Helping Admins Write Grounded Rules for WhatsApp Groups

Sudhamshu M. Hosamane
SETS Fellow, Cornell Tech
PhD Student, Rutgers University, New Brunswick

Background and Motivation

Rules are the quiet infrastructure of online groups: they set expectations, reduce friction, and help communities scale. On public platforms such as Reddit, Discord, and Twitch, visible, collaboratively authored rule sets are linked to lower harassment and toxicity, healthier newcomer participation, and more sustainable moderator workloads. Large comparative studies also show that many thriving communities make their rules explicit and adapt them to local context—often blending *prescriptive* ("do this") and *restrictive* ("don't do that") guidance. A wave of tooling has grown around this ecology, from rule debugging and previews (*ModSandbox*) to executable community procedures (*PolicyKit*) and low-code policy authoring (*Pika*), alongside creator-facing filters and audit.
End-to-end encrypted messaging apps are different. WhatsApp groups rarely surface rules prominently; norms are informal, negotiated through existing offline ties, and enforcement—if it happens—leans on socially costly, one-to-one nudges by volunteer admins. Without clear places to display shared expectations, even well-meaning groups struggle to translate broad values (civility, relevance, privacy) into concrete, situational guidance.

This fellowship investigated that gap through two complementary steps. First, we present a landscape analysis—drawing from large-scale WhatsApp group metadata and a targeted survey of Indian users and admins—to examine how common group rules are, what they typically say, and which problems most often force admins to intervene. These findings inform the design of a lightweight, context-aware assistant for WhatsApp admins that can suggest rules grounded in observed practice. Second, we outline a pre-analysis plan for two experiments: (i) testing whether rules generated with full contextual information are more relevant, adoptable, and enforceable than generic LLM-generated rules, and (ii) comparing three conditions for rule-setting—admins alone, admins assisted by AI, and AI alone—in a simulated WhatsApp group, measuring outcomes such as cognitive burden, relevance for enforcement,

and ease of authoring. Together, these steps aim to reduce moderation workload while making expectations clearer and more actionable for group members.

Analysis of WhatsApp Group Rules

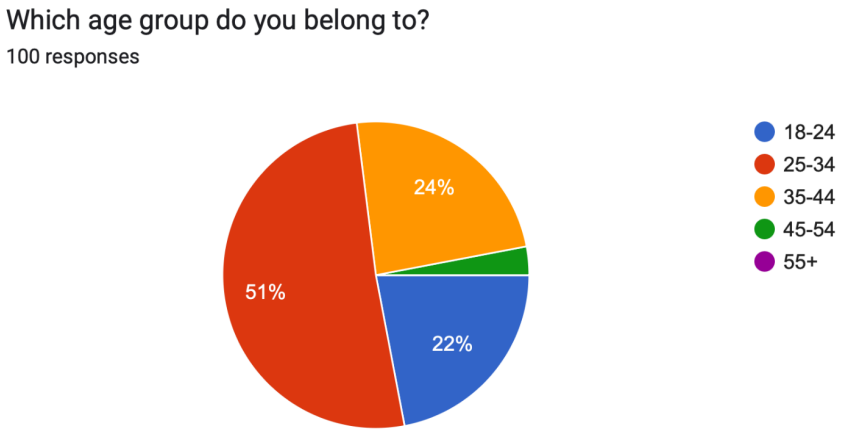
Analyzing donated WhatsApp groups–metadata.

WhatsApp Explorer is a tool to collect WhatsApp group data in a privacy preserving manner. We analyzed the metadata (group name, creation date, description, group size etc.) of 65,984 groups voluntarily donated by participants for various other projects using this tool. Although many of these were used for formal communication and consisted of members without strong interpersonal ties (e.g., friends and family), very few groups use the description field: only **152** (0.2%) groups had a non-empty description, and just **15** of those descriptions actually contain rules. Most descriptions read like bulletin boards rather than charters: the dominant themes observed were logistics—dates/venues/links and a brief statement of purpose—with "about the group" blurbs far outnumbering any governance text. The groups spanned geographies and languages (e.g., U.S. diaspora communities in San Francisco and Houston, Hindi-language study groups in India, East African marketplaces, Spanish language buy/sell chats, etc.), but the pattern holds across contexts.

Where rules do appear, they cluster in communities with clear content boundaries—marketplaces, food/hobby, and student cohorts. These rules skew *restrictive* (e.g., "only post X," "no spam/off-topic," "no adult content"; ≈ 12/15) over *prescriptive* ("please introduce yourself when you join"; ≈2/15). Linguistically, rule-bearing descriptions are both longer and more directive, with more imperative cues (as previously described) compared to other descriptions. This finding suggests that rule-bearing descriptions are not only more verbose but also explicitly signal behavioral boundaries, whereas most descriptions—used for logistics or announcements—contain little to no governance-oriented language.

