PLURULE: A Benchmark for Moderating Pluralistic Communities on Social Media

Zoher Kachwala[1*], Bao Tran Truong[1,2], Rasika Muralidharan[1], Haewoon Kwak[1], Jisun An[1], Filippo Menczer[1]
[1]Observatory on Social Media, Indiana University, USA
[2]Center Synergy of Systems, TUD Dresden University of Technology, Germany

Abstract

Social media are shifting towards pluralism — community-governed platforms where groups define their own norms. What violates rules in one community may be perfectly acceptable in another. Can AI models help moderate such pluralistic communities? We formalize the task as a multiple-choice problem, mirroring how human moderators operate in the real world: given a comment and its surrounding context, identify which specific rule, if any, is violated. We introduce PLURULE, a multimodal, multilingual benchmark for detecting 13,371 rule violations across 1,989 Reddit communities spanning 2,885 rules in 9 languages. Using this benchmark, we show that state-of-the-art vision-language models struggle significantly: even GPT-5.2 with high reasoning performs only slightly better than a trivial baseline. We also find that bigger models and increased context provide marginal gains, and universal rules like civility and self-promotion are easier to detect. Our results show that moderation of pluralistic communities on social media is a fundamental challenge for language models. Our code[1] and benchmark[2] are publicly available.

1 Introduction

Ensuring platform safety and encouraging constructive participation are among the most persistent challenges of social media governance (Gillespie, 2020). On centralized platforms such as X (Transparency, 2025), YouTube (Google, 2025), and Meta (Meta, 2024), content moderation is increasingly carried out by a combination of human moderators and automated detection algorithms. These systems typically focus on narrowly defined categories — such as unlawful content, incivility, hate speech, and harassment — that are assumed to be

Input Prompt

Subreddit Info:

Subreddit: r/santamonica: Welcome to Santa Monica, California

Anything related to Santa Monica, where the rent is high but the temperatures aren't.

Rules:

Rule 1: Santa Monica-related only

Please keep posts to Santa Monica-specific issues. Venice and West LA issues can be posted in /r/LosAngeles

Rule 2: Respect other redditors

Respect other r/SantaMonica users, as both individuals and as groups. If you disagree with someone here, do so gracefully . . .

. . . (other rules). . .

Rule 6: No calling people paid shills just because they disagree with you.

Accusing people of being paid shills with no proof just because they disagree with you can and will result in bans.

. . . (other rules). . .

Submission:

Homeowners association sues City and Fairmont Miramar over hotel redevelopment approval
USER1, Tue, Dec 15, 2020, 1:37PM

[URL]

Discussion:

Comment 1: USER1, Tue, Dec 15, 2020, 1:38PM

And people wonder why everything is so expensive here. . .

. . . (other comments). . .

Comment 4 [TARGET COMMENT]: USER4, Mon, Dec 21, 2020, 3:58PM

You must work for Dell. He's been ripping off Santa Monica for years by avoiding taxes on the Fairmont Miramar. Look it up. He's a right wing a-hole

Question: Does the [TARGET COMMENT] violate a rule?

(a) Santa Monica-related only; (b) No rules broken; (c) Respect other redditors; (d) Appealing bans; (e) No calling people paid shills just because they disagree with you; (f) No low-effort posts; (g) No hateful speech directed at other people, including about the homeless; (h) No spam

Model Response

. . . (reasoning trace). . .

Final answer: (c) Respect other redditors. ✗

Figure 1: A PLURULE example. GPT-5.2 (high reasoning) receives the full context of the target comment, then selects which rule is violated. Here, the correct answer is (e) but GPT-5.2 selects (c). Full text in the Appendix.

*Correspondence: zkachwal@iu.edu, rasimura@iu.edu
[1]https://github.com/osome-iu/PluRule
[2]https://hf.co/datasets/osome-iu/PluRule

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34452–34471
July 2-7, 2026 ©2026 Association for Computational Linguistics

universal across all users and communities. By privileging mainstream norms, they overlook the values, languages, and forms of expression used by minority communities, leading to higher rates of content removal for marginalized groups (Lingel and Golub, 2015; Jiang et al., 2020; Griffin, 2024; Celeste et al., 2023). Centralized platform rules therefore fail to account for the diverse experiences and contextual meanings that vary across communities (Díaz and Hecht-Felella, 2021).

Bucking this trend, some platforms have adopted community-governed structures that allow groups to define their own norms. Reddit, for instance, hosts hundreds of thousands of topic-based communities (subreddits), each with its own rule set in addition to platform-wide guidelines (Reddit, 2025). While these pluralistic structures empower communities, they also place a substantial burden on volunteer moderators. On Reddit alone, the estimated value of this uncompensated labor exceeded $3.4 million in 2020 (Li et al., 2022b). Unsurprisingly, moderators are often eager to adopt automated tools that can reduce their burden (Robinson, 2025; Dosono and Semaan, 2019; Hill, 2019; Lloyd et al., 2025).

However, the contextual nature of community-specific rules poses a fundamental challenge for automation. What violates a rule in one community may be perfectly acceptable in another (Chandrasekharan et al., 2019; Li et al., 2022a). A satirical insult about someone's appearance, for instance, is encouraged in r/RoastMe but would violate civility rules in most other communities. Similarly, self-promotion that constitutes spam in most subreddits is required in creative showcase communities. Effective moderation requires understanding not just the rule text, but the implicit norms, values, and purposes that each community has developed over time.

Given these contextual complexities, the question arises whether modern AI systems can effectively assist with pluralistic moderation. The central challenge is whether language models can recognize that identical content may be acceptable in one community but violate rules in another. Even similar rules may be interpreted differently depending on local community norms (Selbst et al., 2019; Birhane et al., 2021).

To investigate this question empirically, we formalize the detection of rule violations as a multiple-choice task that mirrors how human moderators operate in practice (Figure 1). We introduce PLURULE, the first multimodal, multilingual benchmark for moderating pluralistic communities on social media. The benchmark comprises 13,371 moderation instances with 72,675 comments and 3,643 images, spanning 1,989 subreddits with 2,885 distinct rules across 9 languages. PLURULE incorporates substantial diversity along two dimensions: 25 semantically-derived subreddit categories (e.g., politics, gaming, music) and 27 rule categories (e.g., civility, self-promotion, spoilers).

Using PLURULE, we evaluate state-of-the-art vision-language models (VLMs) on the detection of rule violations under different context conditions. Our results reveal substantial limitations: even GPT-5.2 with high reasoning effort achieves only 58% accuracy, barely exceeding a trivial baseline that always predicts no violation (50%). Providing additional context — the discussion thread, original submission, participant labels, and images — improves GPT-5.2's performance by only 2–3 percentage points. Open-weight models like Qwen3-VL-Instruct and Qwen3-VL-Thinking perform even worse, failing to surpass baseline performance. Performance breakdown by rule category reveals that models successfully detect universal violations such as civility (69%) and self-promotion (63%), but fail on rules that require contextual understanding; low-effort (43%), evidence-based (47%), and relevance (44%) all fall below baseline. These results reveal a critical gap: current VLMs can enforce universal norms but cannot adapt to the diverse, context-dependent standards that define pluralistic moderation.

2 Related Work

Existing datasets for content moderation focus on narrow categories such as toxic speech (Hoang et al., 2024), hate speech (Nghiem and Daumé Iii, 2024), or misogyny (Sheppard et al., 2024). Automated systems trained on these datasets are limited to detecting broadly unacceptable content under singular global standards of appropriateness. This assumption breaks down in decentralized platforms, where different demographic groups significantly diverge about what is considered respectful, emotionally appropriate, or toxic (Sachdeva et al., 2022; Ali et al., 2025). Moderation on such platforms must account for pluralism. On Reddit, for example, rules extend beyond toxicity (Binns et al., 2017; Matias, 2019) to include locally defined norms around formatting, tone, and ideological or topical relevance (Chandrasekharan and Gilbert, 2019).

Even the enforcement of similar norms can vary widely across communities (Chandrasekharan et al., 2018). On Reddit in particular, moderators routinely interpret rules and assess the appropriateness of content relative to local community values rather than mechanically executing fixed policies (Li et al., 2022a; Fiesler et al., 2018; Matias, 2019). Consequently, within a community, multiple moderators can diverge when guidelines are broad or context-dependent (Binns et al., 2017; Chandrasekharan et al., 2019). Across communities, the same content may be acceptable in one context while violating norms in another — a distinction that models trained on aggregated data from multiple communities often fail to capture (Sap et al., 2022; Raji et al., 2020).

Previous work attempts to model the community-dependent nuance of moderation, but does not address its context-dependent nature. Chandrasekharan and Gilbert (2019) identify a small set of recurring "macro" norms shared across communities. Park et al. (2021) introduce a text-only dataset that collapses thousands of community-specific rules into coarse-grained types. This approach abstracts thousands of individual subreddit rules into a limited number of universal categories, obscuring differences that define each community. (He et al., 2024) provide models with individual rules for binary yes/no judgments.

PLURULE advances beyond prior work along three key dimensions. First, it explicitly models pluralism: instead of applying a fixed set of universal categories, models must reason over distinct, community-defined rules. Second, it frames moderation as a rule identification task (multiple-choice) rather than binary classification. This mirrors real-world moderator workflows and enables more fine-grained evaluation. Finally, PLURULE is multilingual and multimodal, capturing the visual (Gomez et al., 2020) and linguistic (Blodgett et al., 2016) diversity of online communities often overlooked by text-only benchmarks.

3 PLURULE Benchmark

PLURULE formalizes the task of moderating pluralistic communities on Reddit as a multiple-choice question (Figure 1). Given a comment from a specific community (subreddit), models must identify which specific rule, if any, has been violated.

For each comment, models receive the community's rules along with the surrounding context that moderators consider when making decisions. The context includes: (1) the discussion thread that precedes the comment; (2) the submission post to which the comment responds, including any images; and (3) anonymized identifiers of the participants in the discussion.

