PLURULE: A Challenging Benchmark for Detecting Rule Violations of Pluralistic Communities on Social Media

Anonymous ACL submission

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 detect rule violations of 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 17,313 rule violations across 2,419 Reddit communities spanning 3,692 pluralistic rules in 10 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 pluralistic moderation of social media is a fundamental challenge for language models.

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 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

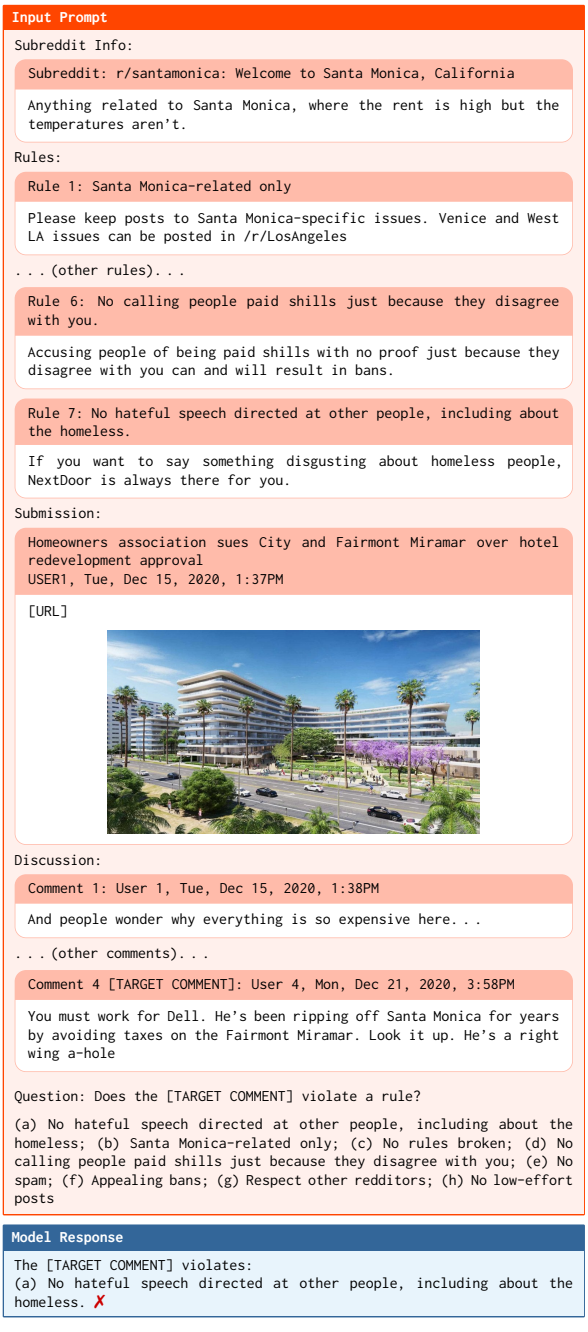

Figure 1: A PLURULE example. The model receives the full context of the target comment, then selects which rule is violated. Here, the correct answer is (d) but the model selects (a).

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 detecting violations of pluralistic community rules. The benchmark comprises 17,313 moderation instances with 92,305 comments and 4,843 images, spanning 2,419 subreddits with 3,692 distinct rules across 10 languages. PLURULE incorporates substantial diversity along two dimensions: 27 semantically-derived subreddit categories (e.g., politics, gaming, music) and 31 rule categories (e.g., civility, self-promotion, spoilers).

Using PLURULE, we evaluate state-of-the-art vision-language models (VLMs) on the detection of pluralistic 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 (67%), but fail on rules that require contextual understanding; sourcing requirements (30%), content neutrality (40%), and topic relevance (43%) all fall well 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 is inherently community-dependent and must account for pluralistic, community-specific rules. 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 detecting rule violations of 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 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 each comment, 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 version of the Pushshift Reddit archives (Cohen and Lo, 2014), spanning from December 2005 to February 2023 and 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.

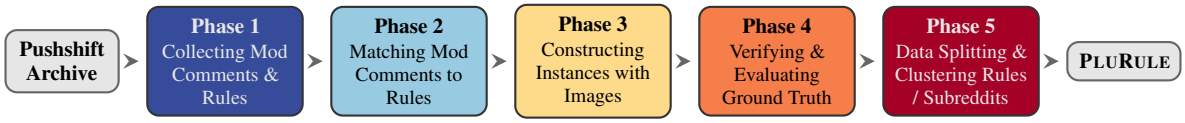

Figure 2: PLURULE construction pipeline

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 99th percentile of similarity scores (0.78): 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. From 9 million moderator comments, 174,412 ambiguous cases are discarded, yielding 775,594 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 84.9 million comments. We then build comment trees representing the reply structure. From each tree, we extract a *violating thread*: the path from the rule-violating comment up 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 pair each violating thread with a *compliant thread* — a discussion branch from the same submission that received no moderator action. We first select candidate compliant threads with depth equal to the violating thread (n) or, if not possible, $n-1$.

We then apply five 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.

After filtering, we rank candidate compliant threads for each violating thread using three criteria to maximize the shared context between the two: (1) higher thread depth to prioritize n over $n-1$; (2) higher number of common ancestors; 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 the benchmark instances, we finally collect full submission objects from Pushshift for all thread pairs. We remove thread pairs for submissions with NSFW content, crossposts, and videos. For the remaining submissions, we download images using a priority hierarchy: gallery images,

