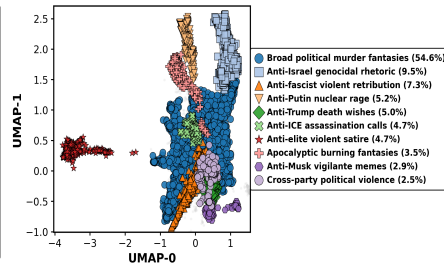
```
<chunk summary 1>
<chunk summary 2>
⋮
```

and video into a single joint embedding vector. The embedding model follows the Qwen3-VL chat template, in which the multimodal instance—text, image(s), and video in any combination is passed as a single user turn, and the final hidden state at the sequence’s last token (via last-token pooling) is taken as the fixed-length (2048-d) representation of the entire post. These embeddings were then projected into a low-dimensional space using UMAP. We explored a range of *n_neighbors* and *n_components*, using Trustworthiness score to assess the preservation of local structure. Based on these evaluations, *n_components* = 5 was selected as a reasonable trade-off for dimensionality reduction, and this setting was used for subsequent HDBSCAN clustering. Since hyperparameter optimisation requires running UMAP and HDBSCAN across dozens to hundreds of configurations, processing the full corpus at each trial would be both computationally prohibitive and memory intensive. Therefore, when a label corpus exceeded 300,000 posts, a uniform random subsample of that size was drawn.

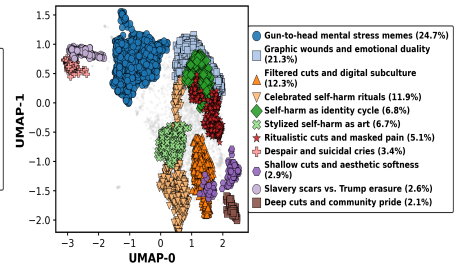
Clusters were obtained from the UMAP embeddings using HDBSCAN (Campello, Moulavi, and Sander 2013), which additionally produces a soft membership probability for each post reflecting its confidence of belonging to its assigned cluster. Rather than exhaustive grid search, combinations of UMAP *n_neighbors*, HDBSCAN *min_cluster_size*, and *min_samples* were