Each *moderation instance* in PLURULE consists of a pair: a violating comment and a compliant comment with overlapping context from the same submission. Models are evaluated on both comments separately. For both comments, models are presented with answer options consisting of all subreddit rules plus a "No rules broken" option, labeled (a), (b), (c), etc. Each comment's answer options are deterministically shuffled using a seed based on the comment ID to prevent models from exploiting positional bias. The correct answer for violating comments is the violated rule; for compliant comments, it is "No rules broken." Since half the comments violate a rule and half do not, always predicting "No rules broken" yields a majority baseline of 50% accuracy.

4 PLURULE Construction

We select Reddit as a platform because moderation actions are public: a moderation action occurs when a human moderator leaves a comment explaining a rule violation (e.g., "Your comment violates Rule 2"). We construct PLURULE by starting from such moderator comments in the Pushshift Reddit archives (Baumgartner et al., 2020) and transforming them into structured benchmark instances with verified rule labels, contrastive pairs, and semantic clustering. Below we describe the five-phase pipeline for this construction process.

4.1 Phase 1: Data Collection

We start from a publicly hosted, extended version of the Pushshift Reddit archives (Cohen and Lo, 2014), containing approximately 15 billion comments across 40 thousand subreddits. From these archives, we extract comments by moderators, flagged by a "distinguished" field in the comment object. To focus only on comment (not post) violations, we exclude top-level replies to submission posts. We filter out accounts with usernames that match bot-related keywords, e.g., "bot," "automod." This yields approximately 10 million moderator comments across 40 thousand subreddits.

Figure 2: PLURULE construction pipeline

We then query the Reddit API to collect each subreddit's full object. This helps to retrieve its current rules, infer its official language, and determine if it is NSFW (not-safe-for-work). Each subreddit must have at least one moderator comment and at least two explicit rules. We also exclude communities with adult content (NSFW). After filtering for these criteria, we obtain 17,468 subreddits with 131,400 rules and approximately 9 million moderator comments.

4.2 Phase 2: Rule Matching

Moderators often reference rules in their comments (e.g., "Rule 3 – No personal attacks"). Despite this, linking historical moderator comments to specific rule violations is a key challenge because rules evolve over time: new ones are added, old ones are deleted, and numbering and wording of existing rules can change. Since the Reddit API provides only present-day rules (as of November 2025), we match the full text of moderator comments to the full text of current rules.

We use Qwen3-Embedding-8B, a multilingual text embedding model, to encode all 9 million moderator comments and 131,400 rules as dense vectors. For each comment, we compute cosine similarity against all rules of the subreddit to which it belongs (7.5 rules on average), producing 90 million comment-rule scores.

We apply two thresholds to infer high-quality labels. First, a *match threshold* at the 99.2nd percentile of similarity scores (0.79): only comment-rule pairs above this threshold count as matches. Second, an *ambiguity threshold* at the 98th percentile (0.75): if multiple rules for a single comment exceed this threshold, we discard the comment entirely to avoid inferring ambiguous labels. The 9 million moderator comments yielded 174,412 ambiguous cases and 672,493 matched comments. In Phase 4, after additional filtering, we verify the quality of these matches.

4.3 Phase 3: Instance Construction

We wish to capture the complete conversational context of each violation, i.e., the full comment thread leading to the rule-violating comment — the one to which the moderator replied. We collect all comments from the same submission by matching submission IDs in the Pushshift archives, yielding 73.8 million comments. We then build comment trees representing the reply structure. From each tree, we extract a *violating thread*: the path up from the rule-violating comment, through its parent comments, to the root submission.

Effective moderation requires a capability to discriminate between similar rule-violating and compliant comments within the same discussion. To this end, we create a *moderation instance* by pairing each violating thread with a *compliant thread* — a discussion branch from the same submission that received no moderator action. We first collect candidate compliant threads whose leaf comment lies at either depth n (same as the violating thread) or depth $n-1$.

We then apply six filtering criteria. For both violating and compliant threads, we exclude: (1) deleted/removed content or deleted users to ensure complete discussion context; (2) media in comments to limit images to submissions only; and (3) any moderator-authored comments in the thread to avoid back and forth discussions between a moderator and a user. For violating threads specifically, we exclude (4) edited leaf comments that became compliant after moderator intervention. For compliant threads specifically, we exclude (5) leaf comments with moderator replies to ensure no moderator flagged these comments as violations. At the instance level, we further exclude (6) instances whose moderator comment was posted on or after March 1, 2023 to adhere to the Pushshift coverage window.

After filtering, we rank candidate compliant threads for each violating thread using three criteria to maximize the shared context between the two: (1) higher number of common ancestors; (2) higher thread depth to prioritize n over $n-1$; and (3) lower vote score, to select less popular content that nevertheless complied with community rules. We select the compliant thread that ranks highest.

To complete moderation instances with the con-

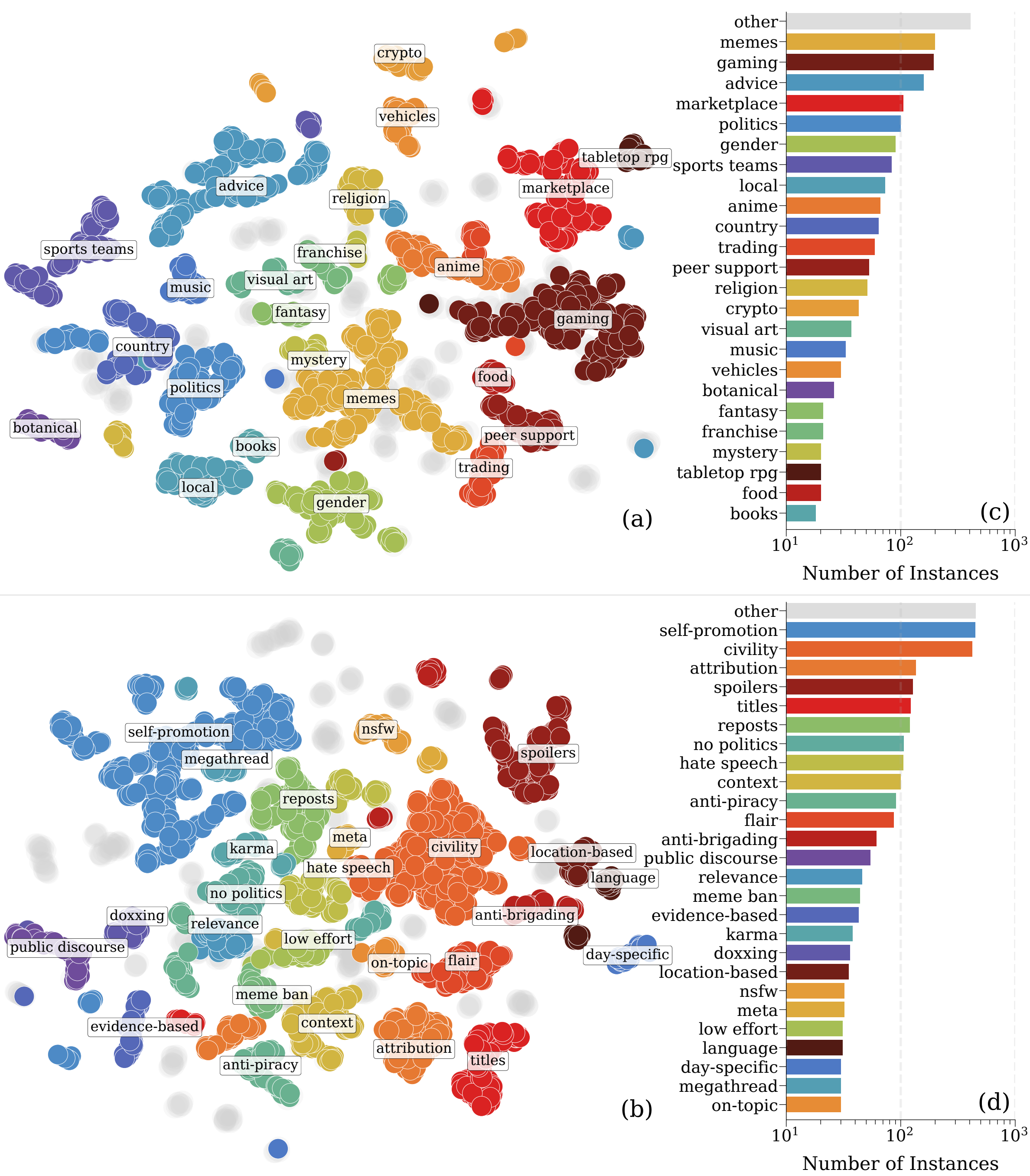

Figure 3: Diverse communities and rules in the PLURULE benchmark. Left: 2D UMAP visualizations of (a) 1,989 subreddits and (b) 2,885 rules, with colors indicating cluster assignments by HDBSCAN. Grey points represent unclustered items categorized as "other". Right: Distributions of 13,371 instances across (c) 25 subreddit clusters and (d) 27 rule clusters, with bar colors matching the clusters.

Split	Instances	Comments	Images	Subreddits / Clusters	Rules / Clusters	Languages
Train	9,155	51,968	2,077	861 / 25	1,336 / 27	9
Val	1,382	7,631	376	537 / 25	586 / 27	9
Test	2,834	13,076	1,190	1,989 / 25	2,039 / 27	9
Total	**13,371**	**72,675**	**3,643**	**1,989 / 25**	**2,885 / 27**	**9**

Table 1: PLURULE statistics. Each instance contains one rule-violating and one compliant thread from the same submission. All 1,989 subreddits appear in the test set. We count only the 9 languages with at least 10 instances each. There are 22 languages with at least one instance.

text from which the discussions originated, for each thread pair we collect the corresponding submission objects from Pushshift. We filter out instances whose submissions contain NSFW content, cross-posts, or videos. For the remaining submissions, we download images using a priority hierarchy: gallery images, direct URLs, video thumbnails, and Reddit-cached previews as fallbacks. Each download is validated for image content type and capped at 50 MB. We also exclude submissions with deleted/removed content, deleted users and those posted by moderators.

From 672,493 matched comments, we build 378,334 comment trees and successfully create 16,289 instances, totaling 32,578 threads. Most failures for threads and submissions stem from deleted/removed content or deleted users.