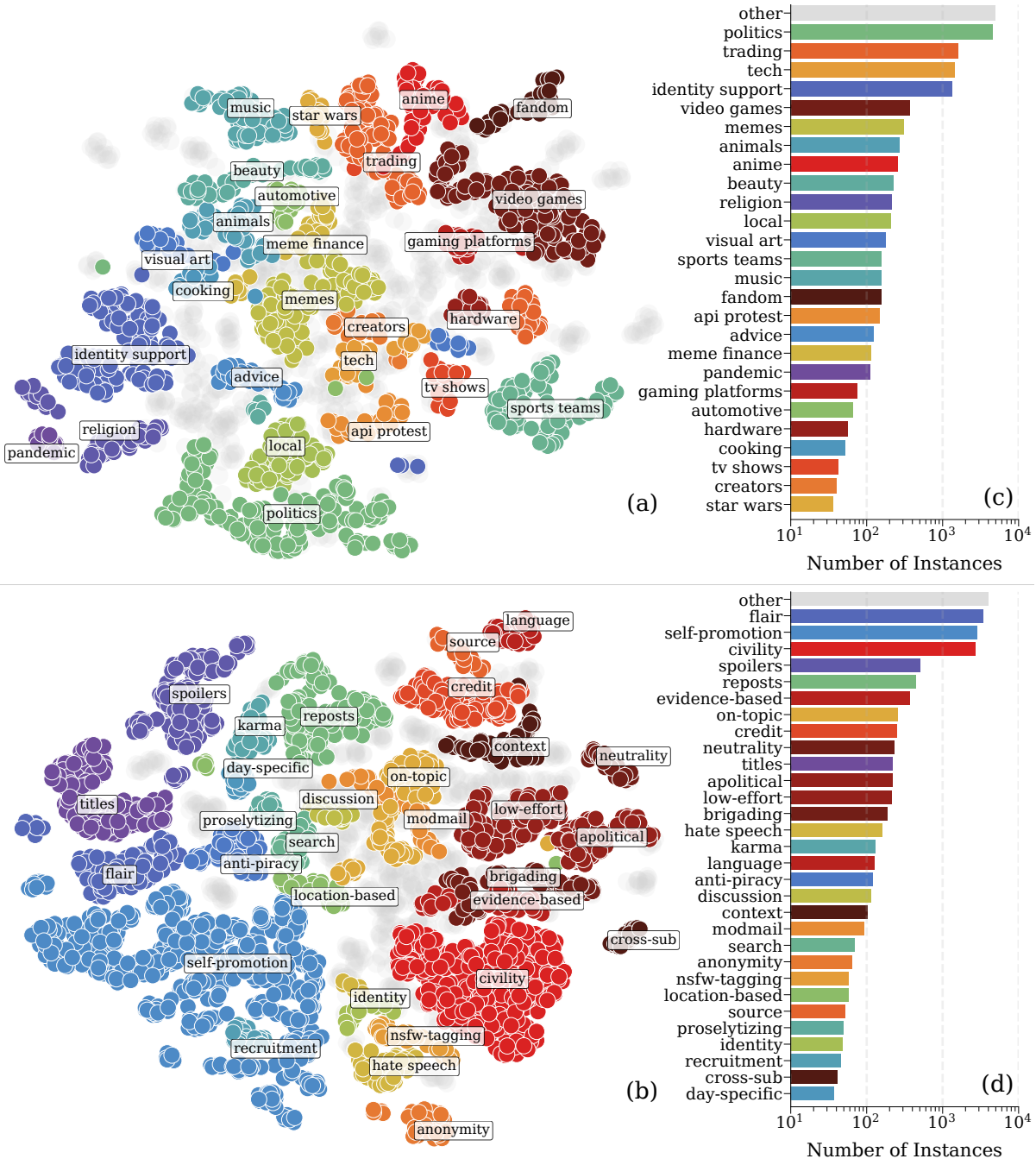

Figure 3: Two dimensions of pluralism in the PLURULE benchmark. Left: 2D UMAP visualizations of (a) 2,419 subreddits and (b) 3,692 rules, with colors indicating cluster assignments by HDBSCAN. Grey points represent unclustered items categorized as "other". Right: Distributions of 17,313 instances across (c) 27 subreddit clusters and (d) 31 rule clusters, with bar colors matching the cluster colors.

Split	Instances	Comments	Images	Subreddits / Clusters	Rules / Clusters	Languages
Train	11,979	66,280	2,882	1,104 / 27	1,801 / 31	10
Val	1,809	9,810	512	703 / 27	779 / 31	10
Test	3,525	16,215	1,449	2,419 / 27	2,491 / 31	10
Total	**17,313**	**92,305**	**4,843**	**2,419 / 27**	**3,692 / 31**	**10**

Table 1: PLURULE statistics. Each instance consists of one rule-violating and one compliant comment from the same submission. All 2,419 subreddits appear in the test set. We count only the 10 languages with at least 10 instances each. There are 24 languages with at least one instance.

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 775,594 matched comments, we build 448,379 comment trees and successfully create 21,238 instances, totaling 42,476 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 81.5% verification rate (17,313 out of 21,238 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 78 of the comments in the sample, all three annotators agreed on the label. In 16 cases, the label was assigned based on a majority (two of the three annotators agreed). In the six 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 95% overall accuracy: 100% on full-agreement cases (78/78), 81.25% on majority-agreement cases (13/16), and 66.67% on adjudicated cases (4/6).

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 pluralistic 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 A). Figure 3 shows that rules cluster more cleanly than subreddits, reflecting their greater semantic coherence compared to the diversity of community topics.

We label each cluster using Qwen3-30B-A3B-Thinking (see Appendix A) followed by manual refinement. We assign 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 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.

Models	Qwen3-VL-4B		Qwen3-VL-8B		Qwen3-VL-30B		GPT-5.2	
Variants	Instruct	Thinking	Instruct	Thinking	Instruct	Thinking	Low	High
Comment Only	**49.3**	37.7	**50.9**	41.4	50.1	45.4	54.5	55.0
+Discussion	48.9 (-0.4)	40.2 (+2.5)	50.9 (+0.0)	44.6 (+3.2)	50.7 (+0.6)	46.8 (+1.4)	55.1 (+0.6)	55.8 (+0.8)
+Submission	47.6 (-1.3)	44.6 (+4.4)	49.8 (-1.1)	46.9 (+2.3)	51.6 (+0.9)	48.7 (+1.9)	57.1 (+2.0)	57.5 (+1.7)
+User	48.0 (+0.4)	44.5 (-0.1)	50.4 (+0.6)	**47.2** (+0.3)	52.2 (+0.6)	**49.2** (+0.5)	**57.8** (+0.7)	57.6 (+0.1)
+Images	47.2 (-0.8)	**44.7** (+0.2)	50.8 (+0.4)	45.7 (-1.5)	**52.7** (+0.5)	48.8 (-0.4)	57.3 (-0.5)	**58.2** (+0.6)
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 ±1.1%. The baseline corresponds to always predicting "no rules broken."

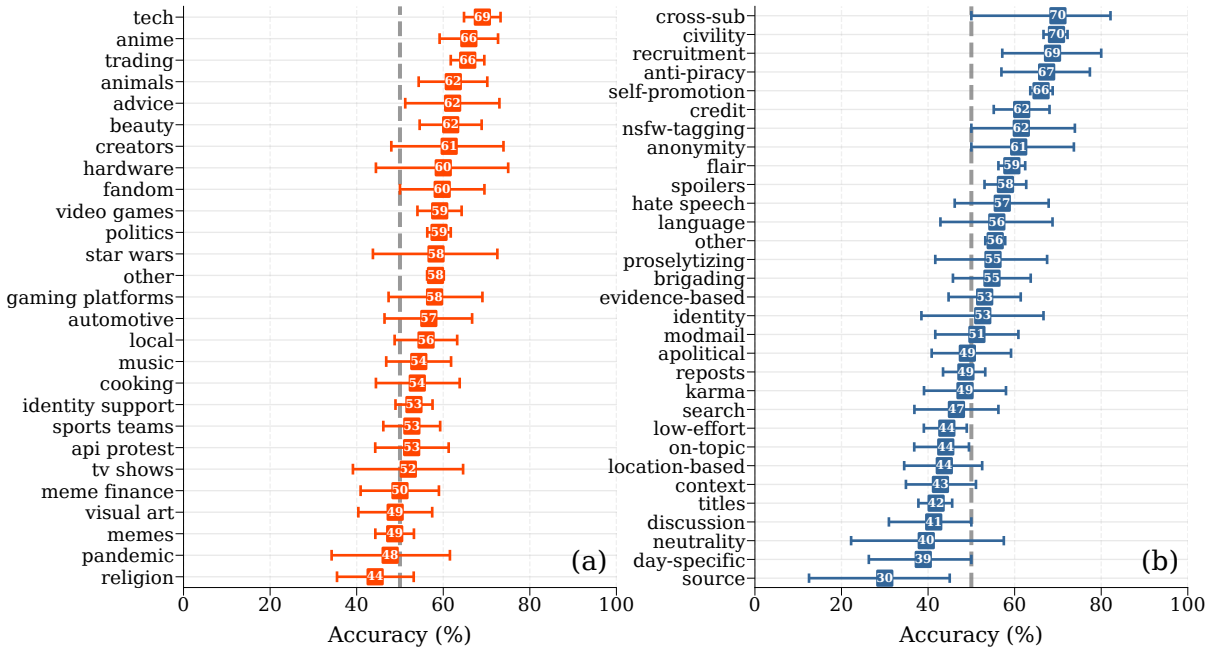

Figure 4: Accuracy for GPT-5.2 (high reasoning) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

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 B for violating/compliant comment breakdowns.) GPT-5.2 substantially outperforms all Qwen variants, with high reasoning effort achieving 58.2% accuracy — 8 points above the 50% baseline — while Qwen models barely exceed the baseline regardless of scale.