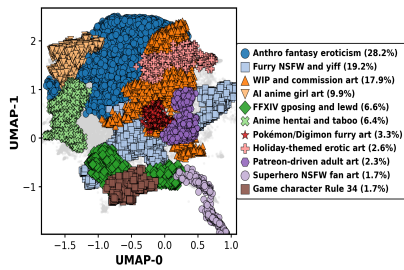
(a) Rude



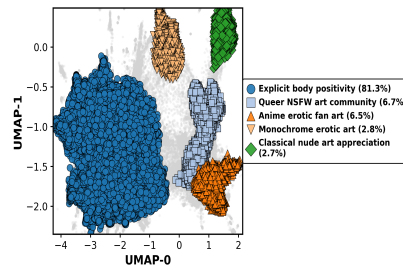
(b) Threat



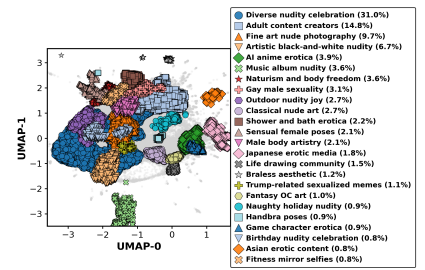
(c) Self-harm



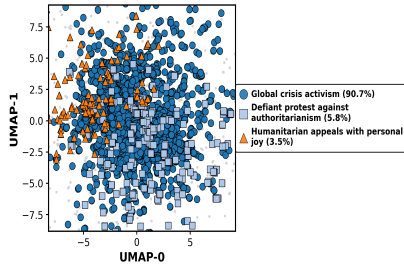
(d) Sexual-figurative



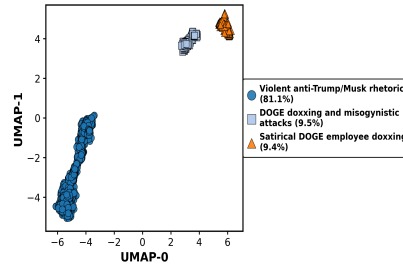
(e) Sexual



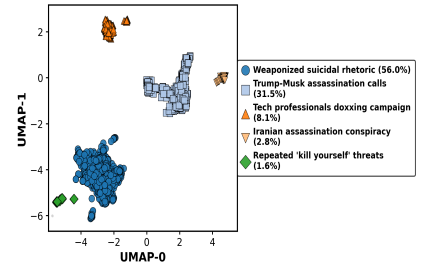
(f) Nudity



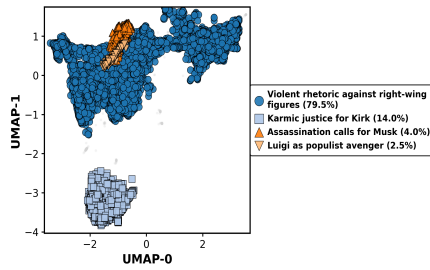
(g) Spam



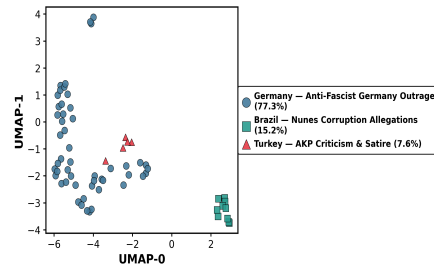
(h) !warn



(i) !hide



(j) !takedown



(k) !hide (country-specific)

Figure 7: Visualizations of clusters for different labels (first two components of UMAP)

Table 12: Prompt used for cluster naming with Qwen3-VL-32B-Instruct. All cluster summaries for a label are presented jointly so names reflect within-label distinctions.

System Prompt:
You are an expert content-taxonomy analyst. You will receive a list of cluster summaries — all clusters belong to the same parent label. Your task is to assign each cluster a short, contrastive name (2–5 words) that:
1. Clearly signals what makes that cluster different from the others.
2. Stays specific and avoids generic phrases like “related content” or “posts about”.
3. Is understandable by a non-specialist moderator.
Respond only with a JSON object mapping cluster_id (integer key as string) to cluster_name (string). No explanation, no markdown fences, no extra keys. Example format: {"0": "Racial slurs in sport", "1": "Anti-immigrant rhetoric", ...}
User Prompt:
Parent label: <label>
Below are the cluster summaries (one per cluster):
— CLUSTER 0 (n=<n>)—
<cluster 0 summary>
— CLUSTER 1 (n=<n>)—
<cluster 1 summary>
:
:
Now generate a short, contrastive name for every cluster listed above. Return only the JSON object described in the system prompt.

evaluated via Bayesian optimisation using Optuna with a Tree-structured Parzen Estimator (TPE) sampler (Akiba et al. 2019). Search ranges were adapted to corpus size. Each configuration was assessed using the Density-Based Clustering Validation (DBCV) score (Moulavi et al. 2014), which measures intra-cluster density relative to inter-cluster separation. The configuration with the highest DBCV score was selected for final clustering. Table 9 reports the selected hyperparameters, and DBCV scores for all labels.

While quantitative metrics such as DBCV provide an intrinsic measure of cluster coherence, we also obtained qualitative descriptions and names for each cluster using the LLM-based pipeline described in the following section.

E.2 Interpreting clusters

To characterise the thematic content of each cluster, we employ a two-stage automated pipeline: we first draw a sample from each cluster, then pass the sampled posts to a vision-language model to generate a concise thematic description. Later, we also validate our pipeline with human evaluation on a subset of clusters.