4.4 Phase 4: Verification

We use a large language model, Qwen3-30B-A3B-Instruct, to verify the rule matches inferred in Phase 2 for the benchmark instances. For each instance, we present the model with the *moderator comment* and *matched rule* to classify the comment as: (a) stating a violation of the rule, (b) discussing the rule, or (c) unrelated to the rule. We retain as ground-truth labels only matched rules classified as (a), achieving an 82.1% verification rate (13,371 out of 16,289 instances). This step filters out incorrect matches and cases where moderators mentioned a rule without enforcing it.

To evaluate the accuracy of these ground-truth labels, three authors independently annotated 100 moderator comments sampled randomly from English subreddits. For each moderator comment, annotators selected which subreddit rule was violated from the available options — the same task performed by the matching pipeline. For 85 of the comments in the sample, all three annotators agreed on the label. In 12 cases, the label was assigned based on a majority (two of the three annotators agreed). In the 3 remaining cases, there was no majority agreement and the label was adjudicated after further inspection by one annotator. Comparing the pipeline's labels against this human-established ground truth, we found 96% overall accuracy: 100% on full-agreement cases (85/85), 66.67% on majority-agreement cases (8/12), and 100% on adjudicated cases (3/3).

4.5 Phase 5: Data Splitting and Clustering

We split the instances into training, validation, and test sets using a strategy based on the number of instances per subreddit. For subreddits with a single instance, we allocate the instance to the test set. For subreddits with two instances, we allocate one instance to the training set and one to the test set. For subreddits with 3–9 instances, we allocate one each to test and validation sets, and the remaining to the training set. For subreddits with 10 or more instances, we use a 80/10/10 split for the training, validation, and test sets. This ensures all communities appear in the test set while preventing any single community from dominating the evaluation.

To analyze model accuracy across communities, we cluster subreddits and rules based on their semantic embeddings. For subreddits, we embed the subreddit name, title, and description. For rules, we embed the concatenation of short name, description, and violation reason. We apply UMAP for dimensionality reduction using cosine distance on the 4,096-dimensional Qwen3-Embedding-8B, then HDBSCAN for density-based clustering (see Appendix B). We visualize the resulting clusters in Figure 3.

We labeled each cluster using Qwen3-30B-A3B-Thinking (see Appendix B) followed by manual refinement. We assigned these cluster labels to all instances, enabling both fine-grained and category-level evaluation. Table 1 provides full statistics of the PLURULE dataset.

5 Evaluation

5.1 Experimental Setup

For each instance, models receive the subreddit description, complete rule set, and surrounding context, then select the correct answer from the multiple-choice options. We report accuracy on the test set, with the 50% baseline corresponding to always predicting "No rules broken." We compute 95% confidence intervals via bootstrap resampling with 100 thousand iterations.

We evaluate three open-weight Vision-Language Models from the Qwen3-VL family for their diversity in sizes (4B, 8B, and 30B) and OpenAI's flagship model GPT-5.2. For each Qwen model, we test both Instruct and Thinking variants. For GPT-5.2, we test with low and high reasoning effort. Qwen models use temperature 0 and seed 0 for reproducibility.

We use a two-stage evaluation pipeline. In

Models	Qwen3-VL-4B		Qwen3-VL-8B		Qwen3-VL-30B		GPT-5.2	
Variants	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Low	High
Comment Only	**49.6**	37.4	**51.0**	40.3	50.2	46.1	54.1	55.0
+Discussion	49.2 (-0.4)	39.8 (+2.4)	50.7 (-0.3)	43.9 (+3.6)	51.0 (+0.8)	48.2 (+2.1)	55.3 (+1.2)	56.2 (+1.2)
+Submission	48.3 (-0.9)	44.9 (+5.1)	49.2 (-1.5)	**47.2** (+3.3)	51.1 (+0.1)	49.1 (+0.9)	56.8 (+1.5)	57.3 (+1.1)
+User	48.9 (+0.6)	**45.0** (+0.1)	50.0 (+0.8)	46.7 (-0.5)	**52.4** (+1.3)	49.4 (+0.3)	**57.4** (+0.6)	**57.7** (+0.4)
+Images	48.4 (-0.5)	45.0 (+0.0)	49.8 (-0.2)	44.9 (-1.8)	52.3 (-0.1)	**49.5** (+0.1)	**57.4** (+0.0)	57.6 (-0.1)
Baseline	50.0							

Table 2: Accuracy (%) across models and context levels. Numbers in parentheses show differences from the previous row. Best-performing contexts for each model variant are highlighted in bold. 95% CI for all values do not exceed $\pm 1.3\%$. The baseline corresponds to always predicting "no rules broken."

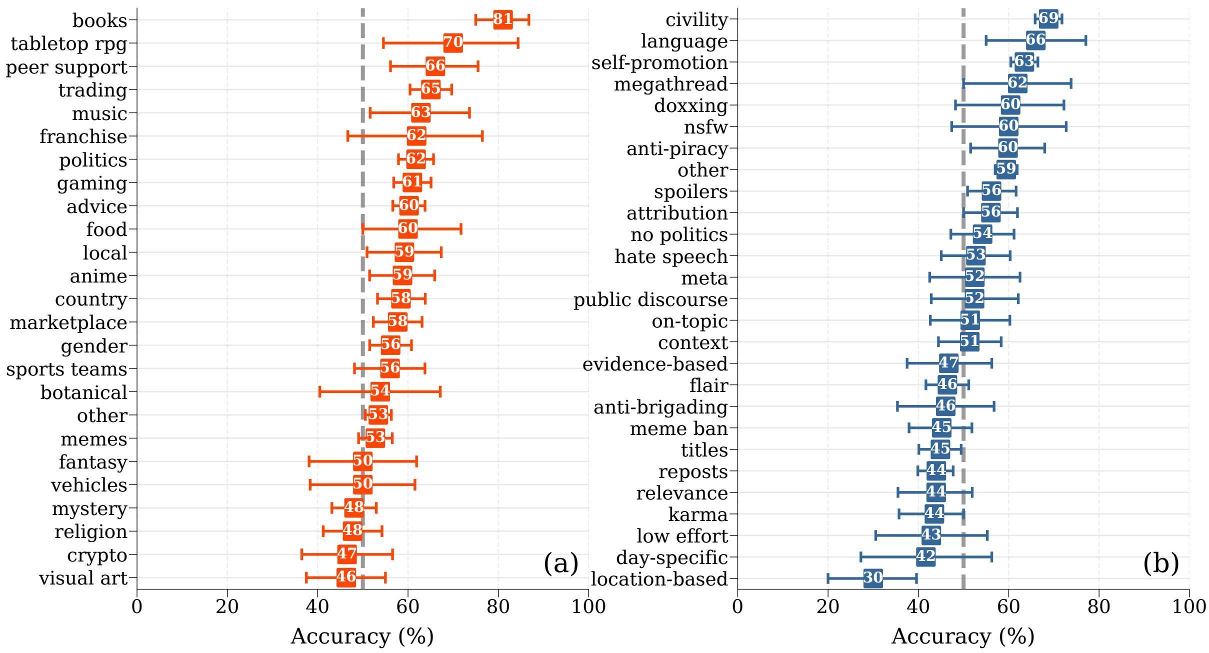

Figure 4: Accuracy for GPT-5.2 (high reasoning) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline. Results for other models are in Appendix C.

Stage 1, the model generates a free-form response to the input (see Fig. 1). In Stage 2, we append "Final Choice:" to prompt the model for its final answer, then extract the selected option (a–h) using a regular expression. For GPT-5.2, Stage 2 uses Qwen3-VL-30B-Instruct for answer extraction.

To understand which contextual signals aid rule violation detection, we test five cumulative context levels, where each level adds information to the previous: (1) **Comment Only** — the target comment; (2) **+ Discussion** — the full comment thread leading to the target comment; (3) **+ Submission** — title and body text of the original post that initiated the discussion; (4) **+ User** — anonymized author labels (USER1, USER2, etc.) to track participants; and (5) **+ Images** — media from the submission, when available. All levels include the subreddit description and complete rule set as baseline. We further analyze performance by subreddit cluster and rule cluster to identify which community types and rule categories pose the greatest challenges.

5.2 Results

Table 2 reports accuracy across model sizes, context configurations, and reasoning variants (see Appendix C for violating/compliant comment breakdowns.) GPT-5.2 substantially outperforms all Qwen variants, with high reasoning effort achieving 57.7% accuracy — nearly 8 points above the 50% baseline — while Qwen models barely exceed the baseline regardless of scale. If we were to weigh violating comments more heavily than compliant ones, the results are worse across all models (see Appendix C).

Focusing on GPT-5.2 with high reasoning, additional context provides limited signal for rule violation detection: using the full context improves performance by only 2.7 percentage points above comment-only, with discussion comments yielding the largest incremental gain (+1.2 points).

Extended reasoning does not necessarily help with this task: Qwen Thinking variants underperform their Instruct counterparts, while the differences between GPT-5.2 with low and high reasoning are not significant.

Figure 4 breaks down accuracy by subreddit and rule clusters for GPT-5.2 (high reasoning) using full context. See Appendix C for breakdowns by violating/compliant comments and languages. Performance varies substantially across community and rule clusters. For example, civility (69%), language (66%), and self-promotion (63%) rules are detected reliably, but low effort (43%), relevance (44%), and evidence-based (47%) perform worse than baseline. To help interpret the higher accuracy in some rule clusters, we examined the number of rules in these clusters and the number of subreddits that contain these rules. As reported in Appendix C, civility and self-promotion contain far more rules than any other cluster, and those rules appear in far more subreddits than rules from other clusters. This suggests that civility and self-promotion are universal rule types, and that models are better at detecting violations of such rules than rules requiring local context. We report similar results obtained with other models in Appendix C, with additional breakdowns by violating/compliant comments and languages.

6 Discussion

PLURULE provides a testbed for measuring progress toward models that can help moderate diverse community standards. Unlike prior datasets that use coarse-grained categories or single-rule binary classification, PLURULE requires models to distinguish among all of a community's rules simultaneously — mirroring the decision space faced by human moderators. We evaluate whether a single model can serve as an expert moderator across thousands of communities.

Our results reveal a gap: models succeed on universal violations like civility and self-promotion, but struggle with rules that vary across communities. The two-dimensional variability — 46–81% accuracy across community types and 30–69% across rule categories — suggests that VLMs may be internalizing universal standards from training.