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 3.2 percentage points above comment-only, with submissions yielding the largest gain (+1.7 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 B for breakdowns by violating/compliant comments and languages. Performance varies substantially across community clusters: tech (69%), anime (66%), and trading (66%) subreddits achieve the highest accuracy, while religion (44%), pandemic (48%), and memes (49%) fall below baseline. Rule clusters show even greater variability: civility (70%), recruitment (69%), and self-promotion (66%) rules are detected reliably, but source requirements (30%), neutrality policies (40%), and on-topic rules (44%) perform worse than baseline. Similar results are obtained with other models (Appendix B).

Civility and self-promotion appear in far more subreddits and contain far more member rules than any other category (Appendix B). These universal categories consistently rank among the highest-accuracy rule clusters. In contrast, pluralistic rules requiring local context remain challenging.

6 Discussion

PLURULE provides a testbed for measuring progress toward models that can adapt to 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 fundamental gap: models succeed on universal violations like civility and self-promotion, but struggle with pluralistic rules that vary across communities. The two-dimensional variability — 44%–69% accuracy across community types and 30%–70% 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 may help models learn local norms, though scalability across thousands of communities remains 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?

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, and 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

PLURULE contains only publicly posted content from the Pushshift archives. We anonymize all usernames by replacing them with generic labels (USER1, USER2, etc.). We do not collect or release any private user metadata such as IP addresses, email addresses, or account history.

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.

References

Dalia Ali, Dora Zhao, Allison Koenecke, and Orestis Papakyriakopoulos. 2025. Operationalizing pluralis-

tic 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 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 = 0.400 with 26 clusters (62 items per cluster on average). We treat the remaining subreddits as the 'other' cluster (33.5% noise). For rules, optimal parameters yield DBCV = 0.524 with 30 clusters (100 items per cluster on average). We treat the remaining rules as the 'other' cluster (18.6% 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 31 rule clusters (e.g., civility, self-promotion, spoiler policy, flair enforcement) and 27 subreddit clusters (e.g., politics, meme culture, video games, trading).

B Extended Results

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

GPT-5.2 Analysis. Figure 5 shows recall by subreddit and rule clusters for GPT-5.2 (high reasoning) with full context. Figure 6 shows instance distribution and accuracy across 10 languages.

Qwen3-VL Analysis. For each Qwen3-VL model (4B, 8B, 30B), we provide three figures: accuracy by subreddit and rule clusters (Figures 7, 10, 13), recall breakdowns (Figures 8, 11, 14), and language analysis (Figures 9, 12, 15).

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 16, 17, 18, and 19 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.

Model	Variant	Metric	Comment	+Discussion	+Submission	+User	+Images
Qwen3-VL-4B	Instruct	V. Recall	23.8	26.8 (+3.0)	29.6 (+2.8)	29.6 (+0.0)	**31.3** (+1.7)
		C. Recall	**74.9**	70.9 (-4.0)	65.5 (-5.4)	66.4 (+0.9)	63.1 (-3.3)
		Accuracy	**49.3**	48.9 (-0.4)	47.6 (-1.3)	48.0 (+0.4)	47.2 (-0.8)
	Thinking	V. Recall	25.8	27.1 (+1.3)	27.7 (+0.6)	**29.6** (+1.9)	28.9 (-0.7)
		C. Recall	49.7	53.4 (+3.7)	**61.5** (+8.1)	59.4 (-2.1)	60.5 (+1.1)
		Accuracy	37.7	40.2 (+2.5)	44.6 (+4.4)	44.5 (-0.1)	**44.7** (+0.2)
Qwen3-VL-8B	Instruct	V. Recall	27.7	28.0 (+0.3)	28.0 (+0.0)	29.2 (+1.2)	**30.9** (+1.7)
		C. Recall	**74.1**	73.8 (-0.3)	71.6 (-2.2)	71.5 (-0.1)	70.8 (-0.7)
		Accuracy	**50.9**	50.9 (+0.0)	49.8 (-1.1)	50.4 (+0.6)	50.8 (+0.4)
	Thinking	V. Recall	33.0	33.9 (+0.9)	33.5 (-0.4)	**35.1** (+1.6)	33.6 (-1.5)
		C. Recall	49.8	55.3 (+5.5)	**60.4** (+5.1)	59.3 (-1.1)	57.8 (-1.5)
		Accuracy	41.4	44.6 (+3.2)	46.9 (+2.3)	**47.2** (+0.3)	45.7 (-1.5)
Qwen3-VL-30B	Instruct	V. Recall	31.3	31.7 (+0.4)	31.5 (-0.2)	32.7 (+1.2)	**33.7** (+1.0)
		C. Recall	68.9	69.6 (+0.7)	71.6 (+2.0)	71.7 (+0.1)	**71.8** (+0.1)
		Accuracy	50.1	50.7 (+0.6)	51.6 (+0.9)	52.2 (+0.6)	**52.7** (+0.5)
	Thinking	V. Recall	38.4	41.4 (+3.0)	40.2 (-1.2)	**41.5** (+1.3)	41.0 (-0.5)
		C. Recall	52.4	52.1 (-0.3)	**57.3** (+5.2)	56.9 (-0.4)	56.6 (-0.3)
		Accuracy	45.4	46.8 (+1.4)	48.7 (+1.9)	**49.2** (+0.5)	48.8 (-0.4)
GPT-5.2	Low	V. Recall	40.7	41.8 (+1.1)	42.1 (+0.3)	**43.4** (+1.3)	42.0 (-1.4)
		C. Recall	68.3	68.3 (+0.0)	72.2 (+3.9)	72.3 (+0.1)	**72.6** (+0.3)
		Accuracy	54.5	55.1 (+0.6)	57.1 (+2.0)	**57.8** (+0.7)	57.3 (-0.5)
	High	V. Recall	41.6	43.1 (+1.5)	43.1 (+0.0)	**43.8** (+0.7)	43.7 (-0.1)
		C. Recall	68.4	68.5 (+0.1)	72.0 (+3.5)	71.3 (-0.7)	**72.7** (+1.4)
		Accuracy	55.0	55.8 (+0.8)	57.5 (+1.7)	57.6 (+0.1)	**58.2** (+0.6)
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 $\pm 1.7\%$.