Sampling for LLM summarisation: Since clusters vary substantially in size — from hundreds to tens of thousands of posts — passing all posts to the model is neither feasible (Ghali et al. 2025) nor necessary: large clusters exceed any model’s context window, and the semantic redun-

dancy inherent in dense clusters means a statistically sufficient sample captures the same thematic signal at a fraction of the inference cost. For each cluster, a statistically justified sample was drawn using the Cochran (Cochran 1977) finite-population-corrected formula:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}, \quad n_0 = \frac{z^2 p (1 - p)}{e^2} \quad (1)$$

where n is the sample size, $z = 1.96$ is the Z-score corresponding to a 95% confidence level, $p = 0.5$ is the estimated proportion of the population (set to 0.5 for maximum sample size), and $e = 0.05$ is the margin of error. Posts were sampled without replacement, weighted by their HDBSCAN soft membership probability, ensuring that sampled posts span the full thematic range of each cluster rather than concentrating on its dense core.

LLM Summarisation. Sampled posts were passed to Qwen3-VL-32B-Instruct¹² using a map-reduce summarisation strategy, which addresses the context length limit of a single-pass LLM summarisation. Posts were batched into chunks of 32, each summarised independently into 2–3 sentences (map), and the chunk-level summaries were merged into a single 1–2 sentence cluster description in a second pass (reduce). Model inputs included post text and any attached media (images/video), enabling the model to reason over visual content alongside text. All inference used greedy decoding (`do_sample=False`) for deterministic, reproducible outputs, with `max_new_tokens=256` for both the map and reduce stages. The prompts used at each stage are shown in Tables 10 and 11.

Naming Clusters Cluster names were generated by presenting all cluster descriptions for a given label to the model in a single prompt, instructing it to assign short (2–5 word) names that distinguish each cluster from its siblings. This joint generation prevents label collapse — the tendency for independently named clusters to converge on near-identical generic phrases — and produces a taxonomy that is both specific and internally contrastive. Naming inference also used greedy decoding (`do_sample=False`) with `max_new_tokens=512`. The prompt used for contrastive naming is shown in Table 12. LLM-generated descriptions indicate that clusters generally capture dominant content themes per label. Some overlap between clusters exists, but the overall cluster assignments are coherent and meaningful, allowing us to characterise labeled content at a fine-grained level. We provide cluster names and their sizes in Figures 4 and 7.

Human Validation of Cluster Names. To assess the quality of the LLM-generated cluster names, we conducted a human validation study using a custom annotation interface built with Streamlit. For each cluster, annotators were presented with the cluster name, its LLM-generated summary, and up to 10 representative posts, including any attached images and video, sampled using scheme as the LLM summarisation stage. To protect annotators from harmful con-

¹²<https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct>

tent, images and video in sensitive label categories (e.g. graphic-media, self-harm) were avoided.

Annotators rated each cluster name on a five-point Likert scale reflecting the degree to which the name accurately and specifically describes the sampled posts: (5) *very appropriate* — the name precisely captures the cluster’s content; (4) *appropriate* — mostly accurate with minor gaps; (3) *neutral* — partially accurate; (2) *inappropriate* — significant misalignment; (1) *very inappropriate* — the name is misleading or irrelevant. For ratings of 3 or below, annotators were additionally asked to identify the primary issue (too broad, too narrow, misleading, unclear, or other) and optionally suggest an alternative name. Inter-annotator agreement was computed using Krippendorff’s α (Krippendorff 2011) over the ordinal Likert ratings.

Human Validation Results. Three annotators independently rated 31 clusters, yielding 93 ratings in total. Cluster names received a mean appropriateness rating of $\mu = 4.58$ ($SD = 0.63$, median = 5) out of 5. In total, 92.5% of individual ratings were at the *appropriate* or *very appropriate* level (≥ 4), and 90.3% of clusters had a mean rating ≥ 4 . No cluster received a rating of *inappropriate* or below (≤ 2), and the remaining 7.5% of ratings were *neutral* (3), indicating minor misalignment in a small number of cases. Inter-annotator agreement was assessed using Krippendorff’s α with an ordinal metric, yielding $\alpha = 0.686$, which falls within the *substantial* agreement range (Landis and Koch 1977). Of all pairwise rating comparisons across annotators, 95.7% differed by at most one point on the five-point scale, and 63.4% were in exact agreement. The four pairs differing by two points all involved a rating of 3 versus 5, representing cases where one annotator found the name neutral while another found it very appropriate. No pair differed by more than two points. Taken together, these results indicate that the automated pipeline: HDBSCAN clustering followed by LLM-based summarisation and naming, produces cluster names that are largely meaningful and faithful to the underlying post content as perceived by human reviewers.