Beyond content moderation, PLURULE evaluates whether AI systems can respect diverse human communities rather than imposing uniform standards. Fine-tuning on community-specific examples might help models learn local norms, though scalability is a challenge. Retrieval-augmented approaches that condition on historical moderation decisions offer another promising direction. The semantic clustering we provide enables analysis of transfer learning: can models trained on one community or rule type generalize to similar ones?

We release PLURULE in dehydrated form — only IDs and our derived labels — along with scripts to rehydrate the content from the Pushshift archives and rebuild the benchmark end-to-end.

7 Limitations

PLURULE was constructed from publicly available data where moderators left comments citing rule violations. Private moderator communications, removed content, and shadow-banned posts are not accessible, meaning communities that moderate silently are underrepresented. This likely biases the dataset toward less severe violations, as serious offenses are often removed without comment. English-language subreddits dominate due to Reddit's user demographics; findings may not generalize to platforms with different community structures or moderation practices.

Our pipeline matches historical moderator comments (2005–2023) to rule sets retrieved in November 2025. While semantic matching handles rule rewording and renumbering, it cannot account for rules that were added, removed, or fundamentally changed over time. Some matches may therefore be anachronistic.

Finally, certain violations require information unavailable in our dataset. Detecting ban evasion or repeat offenders requires historical user data that we do not collect for privacy reasons.

8 Ethical Considerations

In the PLURULE dataset, all usernames are anonymized by generic labels (USER1, USER2, etc.).

Models trained on PLURULE could potentially be misused to evade moderation by learning what content triggers enforcement. This risk is inherent to any dataset that captures moderation decisions. We believe the research benefits outweigh this risk, as understanding moderation patterns is essential for developing robust systems.

Acknowledgments

We gratefully acknowledge NVIDIA for the GPU access that made this study possible. Z.K. received support from the Institute for Humane Studies under grant nos.IHS018765. H.K. was supported by the Republic of Korea's MSIT (Ministry of Science and ICT), under the Global Research Support Program in the Digital Field Program (RS-2024-00425354) supervised by the IITP (Institute of Information and Communications Technology Planning & Evaluation).

References

Dalia Ali, Dora Zhao, Allison Koenecke, and Orestis Papakyriakopoulos. 2025. Operationalizing pluralistic values in large language model alignment reveals trade-offs in safety, inclusivity, and model behavior. *arXiv preprint arXiv:2511.14476*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *arXiv preprint*. ArXiv:2001.08435 [cs].

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics*, pages 405–415, Cham. Springer International Publishing.

Abeba Birhane, Vinay Uday Prabhu, and Evelyn Kahembwe. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 303–314. ACM.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Edoardo Celeste, Nicola Palladino, Dennis Redeker, and Kinfe Yilma. 2023. Platform policies versus human rights standards. In *The content governance dilemma: Digital constitutionalism, social media and the search for a global standard*, pages 93–129. Springer.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30.

Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit. *arXiv preprint arXiv:1904.03596*.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Joseph Paul Cohen and Henry Z. Lo. 2014. Academic Torrents: A Community-Maintained Distributed Repository. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, XSEDE '14, pages 1–2, New York, NY, USA. Association for Computing Machinery.

Ángel Díaz and Laura Hecht-Felella. 2021. Double standards in social media content moderation. Technical report, Brennan Center for Justice at New York University School of Law. Accessed: 24 Dec 2025.

Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Snowmass Village, CO, USA. IEEE.

Google. 2025. Youtube community guidelines enforcement.

Rachel Griffin. 2024. The heteronormative male gaze: Experiences of sexual content moderation among queer instagram users in berlin. *International journal of communications, network and system sciences*, 18:1266–1288.

Zihao He, Jonathan May, and Kristina Lerman. 2024. CPL-NoViD: Context-Aware Prompt-Based Learning for Norm Violation Detection in Online Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:569–582.

Benjamin Mako Hill. 2019. How Discord moderators build innovative solutions to problems of scale with the past as a guide.

Nhat M. Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. ToXCL: A Unified Framework for Toxic Speech Detection and Explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472, Mexico City, Mexico. Association for Computational Linguistics.

Jialun'Aaron' Jiang, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2020. Characterizing community guidelines on social media platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 287–291.

Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022a. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:584–595.

Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022b. Measuring the Monetary Value of Online Volunteer Work. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:596–606.

Jessa Lingel and Adam Golub. 2015. In face on facebook: Brooklyn's drag community and sociotechnical practices of online communication. *Journal of Computer-Mediated Communication*, 20(5):536–553.

Travis Lloyd, Joseph Reagle, and Mor Naaman. 2025. 'There Has To Be a Lot That We're Missing': Moderating AI-Generated Content on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 9(7):CSCW264:1–CSCW264:24.

J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2):2056305119836778. Publisher: SAGE Publications Ltd.

Meta. 2024. How enforcement technology works.

Huy Nghiem and Hal Daumé Iii. 2024. HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting Community Sensitive Norm Violations in Online Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Reddit. 2025. Reddiquette. `https://support.reddithelp.com/hc/en-us/articles/205926439-Reddiquette`. Accessed: 2025-12-23.

C. J. Robinson. 2025. X Is Using AI Fact-Checkers. https://www.cjr.org/analysis/x-twitter-ai-fact-checkers-community-notes-misinformation-bots.php.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pages 59–68. ACM.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2024. Biasly: An Expert-Annotated Dataset for Subtle Misogyny Detection and Mitigation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 427–452, Bangkok, Thailand. Association for Computational Linguistics.

X Transparency. 2025. Global transparency report.

A Full Example

Figure 5 expands the abridged datapoint in Figure 1 with all seven community rules, the complete discussion thread, and GPT-5.2's verbatim reasoning trace.

B Supplementary Methods

For clustering subreddits and rules, we perform grid searches over UMAP and HDBSCAN parameters: n_neighbors, n_components, min_cluster_size, and min_samples. We maximize DBCV (Density-Based Cluster Validity), which measures cluster separation and coherence.

For subreddits, optimal parameters yield DBCV = 0.448 with 24 clusters (66 items per cluster on average). We treat the remaining subreddits as the 'other' cluster (20.4% noise). For rules, optimal parameters yield DBCV = 0.569 with 26 clusters (94 items per cluster on average). We treat the remaining rules as the 'other' cluster (15.6% noise).

We assign semantic labels to each cluster using Qwen3-30B-A3B-Thinking. For each cluster, we prompt the model to generate 10 candidate labels and select the most common one via majority voting. We then manually verify and refine the labels for consistency. This produces 26 rule clusters (e.g., civility, self-promotion, spoilers, flair) and 24 subreddit clusters (e.g., politics, memes, gaming, trading).

C Extended Results

Recall Breakdowns. Table 3 reports recall for violating and compliant comments across all models and context levels.

Weighted Accuracy. If we were to weigh violating comments more heavily than compliant ones, the results are worse across all models. Table 4 reports a 2:1 weighted accuracy that prioritizes violation detection over compliance: every model's score is lower than in the symmetric Table 2 — e.g., GPT-5.2 (high reasoning) with full context drops from 57.6% to 52.6% — because models consistently recall compliant comments better than violating ones.

GPT-5.2 Analysis. Figure 6 shows recall by subreddit and rule clusters for GPT-5.2 (high reasoning) with full context.

Qwen3-VL Analysis. For each Qwen3-VL model (4B, 8B, 30B), we provide two figures: accuracy by subreddit and rule clusters (Figures 8, 10, 12) and recall breakdowns (Figures 9, 11, 13).

Language Analysis. Figure 7 shows the distribution of PluRule instances across 9 languages and per-language accuracy for all evaluated models (Qwen3-VL-4B/8B/30B instruct and thinking; GPT-5.2 low and high reasoning) with full context.

Universality Correlation. To examine whether models perform better on universal rule clusters, we measure universality in two ways: (a) the number of subreddits containing at least one rule in the cluster, and (b) the number of member rules in the cluster. Figures 14, 15, 16, and 17 plot accuracy against both measures. Spearman correlations are weak ($\rho = 0.11$–0.40), with only Qwen3-VL-8B reaching significance. However, civility and self-promotion — the most universal by both measures — consistently rank among the highest-accuracy clusters, with the exception of self-promotion for Qwen3-VL-4B.

Input Prompt

Subreddit Info:

Subreddit: r/santamonica: Welcome to Santa Monica, California

Anything related to Santa Monica, where the rent is high but the temperatures aren't.

Rules:

Rule 1: Santa Monica-related only

Please keep posts to Santa Monica-specific issues. Venice and West LA issues can be posted in /r/LosAngeles

Rule 2: Respect other redditors

Respect other r/SantaMonica users, as both individuals and as groups. If you disagree with someone here, do so gracefully without personal attacks. And remember, members of Santa Monica's unhoused population use this subreddit, too.

Rule 3: No spam

We allow local businesses to share here if they're offering specials to the r/SantaMonica community. If multiple posts are submitted within short succession on the same topic (e.g. "earthquake!"), they will be consolidated down to 1.

Rule 4: Appealing bans

The moderators of this Subreddit seek to make this a place for civil discourse on Santa Monica and new users are held to a high standards for civility and quality of contributions. Mods enforce the rules using bans that allow people to read the rules. If you were banned and would like to continue participating in this forum, let us know that you've read the rules and why you'd like to continue participating in this forum versus other online forums in your ban appeal.

Rule 5: No low-effort posts

Sometimes low-quality posts get caught in the subreddit's filters. These are removed by auto-moderator and removal is confirmed by moderators. Include an adequate and accurate description in your post; provide sources if asked by the moderators. Don't use this Subreddit as your personal research assistant if you're not interested in searching the history.

Rule 6: No calling people paid shills just because they disagree with you.

Accusing people of being paid shills with no proof just because they disagree with you can and will result in bans.

Rule 7: No hateful speech directed at other people, including about the homeless.

If you want to say something disgusting about homeless people, NextDoor is always there for you.