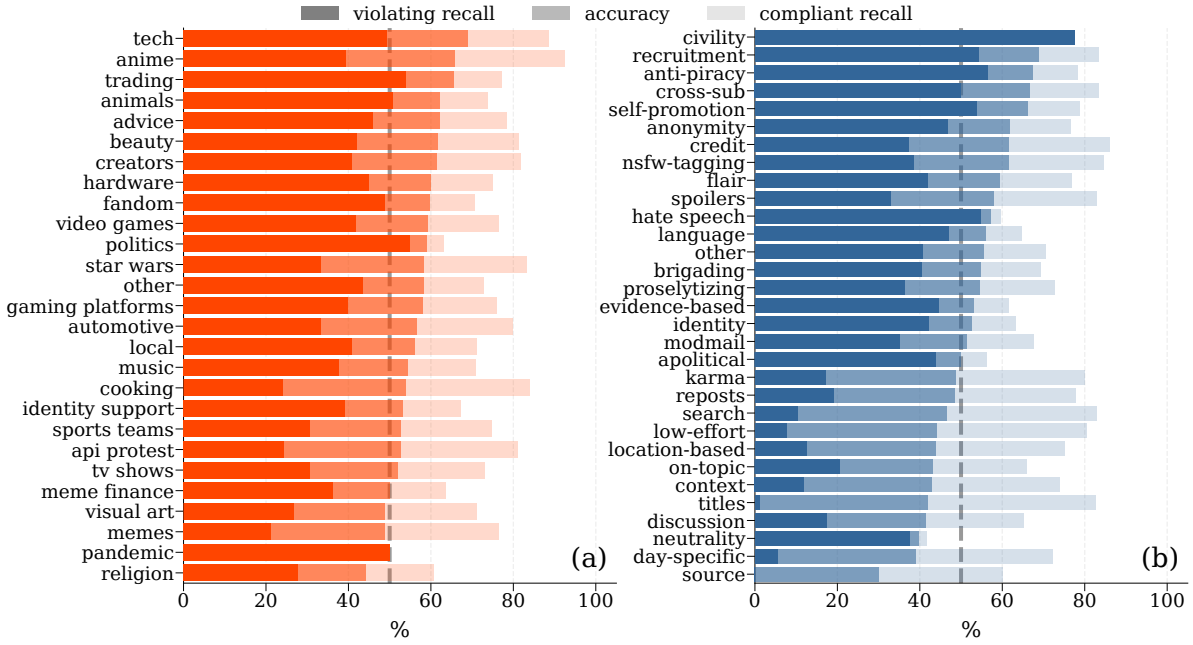

Figure 5: GPT-5.2 (high reasoning) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster on the test set. Stacked bars show violating and compliant recall. Bars sorted by accuracy. Dashed lines indicate the 50% baseline for accuracy.

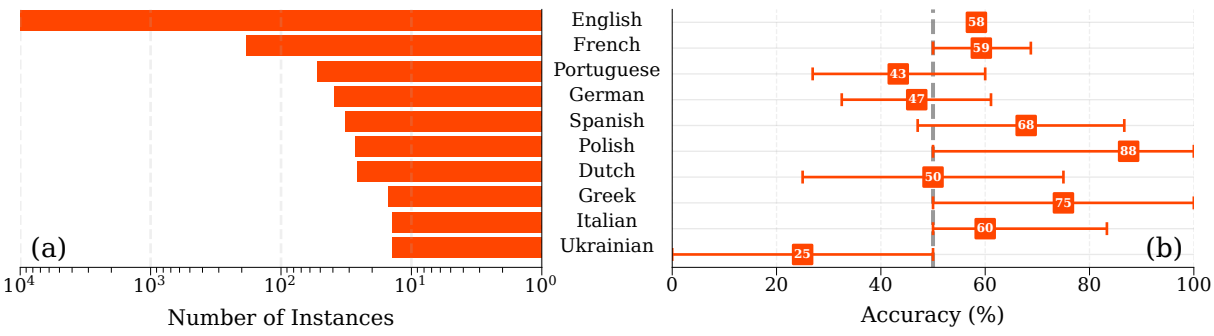

Figure 6: (a) Distribution of PluRule instances across 10 languages and (b) accuracy of the corresponding language in the test set by GPT-5.2 (high reasoning) with full context. Error bars show 95% CI. Dashed line indicates the 50% baseline.

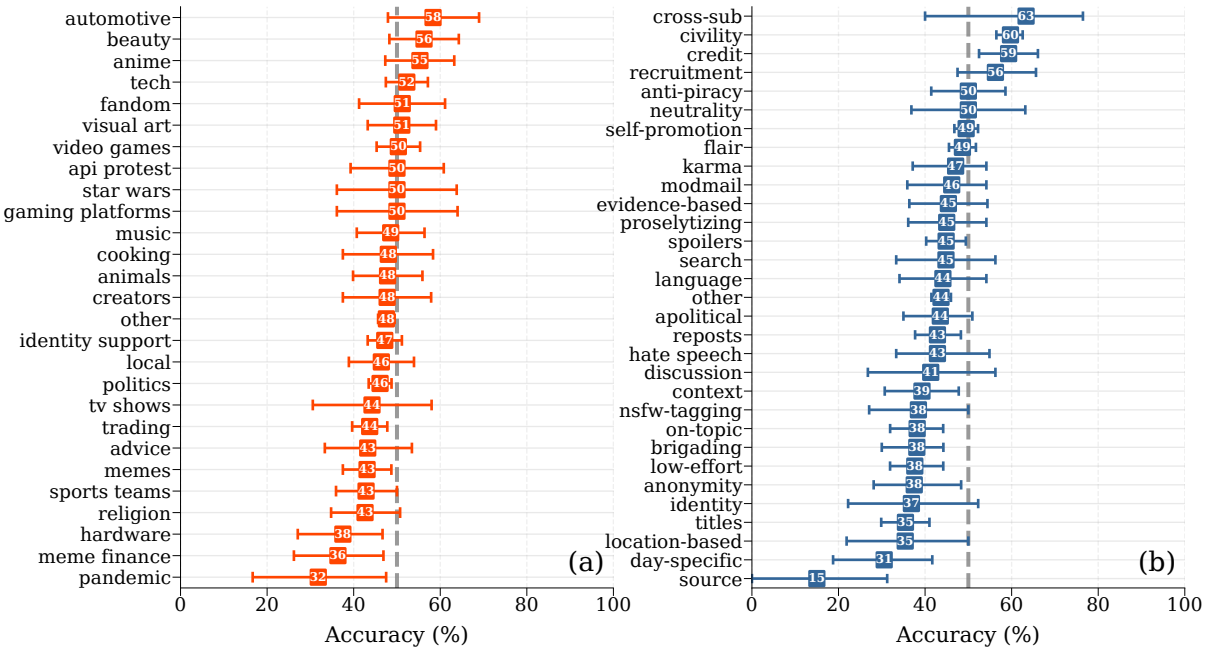

Figure 7: Accuracy for Qwen3-VL-4B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