F Action Labeled Contents and Spam

As discussed in Section 3.2, in addition to category labels, BMS also issues action labels such as *!hide*, *!warn* and *!take-down*, which enforce moderation decisions for all users and override any individual default settings. We also analyze these labeled posts to understand the content and behaviors that trigger such actions.

Takedown: Posts with the label *!takedown* are removed from Bluesky as they violate the platform’s terms of service. Since there is no precise, content-based definition for this label, we analyze the labeled posts to characterize the types of content and behaviors that are taken down by the platform. The clusters (Figure 7j) reveal content centered on extreme political hostility and violent celebration. Most of these posts focus on violent rhetoric and explicit calls for harm against right-wing figures and public personalities (clusters: Violent rhetoric against right-wing figures). Another cluster captures the widespread online glorification of the shooting of Charlie Kirk, marked by *schadenfreude*

and framing his death as karmic justice (cluster: Karmic justice for Kirk). One set of posts centers on *Luigi Manigione* as a symbolic populist avenger, reflecting a broader culture of rage against the elites (cluster: Luigi as populist avenger). Together, these clusters reveal that content taken down on Bluesky is predominantly characterized by celebrations of real-world violence and explicit calls for harm against named individuals.

Hide: When the *!hide* label is applied to a post, it is hidden from user feeds. The clusters for the label *!hide* capture content containing threats, harassment, or promotion of self-harm (Figure 7i in Appendix). Some clusters focus on violent threats against public figures, particularly Donald Trump and Elon Musk (cluster: Trump-Musk assassination calls), as well as conspiracy narratives surrounding alleged Iranian assassination plots against Trump (cluster: Iranian assassination conspiracy). Another cluster captures a coordinated doxxing campaign targeting young tech professionals involved with DOGE.

Warn: The *!warn* label is applied on a post to issue a warning on the content before it is shown. The clusters (in Figure 7h) reveal content centered on extreme political hostility and doxxing. One cluster (Violent anti-Trump/Musk rhetoric) captures violent anti-Trump and anti-Musk sentiment framing both as authoritarian threats, while other clusters contain posts about doxxing activity targeting DOGE employees through the circulation of leaked email lists.

Country-Specific Moderation Service (Hide): These clusters (Figure 7k in the appendix) reveal the contents flagged by country-specific labelers.

Germany: The labeled post clusters are associated with German political discourse and involve political criticism of Friedrich Merz, Sahra Wagenknecht, AfD, and Hubert Aiwanger by labeling them with vulgarities and accusing them of enabling far-right rhetoric.

Turkey: These *!hide* posts focus on criticism and satire of the AKP government and its perceived authoritarian control over state institutions, including commentary on the cancellation of opposition leader İmamoğlu’s diploma.

Brazil: The *!hide* posts are highly specific to the São Paulo municipal elections, focusing on allegations connecting the chief of staff of Mayor Ricardo Nunes and the Primeiro Comando da Capital (PCC), a major organized crime syndicate.

These observations lead to a number of interesting implications for building effective content moderation systems.

Spam: The *spam* label identifies unwanted content and behavior on Bluesky, ranging from bot-like activity and solicitation to coordinated inauthentic posting. However, most post records for this label were not retrieved and are effectively lost, significantly limiting the depth of analysis. The few available clusters reveal predominantly political and activist content, including urgent humanitarian appeals around ongoing geopolitical conflicts (cluster: Global crisis activism), organized resistance and protest against authoritarian regimes (cluster: Defiant protest against authoritarianism), and a blend of humanitarian advocacy with personal expression (cluster: Humanitarian appeals with personal joy).