Submission:

Homeowners association sues City and Fairmont Miramar over hotel redevelopment approval
USER1, Tue, Dec 15, 2020, 1:37PM

https://www.smdp.com/homeowners-association-sues-city-and-fairmont-miramar-over-hotel-redevelopment-approval/199916

Discussion:

Comment 1: USER1, Tue, Dec 15, 2020, 1:38PM

And people wonder why everything is so expensive here. This redevelopment took 10 **years** to get approved and they're still not out of the woods on getting held up by NIMBY lawsuits. These risks get priced in on any development project and it's basically a guarantee a lot of would-be projects never get off the ground in the first place because they know they're risking interminable timelines like this.

Comment 2: USER2, Tue, Dec 15, 2020, 1:48PM

Shrug.
How many more empty storefronts does Santa Monica need?
Third Street Promenade has thousands of square feet of empty available commercial space. The former Barnes & Noble, the former Banana Republic. Dozens of empty storefronts a block away.
This is a pointless multi-million dollar development that no one is asking for.
Meanwhile thousands of unhoused people on the streets. . .

Comment 3: USER3, Tue, Dec 15, 2020, 4:45PM

This is a hotel. Hotels bring more patrons for our local businesses. They also produce a ton of tax revenue.

Comment 4 [TARGET COMMENT]: USER4, Mon, Dec 21, 2020, 3:58PM

You must work for Dell. He's been ripping off Santa Monica for years by avoiding taxes on the Fairmont Miramar. Look it up. He's a right wing a-hole

Question: Does the [TARGET COMMENT] violate a rule?

(a) Santa Monica-related only; (b) No rules broken; (c) Respect other redditors; (d) Appealing bans; (e) No calling people paid shills just because they disagree with you; (f) No low-effort posts; (g) No hateful speech directed at other people, including about the homeless; (h) No spam

Model Response

Reasoning trace: Evaluating user responses. I think I'd choose option (c): "respect other redditors" because calling someone a "right wing a-hole" definitely feels like a personal attack. The question is about whether that behavior violates a rule, and it clearly does. So, saying option (c) is the best choice makes sense to me. It's really important to foster a respectful community, and this kind of language just doesn't fit that.

Final answer: (c) Respect other redditors. ✗

Figure 5: Full PLURULE example with complete input prompt and reasoning trace from GPT-5.2 (high reasoning). The correct answer is (e) No calling people paid shills, since "You must work for Dell" is an unsupported accusation. The model instead locks onto the incivility ("right wing a-hole") and picks (c) Respect other redditors.

Model	Variant	Metric	Comment	+Discussion	+Submission	+User	+Images
Qwen3-VL-4B	Instruct	V. Recall	24.1	26.4 (+2.3)	29.8 (+3.4)	30.7 (+0.9)	**32.6** (+1.9)
		C. Recall	**75.1**	72.0 (-3.1)	66.7 (-5.3)	67.1 (+0.4)	64.3 (-2.8)
		Accuracy	**49.6**	49.2 (-0.4)	48.3 (-0.9)	48.9 (+0.6)	48.4 (-0.5)
	Thinking	V. Recall	24.3	26.1 (+1.8)	27.9 (+1.8)	**30.3** (+2.4)	29.0 (-1.3)
		C. Recall	50.5	53.5 (+3.0)	**62.0** (+8.5)	59.8 (-2.2)	60.9 (+1.1)
		Accuracy	37.4	39.8 (+2.4)	44.9 (+5.1)	**45.0** (+0.1)	45.0 (+0.0)
Qwen3-VL-8B	Instruct	V. Recall	27.9	27.8 (-0.1)	27.7 (-0.1)	29.6 (+1.9)	**29.9** (+0.3)
		C. Recall	**74.1**	73.7 (-0.4)	70.8 (-2.9)	70.4 (-0.4)	69.6 (-0.8)
		Accuracy	**51.0**	50.7 (-0.3)	49.2 (-1.5)	50.0 (+0.8)	49.8 (-0.2)
	Thinking	V. Recall	31.2	33.2 (+2.0)	32.6 (-0.6)	**34.7** (+2.1)	32.3 (-2.4)
		C. Recall	49.3	54.6 (+5.3)	**61.8** (+7.2)	58.6 (-3.2)	57.5 (-1.1)
		Accuracy	40.3	43.9 (+3.6)	**47.2** (+3.3)	46.7 (-0.5)	44.9 (-1.8)
Qwen3-VL-30B	Instruct	V. Recall	30.2	31.4 (+1.2)	31.1 (-0.3)	**32.4** (+1.3)	31.6 (-0.8)
		C. Recall	70.1	70.5 (+0.4)	71.1 (+0.6)	72.4 (+1.3)	**72.9** (+0.5)
		Accuracy	50.2	51.0 (+0.8)	51.1 (+0.1)	**52.4** (+1.3)	52.3 (-0.1)
	Thinking	V. Recall	38.1	**41.0** (+2.9)	40.0 (-1.0)	41.0 (+1.0)	40.3 (-0.7)
		C. Recall	54.1	55.4 (+1.3)	58.3 (+2.9)	57.8 (-0.5)	**58.6** (+0.8)
		Accuracy	46.1	48.2 (+2.1)	49.1 (+0.9)	49.4 (+0.3)	**49.5** (+0.1)
GPT-5.2	Low	V. Recall	40.1	41.2 (+1.1)	40.8 (-0.4)	**42.1** (+1.3)	41.7 (-0.4)
		C. Recall	68.1	69.3 (+1.2)	72.9 (+3.6)	72.7 (-0.2)	**73.1** (+0.4)
		Accuracy	54.1	55.3 (+1.2)	56.8 (+1.5)	**57.4** (+0.6)	**57.4** (+0.0)
	High	V. Recall	40.9	42.8 (+1.9)	42.2 (-0.6)	**43.3** (+1.1)	42.6 (-0.7)
		C. Recall	69.2	69.7 (+0.5)	72.4 (+2.7)	72.1 (-0.3)	**72.7** (+0.6)
		Accuracy	55.0	56.2 (+1.2)	57.3 (+1.1)	**57.7** (+0.4)	57.6 (-0.1)
Baseline		V. Recall			0.0		
		C. Recall			100.0		
		Accuracy			50.0		

Table 3: Violating recall, compliant recall, and accuracy (%) across different models and contexts on the test set. Numbers in parentheses indicate differences compared to values in the previous column. Best-performing contexts for each model variant are highlighted in bold. All values have 95% CI of at most $\pm 1.9\%$.

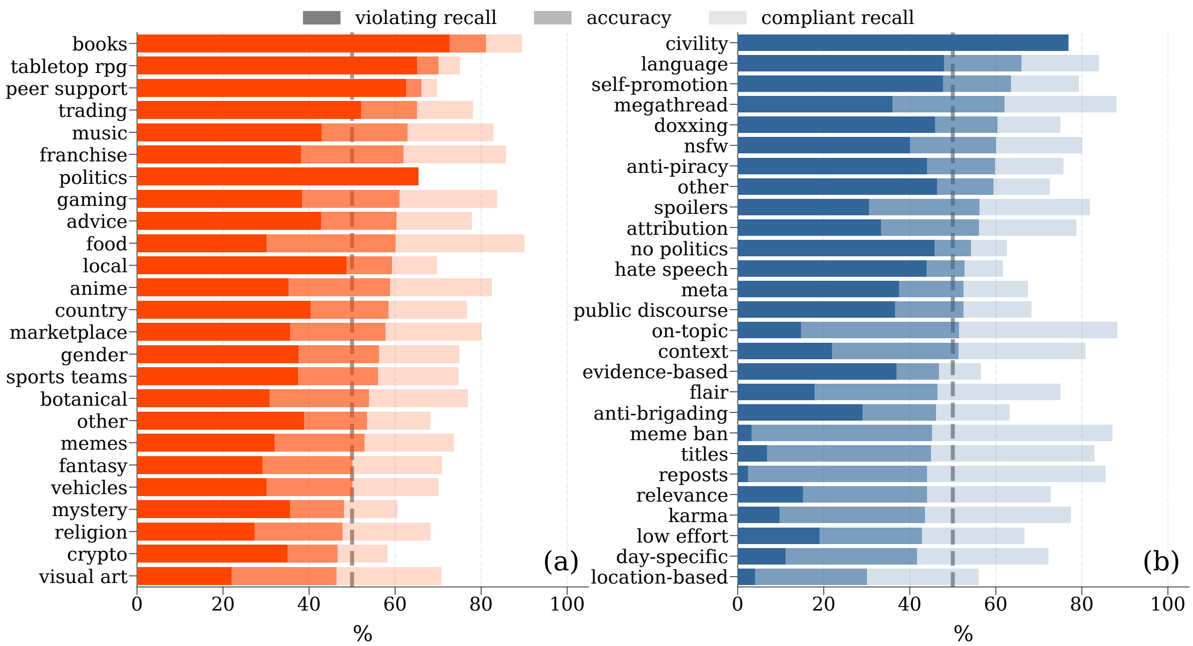

Figure 6: GPT-5.2 (high reasoning) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster on the test set. Stacked bars show violating and compliant recall. Bars sorted by accuracy. Dashed lines indicate the 50% baseline for accuracy.

Models	**Qwen3-VL-4B**		**Qwen3-VL-8B**		**Qwen3-VL-30B**		**GPT-5.2**	
Variants	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Low	High
Comment Only	41.1	33.1	**43.3**	37.3	43.5	43.4	49.4	50.3
+Discussion	41.6 (+0.5)	35.2 (+2.1)	43.1 (-0.2)	40.3 (+3.0)	44.5 (+1.0)	45.8 (+2.4)	50.6 (+1.2)	51.8 (+1.5)
+Submission	42.1 (+0.5)	39.2 (+4.0)	42.0 (-1.1)	42.3 (+2.0)	44.5 (+0.0)	46.1 (+0.3)	51.5 (+0.9)	52.3 (+0.5)
+User	42.8 (+0.7)	**40.1** (+0.9)	43.2 (+1.2)	**42.7** (+0.4)	**45.7** (+1.2)	**46.6** (+0.5)	**52.3** (+0.8)	**52.9** (+0.6)
+Images	**43.1** (+0.3)	39.6 (-0.5)	43.2 (+0.0)	40.7 (-2.0)	45.4 (-0.3)	46.4 (-0.2)	52.2 (-0.1)	52.6 (-0.3)
Baseline	33.3							

Table 4: Weighted accuracy (%) with violating:compliant ratio of 2:1 across models and context levels on the test set. Numbers in parentheses show differences from the previous row. Best-performing contexts for each model variant are highlighted in bold. The no moderation baseline drops to 33.3% under this weighting.