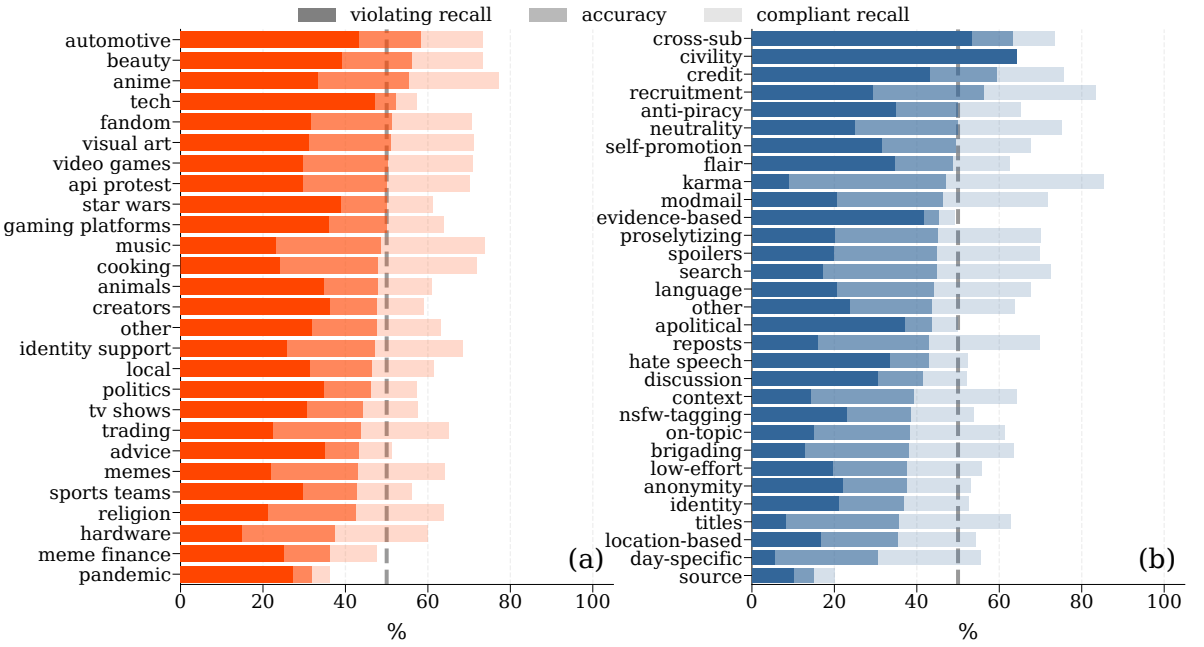

Figure 8: Qwen3-VL-4B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

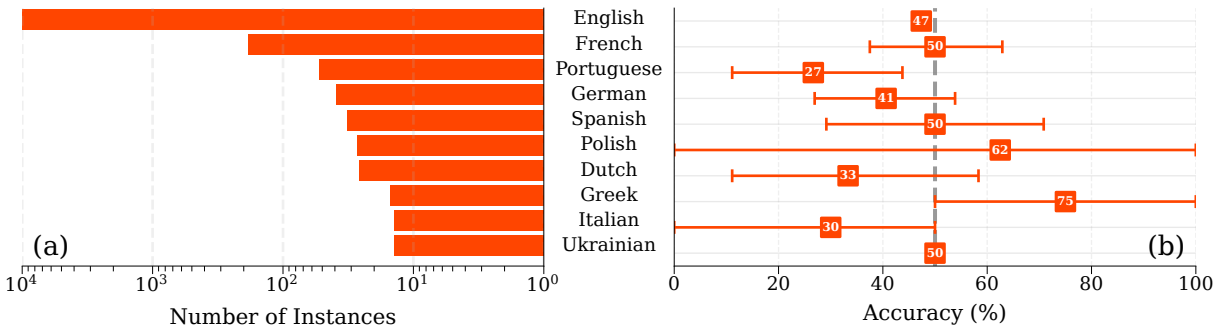

Figure 9: (a) Distribution of PluRule instances across 10 languages and (b) accuracy by language for Qwen3-VL-4B (instruct) with full context. Error bars show 95% CI. Dashed line indicates the 50% baseline.

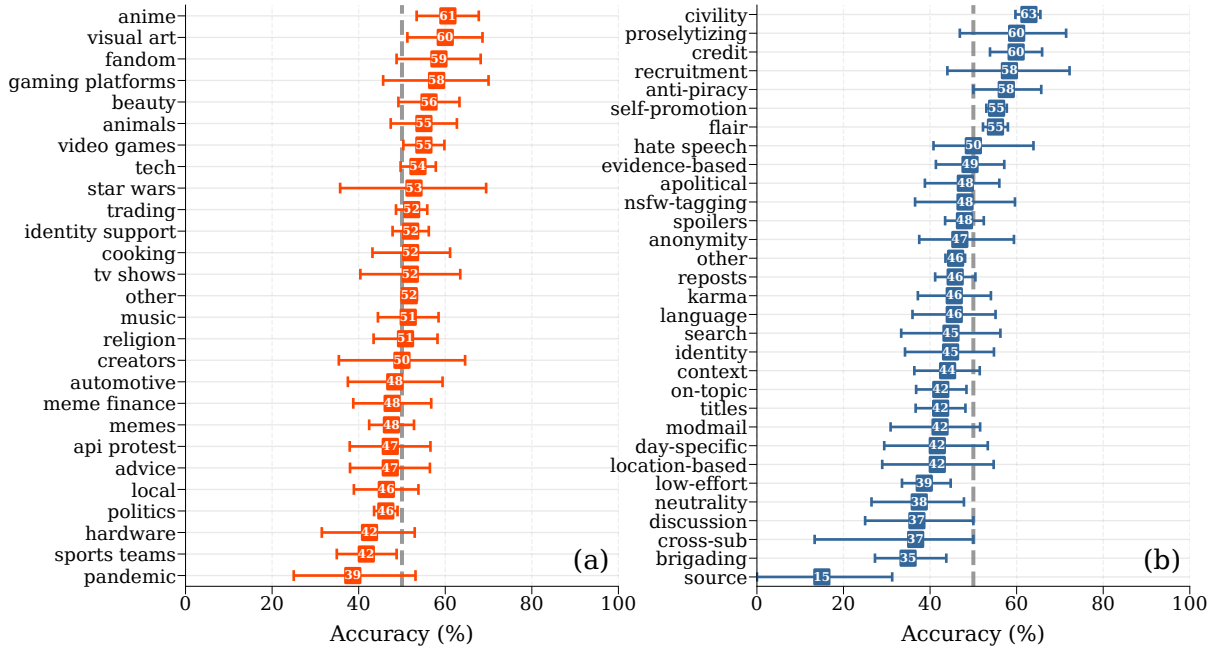

Figure 10: Accuracy for Qwen3-VL-8B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