Survey Insights (India)

To better understand common governance practices in WhatsApp groups, we conducted a short, informal survey with **100** group admins in India, recruited via Facebook Ads. This was not a formal human-subjects study, but rather an exploratory poll to get a broad sense of challenges and rule-setting practices. People reported locations across the country, with clusters around major metros (notably Delhi and Mumbai) and states like West Bengal, plus a long tail of smaller cities and towns. Although over 52% of the people who saw the ads were under 24, admins between 25–44 made up almost 75% of our respondents.



This matches the WhatsApp reality most of us see: group chats are often how young adults coordinate family, work, and community life. Critically, the people answering were mostly the ones who can actually act: 86.7% say they are admins of at least one active group. Formal governance is uneven. Even among this self-selected sample of admins who actively manage groups, fewer than half (48.0%) report having written rules (pinned, starred, or in the description). The rest rely on informal nudges when problems arise (27.6%) or admit to having no rules at all (24.4%).

Among those who shared valid rule text (n = 42), the median rule length is **13** words. Linguistically, rules are *more often restrictive* ("no spam/off-topic," "no politics," "only posts relevant to") than *prescriptive* ("please keep it civil," "be respectful"): Restrictive (22) and Mixed (13) dominate, with few purely Prescriptive (4). In plain terms: most rules draw boundaries rather than set positive norms, which matches how lightweight messaging groups typically manage attention and relevance. We further analyzed whether these rules were written entirely by the admins themselves or if they used the help of AI. We evaluated text samples using two widely used AI-content detection tools, **Quillbot** and **GPTZero**. Both tools identified 8 out of 41 pieces as fully AI-generated and an additional 3 as collaboratively written by humans and AI. The two detectors produced identical classifications in all cases, yielding a 100% agreement rate.

What goes wrong in Indian WhatsApp groups?

Respondents indicated that the top recurring issues in their groups were misinformation (51), spam (44), and conflicts/arguments (32), followed by illegal/unsafe content (16), and privacy breaches (11). These problems show up most in personal spaces (family/friends/neighborhood) and professional groups; learning and hobby groups also feature but less intensely. That mix makes sense: when offline relationships spill into chat, small frictions can escalate, and "forward-as-received" culture fuels low-effort sharing of low-credibility information.

How often do admins intervene?

Among respondents who faced at least one issue (n = 79), intervention was common: 34.6% stepped in 1–2 times over three months, 23.1% about monthly, 14.1% almost weekly, and 7.7% multiple times per week; only 20.5% said never. In other words, even modest-sized groups demand steady, unpaid moderation work—mostly deleting off-topic forwards, cooling arguments, or correcting false claims—burdens we want to alleviate with the help of a rule-setting assistant.

Future Work

Building on these findings, we will run two simple, real-world tests to examine how the assistant can meaningfully help WhatsApp admins set rules.

Experiment 1: Generic vs. Context-Aware Rules

WhatsApp admins of active groups will be invited to evaluate two anonymized sets of rule suggestions. To create these sets, an admin uploads a text-only chat transcript—shared with prior consent from group members—where all personally identifiable information is automatically removed on the client side using established JavaScript libraries. *Set A* contains **eight generic rules** generated solely from the group's type and stated purpose. *Set B* contains **eight context-aware rules** derived from the redacted transcript, incorporating features such as message content, media types, timestamps, topics, and other contextual signals. The two sets are carefully balanced for length and tone, with provenance blinded and near-duplicates merged. Admins are then asked to rank the **top seven** rules according to their perceived relevance, clarity, and ease of implementation.

The trade-off is nontrivial: context-aware rules may capture recurring issues and feel tailored, yet risk overfitting past disputes or amplifying noisy voices. Generic rules are concise, neutral, and privacy-safe, but can miss group-specific pain points.

Measures: Rule selections and rankings by admins, rule adoption rates for each set, stated reasons for rejecting rules (e.g., unclear, hard to enforce), and patterns by group type indicating whether generic or context-aware rules are favored. The aim is to identify when specificity improves usefulness and when a neutral baseline is preferable.

Experiment 2

Another group of participants read short, realistic (mock) WhatsApp threads and generate rule lists under three conditions: **Human-only** (written from scratch), **Human+AI** (edited from AI suggestions), and **AI-only** (produced entirely by prompting the model). A separate set of judges then rates these lists for clarity, contextual fit, enforceability, and legitimacy. The trade-offs are not obvious: human-only rules capture nuance but are effortful; human+AI speeds drafting yet risks anchoring; and AI-only is fast and consistent but may seem generic or less trusted.

Design. Each person sees the same short scenario (spam, misinformation, or conflict). Instructions are the same for all the three arms and time caps are matched across conditions; content is fictional, so no personal data is used.

What we'll measure. Ease of writing and mental effort (7-point scales) for each setting, time to a usable draft, number of edits/ additional prompts, and blinded ratings of the final rules for clarity, enforceability, and coverage of the scenario's problems. We also ask confidence to explain/enforce. The goal of this experiment is to identify which workflow makes rule-setting both easier and better.