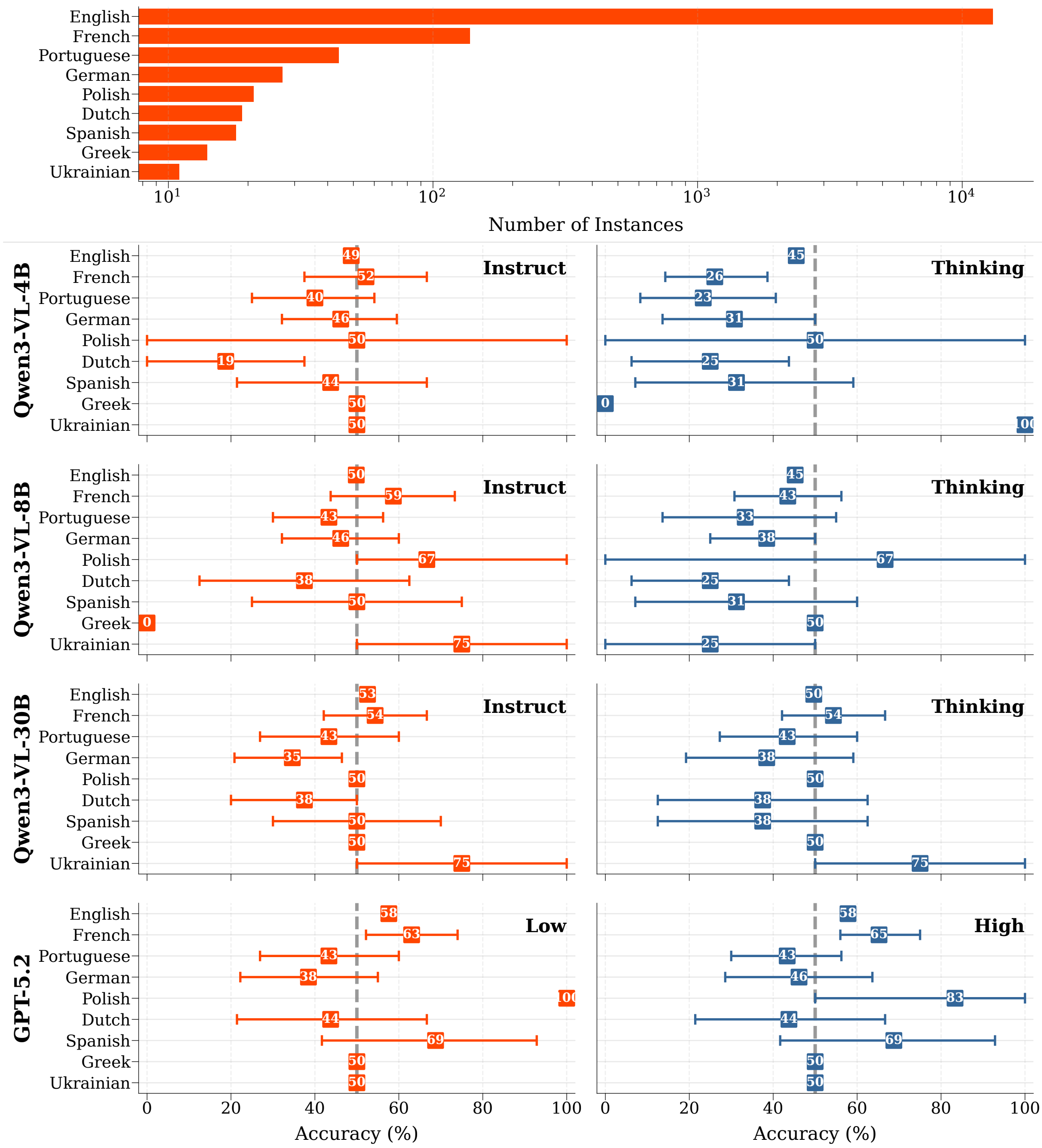

Figure 7: Language analysis across all evaluated models with full context. Top: distribution of PluRule instances across 9 languages. Bottom: per-language accuracy for Qwen3-VL-4B/8B/30B (instruct vs. thinking) and GPT-5.2 (low vs. high reasoning). Error bars show 95% CI. Dashed lines indicate the 50% baseline.

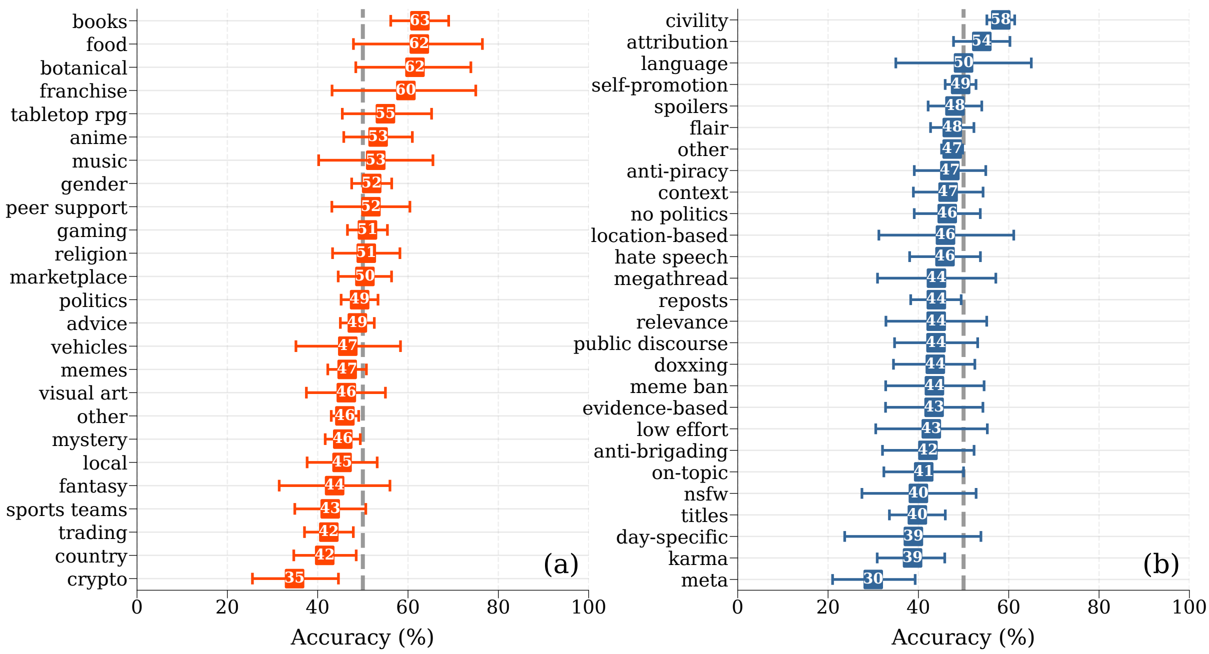

Figure 8: Accuracy for Qwen3-VL-4B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