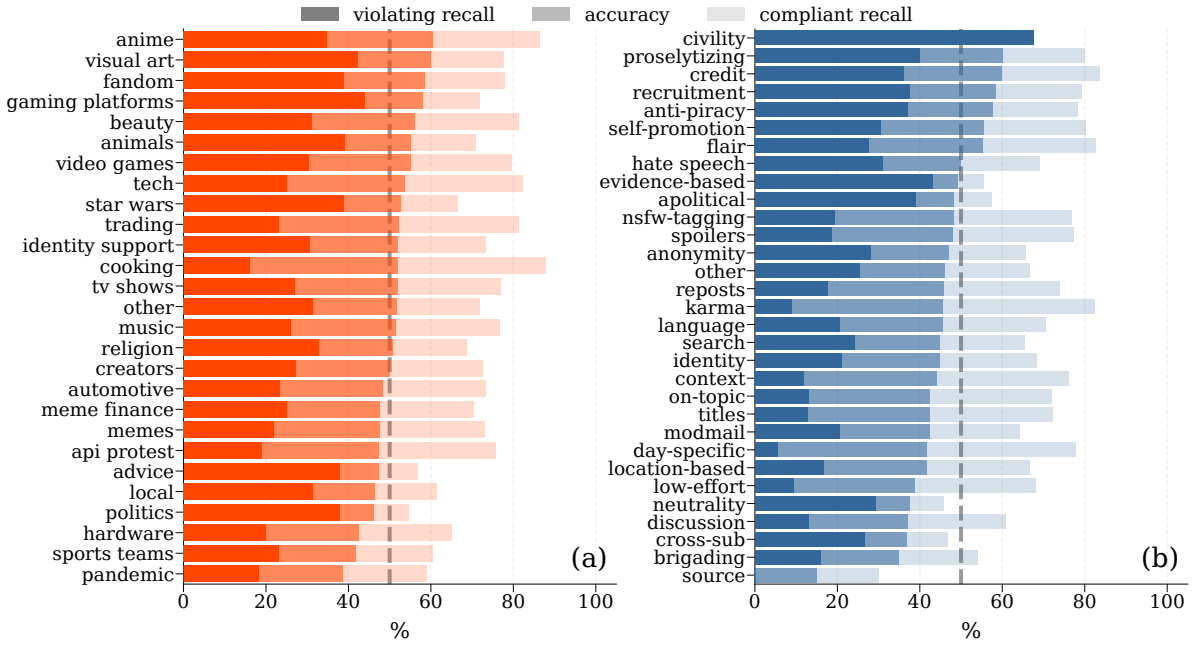

Figure 11: Qwen3-VL-8B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

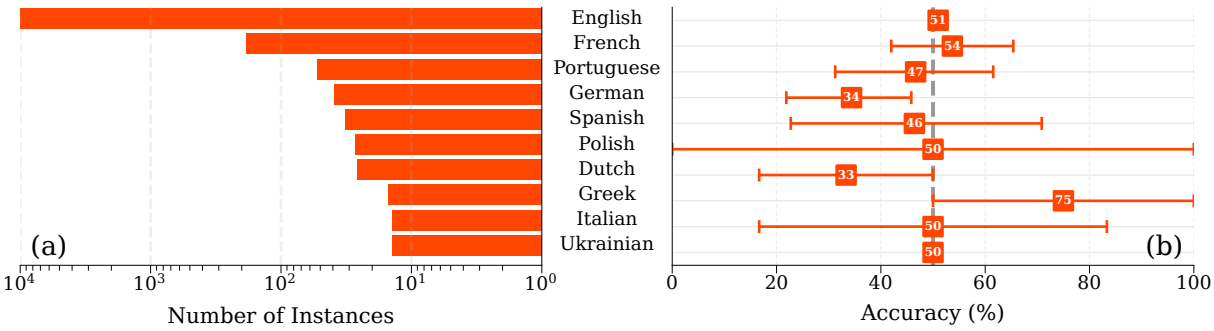

Figure 12: (a) Distribution of PluRule instances across 10 languages and (b) accuracy by language for Qwen3-VL-8B (instruct) with full context. Error bars show 95% CI. Dashed line indicates the 50% baseline.

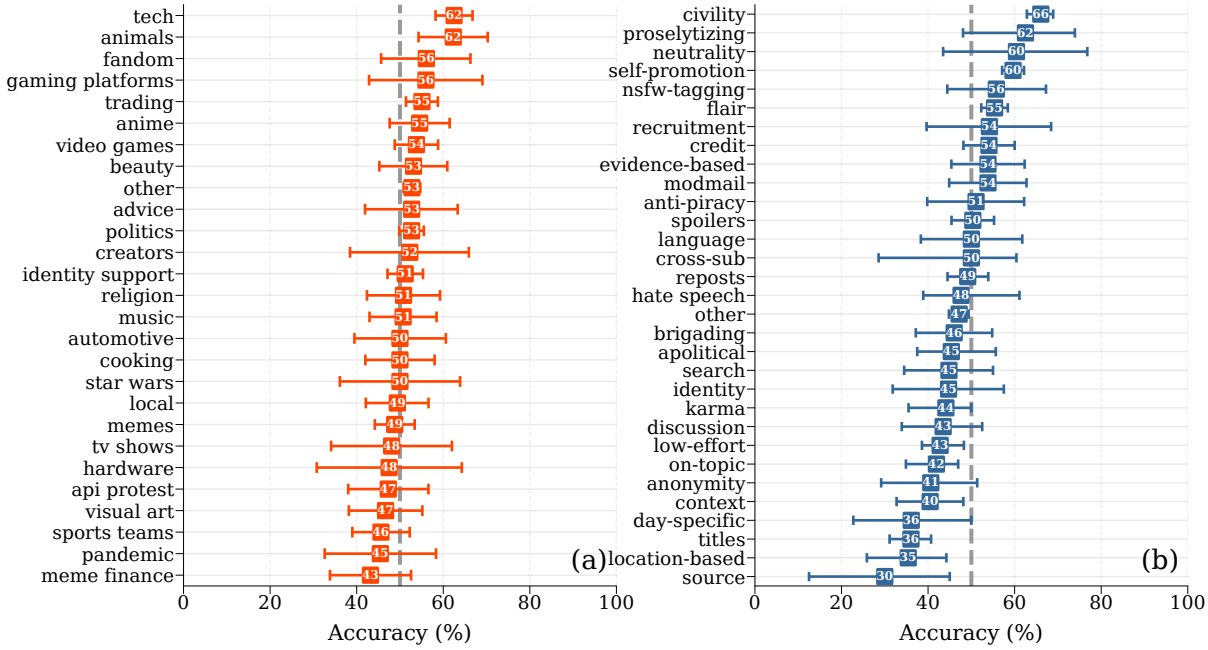

Figure 13: Accuracy for Qwen3-VL-30B (instruct) with full context by (a) subreddit clusters and (b) rule clusters. Error bars show 95% CI. Dashed lines indicate the baseline.