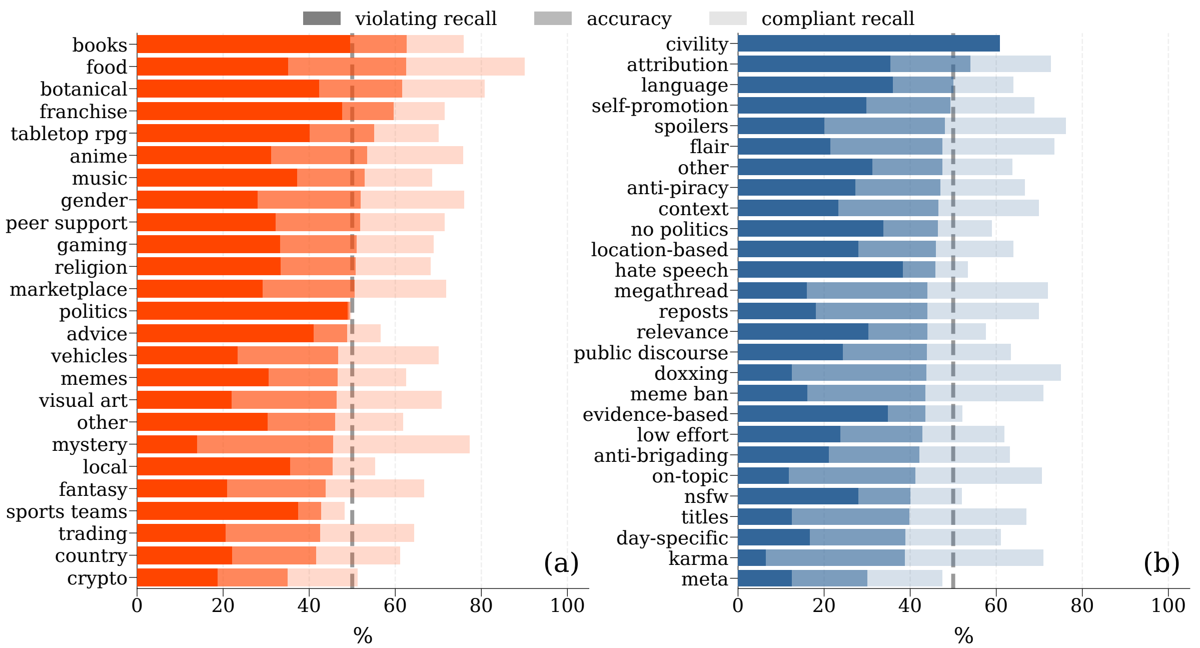

Figure 9: Qwen3-VL-4B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

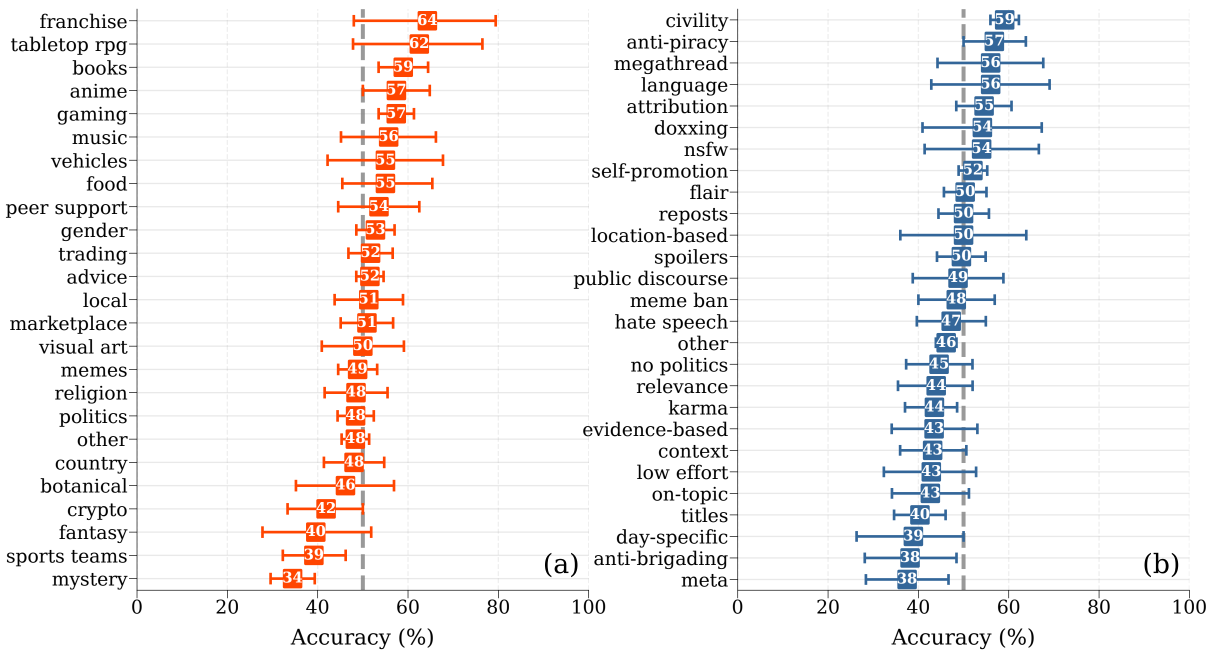

Figure 10: Accuracy for Qwen3-VL-8B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

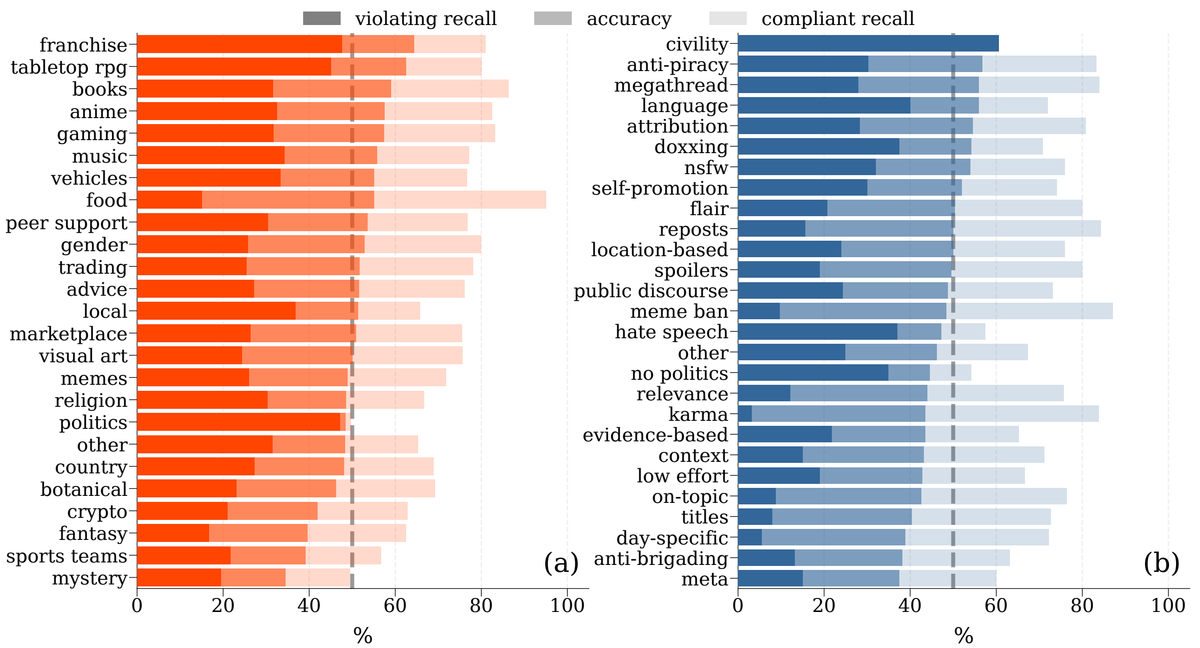

Figure 11: Qwen3-VL-8B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

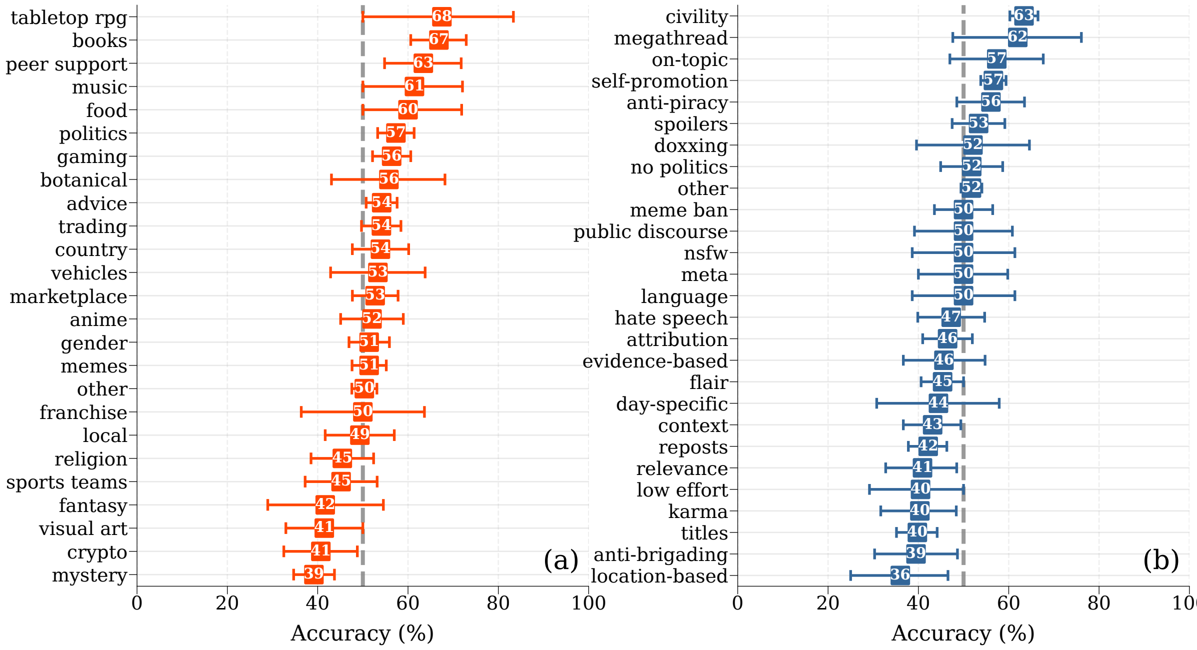

Figure 12: Accuracy for Qwen3-VL-30B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