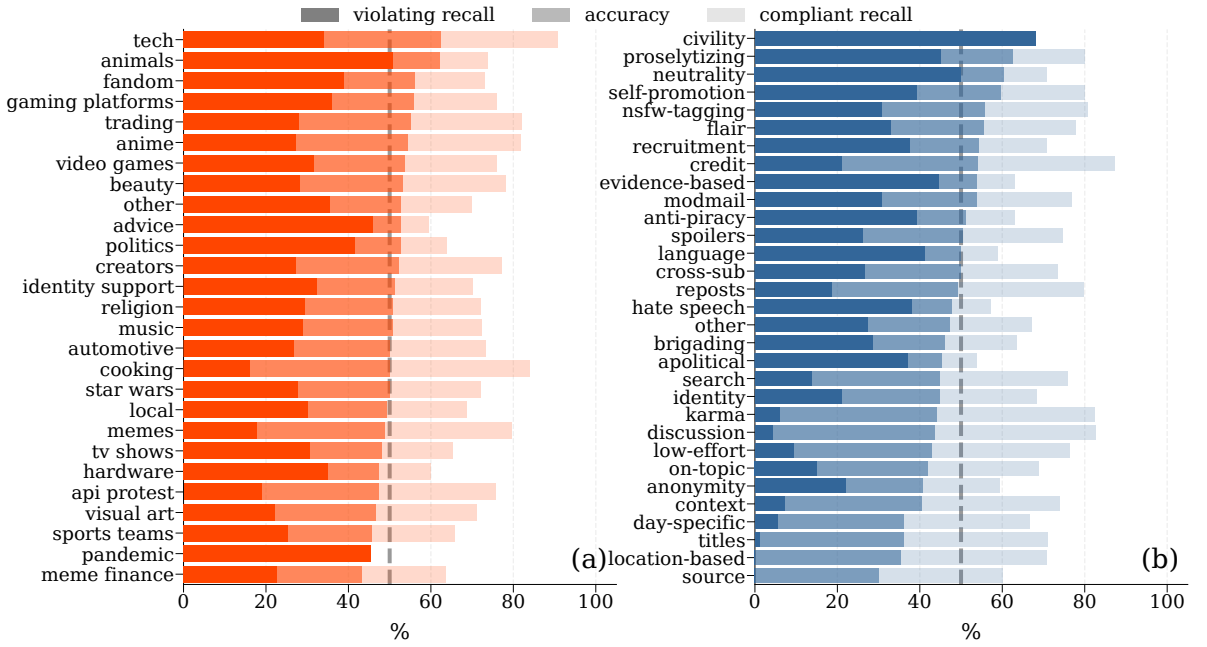

Figure 14: Qwen3-VL-30B (instruct) recall and accuracy with full context by (a) subreddit cluster and (b) rule cluster. Stacked bars show violating and compliant recall. Dashed lines indicate the 50% baseline.

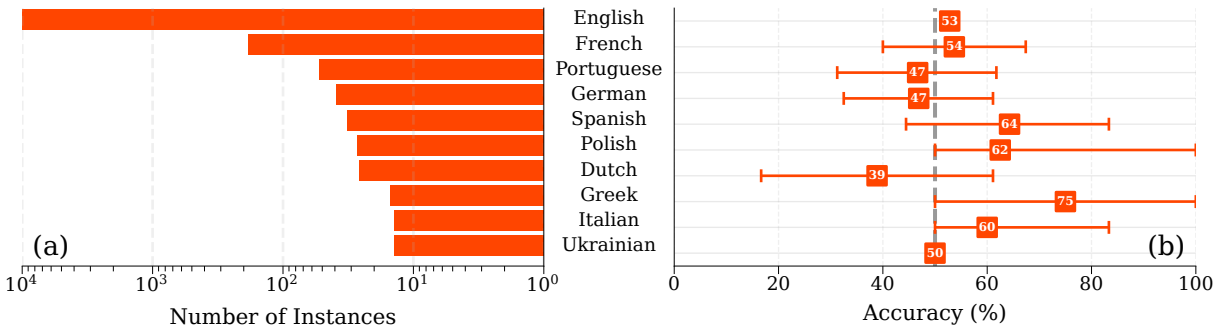

Figure 15: (a) Distribution of PluRule instances across 10 languages and (b) accuracy by language for Qwen3-VL-30B (instruct) with full context. Error bars show 95% CI. Dashed line indicates the 50% baseline.

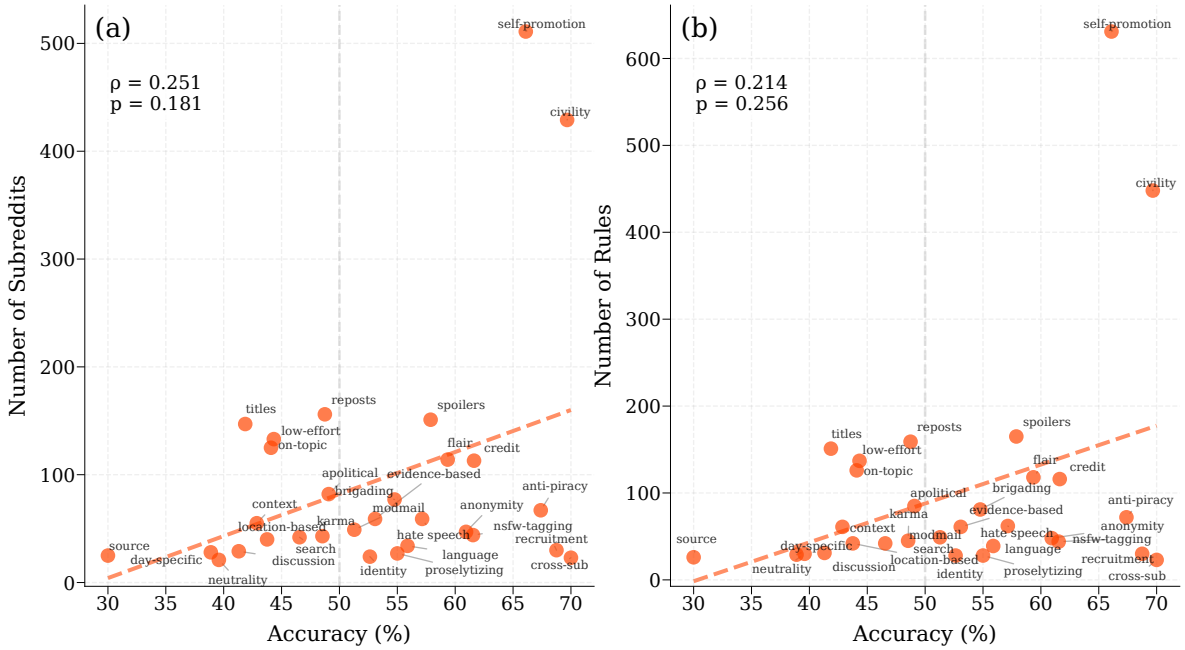

Figure 16: GPT-5.2 (high reasoning) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