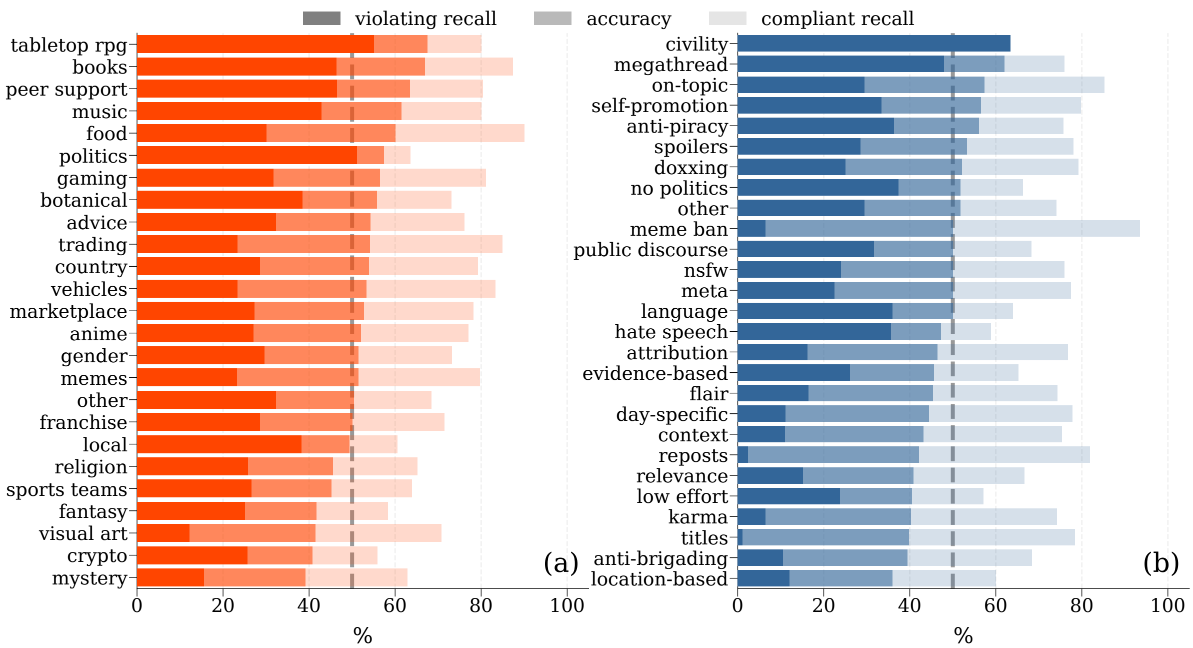

Figure 13: Qwen3-VL-30B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

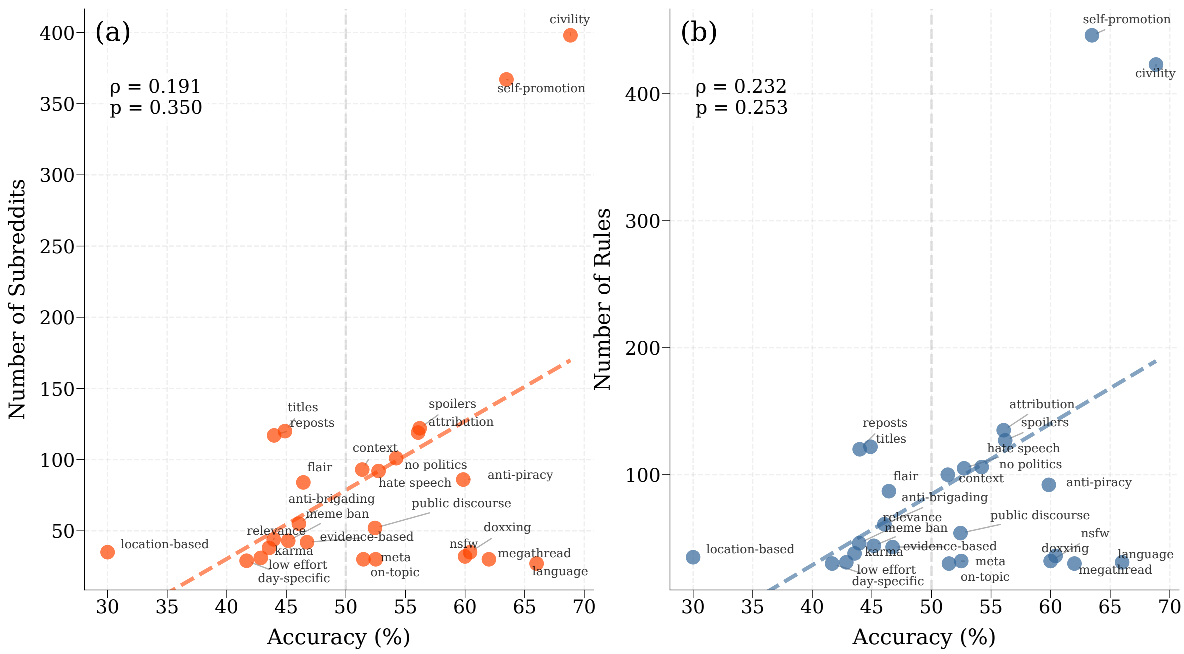

Figure 14: GPT-5.2 (high reasoning) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

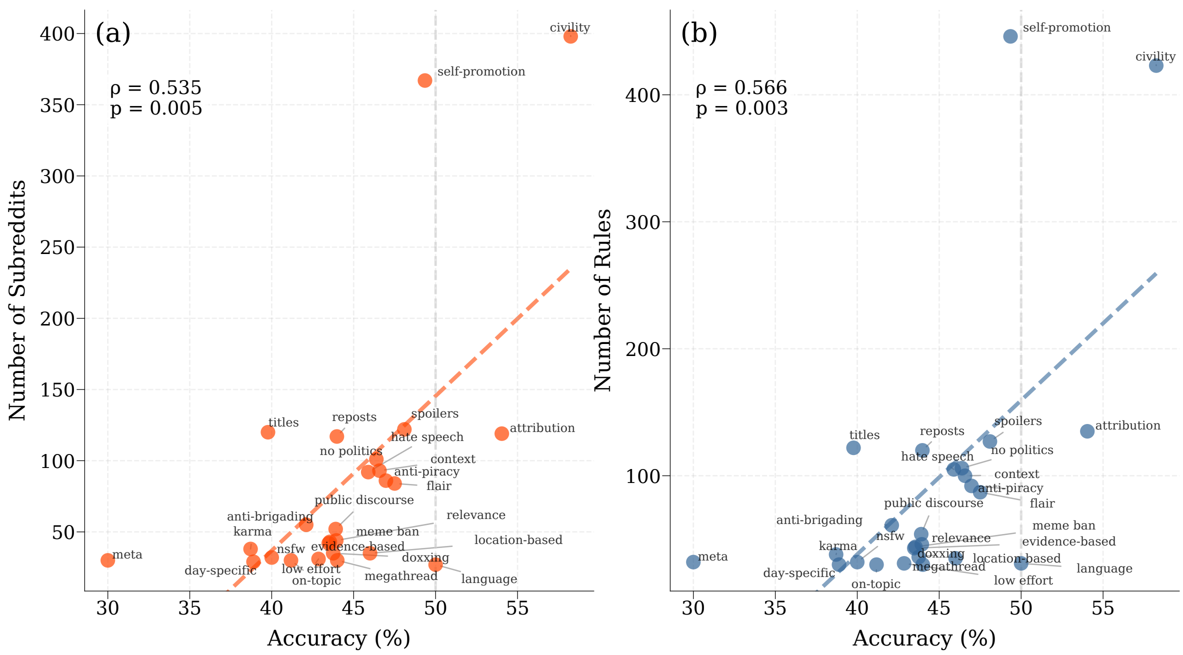

Figure 15: Qwen3-VL-4B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

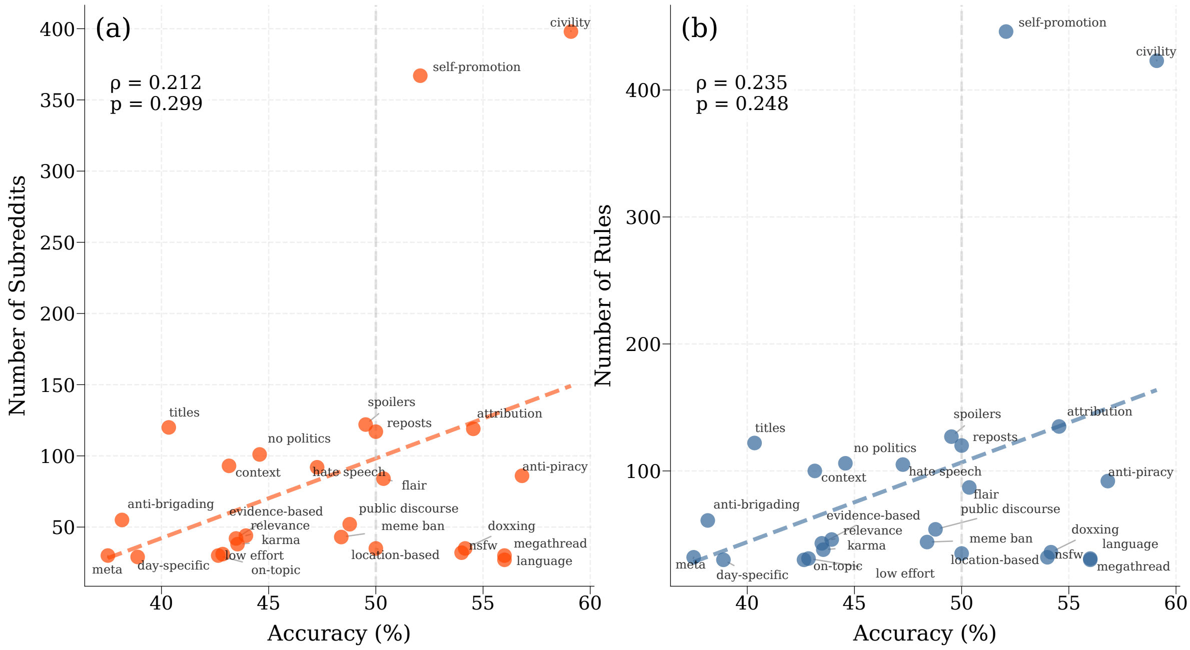

Figure 16: Qwen3-VL-8B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

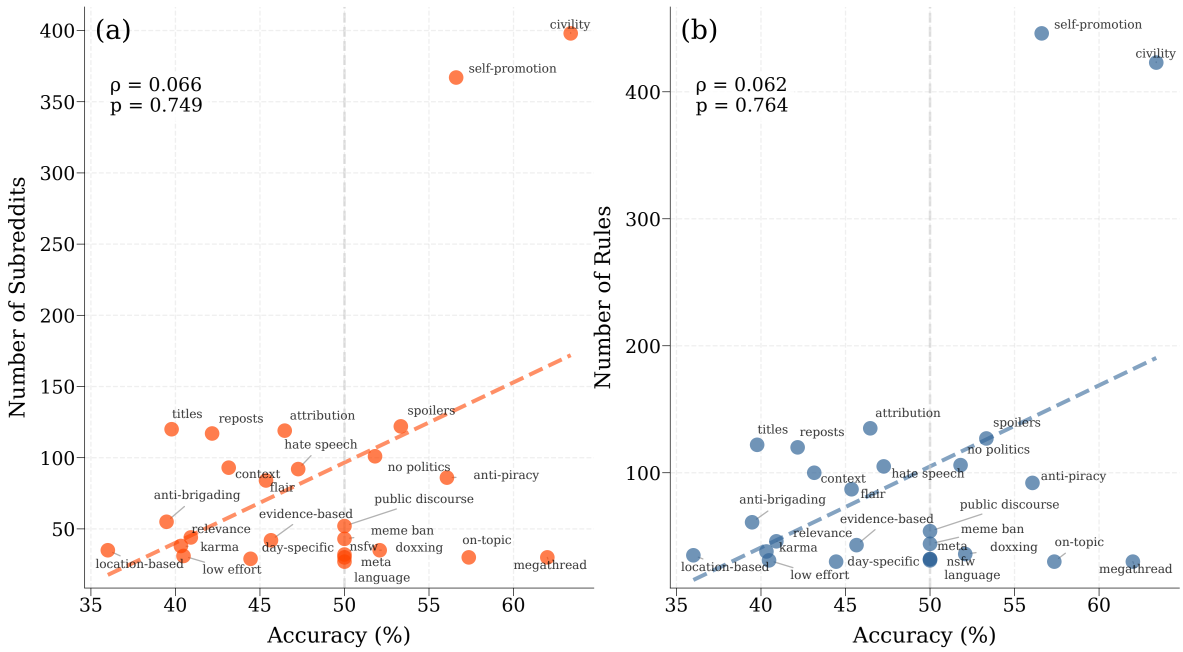

Figure 17: Qwen3-VL-30B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.