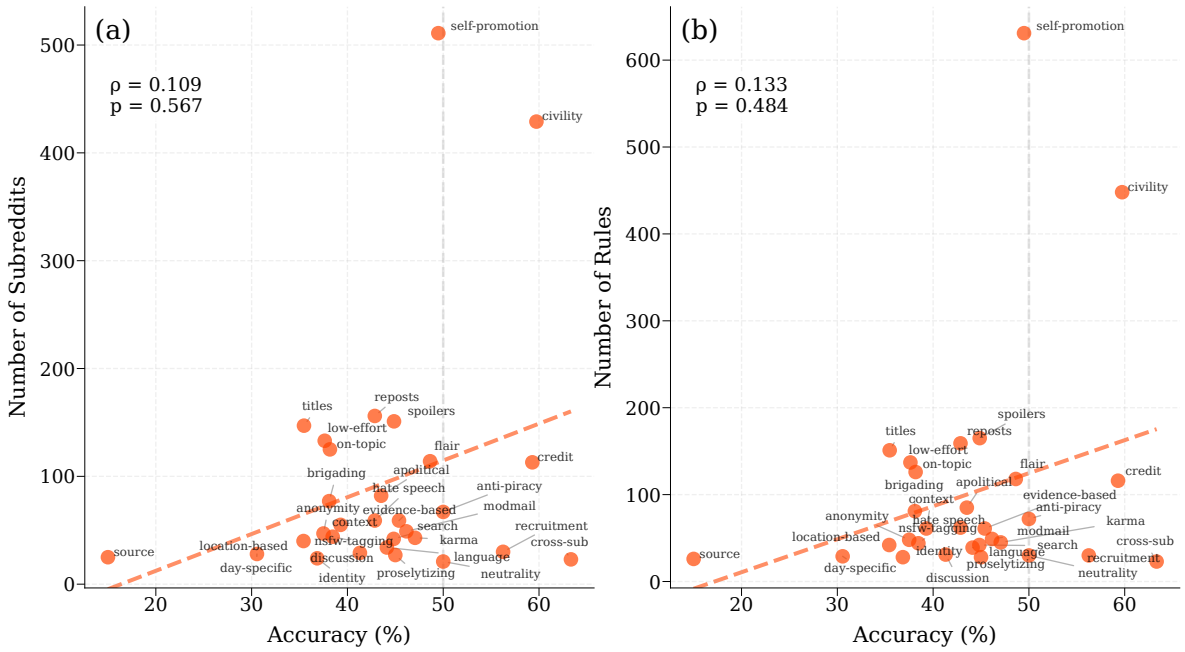

Figure 17: Qwen3-VL-4B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

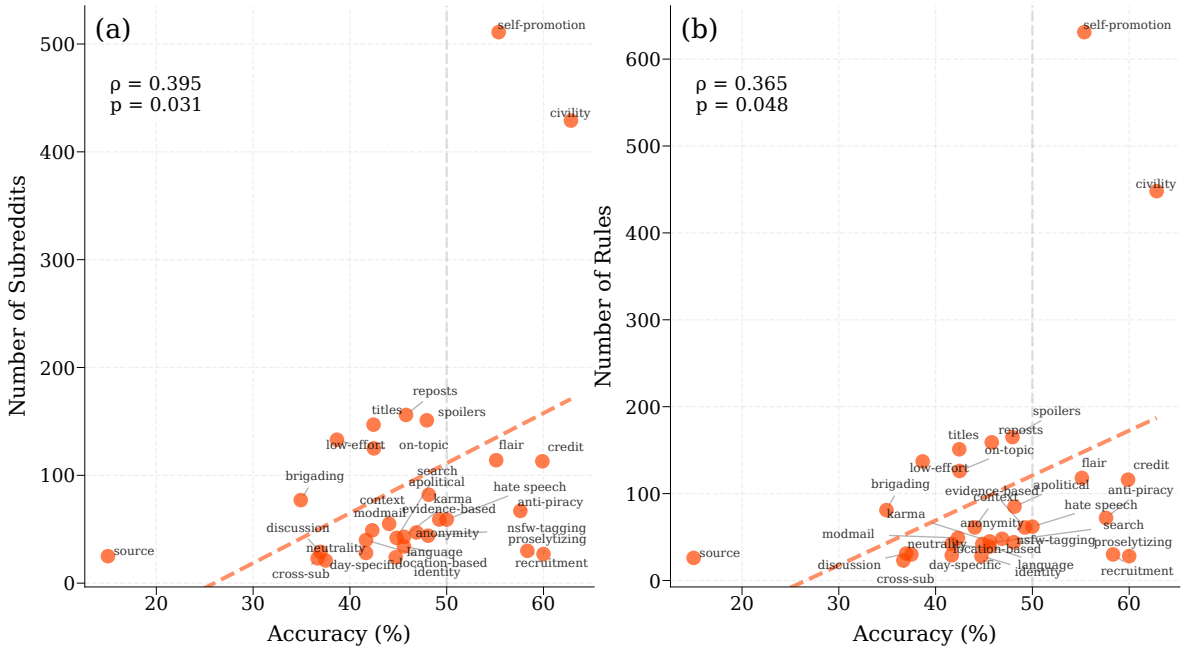

Figure 18: Qwen3-VL-8B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.

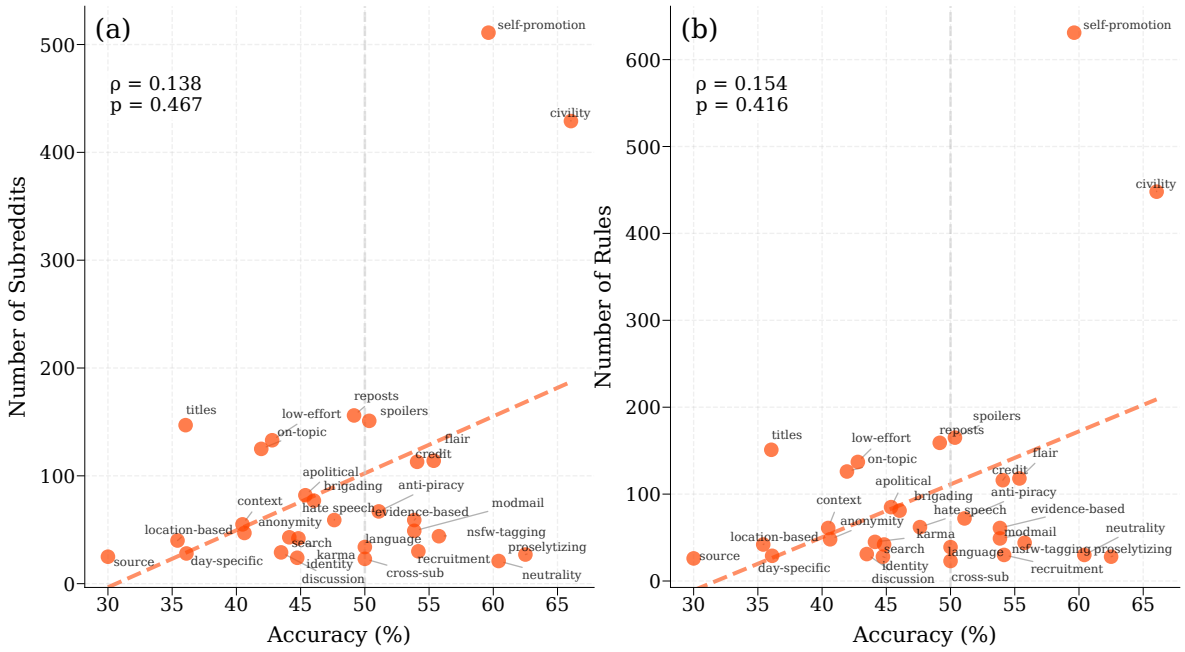

Figure 19: Qwen3-VL-30B (instruct) accuracy with full context compared against (a) the number of subreddits containing rules in each cluster and (b) the number of rules per cluster. Dashed lines indicate the 50% baseline.