Designing and Evaluating a Community Fact-Checking System for WhatsApp

JUAN JOSE ROJAS CONSTAIN, KIRAN GARIMELLA

The proliferation of misinformation on end-to-end encrypted platforms like WhatsApp poses a significant societal challenge, as the privacy architecture precludes traditional, top-down content moderation. While community-driven fact-checking has shown promise on public platforms, its viability within private, high-trust environments remains an untested assumption. This paper introduces and evaluates a community-based fact-checking system designed for WhatsApp, conducting the first field experiment of its kind. We deployed a tool with 38 participants in Colombia, who evaluated real viral content sourced from their own community groups over a two-week period. Our mixed-methods analysis reveals a nuanced picture: while a community can reliably identify true information (78% accuracy), it performs little better than chance at detecting falsehoods (56% accuracy), exposing a critical "pro-truth bias." We further demonstrate that the "wisdom of crowds" is not automatic but can be engineered; weighting votes by demonstrated user expertise allows a small, expert-weighted crowd to achieve high reliability (80%), significantly outperforming a simple majority. This work presents a viable, user-centric model for combating misinformation on encrypted platforms, demonstrating that while community moderation is a promising path, its success hinges on designing socio-technical systems that don't just aggregate all voices, but intelligently identify and amplify expertise.

ACM Reference Format:
Juan Jose Rojas Constain, Kiran Garimella. 2025. Designing and Evaluating a Community Fact-Checking System for WhatsApp. 1, 1 (September 2025), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Encrypted messaging platforms like WhatsApp have become the de facto digital town square for billions of people, particularly in the Global South [12, 26]. They are the primary channels through which news is consumed, communities are organized, and social bonds are maintained [19]. However, this central role comes with a significant societal cost: these platforms have also become potent vectors for the rapid, unchecked spread of misinformation, particularly with offline consequences such as riots and lynchings [3, 4]. The very features that make them invaluable for private communication—end-to-end encryption, small group dynamics, and high-trust social networks—create a fertile ground for falsehoods to flourish, with profound consequences for public health, political stability, and social cohesion [15, 16].

Addressing this problem is uniquely challenging because naive approaches are destined to fail. The intractability is not merely a matter of content but is deeply rooted in WhatsApp's core architecture and its dominant patterns of social use. At least three factors create a socio-technical impasse for traditional content moderation: (i) as a fundamental privacy feature, end-to-end encryption renders centralized, platform-led content review technically impossible, shifting the entire moderation burden onto users and volunteer group administrators; (ii) unlike the interest-based networks of strangers on public platforms like X (formerly Twitter) or Reddit, WhatsApp communication occurs within tight-knit groups of family and friends, a high-trust context that systematically lowers users' cognitive defenses; and (iii) a

Author's Contact Information: Juan Jose Rojas Constain, Kiran Garimella.

Manuscript submitted to ACM

significant portion of misinformation on WhatsApp is not text-based but consists of images, videos, and audio notes with highly local context, which is notoriously difficult for centralized fact-checkers to moderate at scale [10].

Existing interventions are ill-suited to this complex environment. Top-down, professional fact-checking tiplines are reactive and face insurmountable scalability challenges [14]. A more promising paradigm is bottom-up, community-based moderation, operationalized by the "wisdom of crowds" principle that powers systems like X's Community Notes [24]. Foundational research has shown that aggregating the judgments of a small, politically balanced crowd of laypeople can be as accurate as professional fact-checkers [2, 11, 17]. However, this foundational evidence was established under laboratory conditions that are fundamentally disconnected from the reality of WhatsApp. These studies typically use anonymous crowd-workers from Western countries, artificially balance them for political diversity, and ask them to rate generic news headlines. WhatsApp groups, in contrast, are often ideologically homogenous, high-context, and composed of individuals with deep-seated social relationships, where publicly correcting a peer carries significant social friction and interpersonal risk [7, 20, 23]. It remains a critical, unanswered question whether the "wisdom of crowds" holds true in these ecologically valid but socially "messy" environments where the carefully controlled conditions of the lab do not apply.

This paper presents the first field experiment to empirically test the feasibility and accuracy of a crowdsourced fact-checking system on WhatsApp. We developed and deployed a WhatsApp bot with 38 participants in Colombia, asking them to evaluate real, viral content sourced from their own community WhatsApp groups over a two-week period. Through this mixed-methods study, combining quantitative analysis of participants' daily ratings with follow-up qualitative interviews, we investigate the entire lifecycle of a community-based moderation system.

Our mixed-methods approach reveals a nuanced picture of the promises and perils of this moderation paradigm, with three key findings.

- First, we identify a strong pro-truth bias that limits crowd reliability: participants were effective at identifying true content (78% accuracy) but performed little better than chance at identifying falsehoods (56% accuracy), indicating that a community can ratify truth but is highly vulnerable to believing misinformation.
- Second, we find that the "wisdom of crowds" is not an automatic outcome but must be engineered: a simple majority vote offers only modest gains, but weighting contributions by demonstrated expertise allows a small, "smarter" crowd to achieve high accuracy (around 80%), outperforming much larger, unweighted groups.
- Third, our qualitative analysis reveals that expertise, not a specific evaluation strategy, is the key predictor of accuracy, reinforcing the quantitative finding that the central design challenge is not simply to aggregate all opinions, but to identify and amplify the most reliable voices within a community.

By moving the study of crowdsourcing from the lab to the wild, our work provides a crucial, empirically grounded data point for the HCI discourse on governing encrypted spaces. We demonstrate that community-based fact-checking is a promising but not magical solution. Its success hinges on designing socio-technical systems that can navigate inherent cognitive biases and intelligently leverage the distributed expertise that exists within any community.

2 Related Work

The proliferation of misinformation on encrypted messaging platforms like WhatsApp presents a formidable challenge for content moderation [27]. The privacy-preserving nature of end-to-end encryption renders centralized, platform-led interventions technically infeasible, shifting the responsibility of governance onto users themselves. This paradigm shift necessitates a deep understanding of community-driven moderation, a topic with a rich history in the HCI and CSCW

communities. This section situates our work within this literature. We first examine established models of community governance on public online platforms to understand their underlying principles and assumptions. We then critically analyze the specific mechanism of crowdsourced fact-checking, focusing on the conditions under which it succeeds or fails. Finally, we synthesize research that characterizes the unique socio-technical fabric of WhatsApp, arguing that this context fundamentally challenges the assumptions of existing models and reveals a critical gap that our study addresses.

2.1 Models of Community-Based Governance in Online Spaces

Decentralized, user-driven governance is a cornerstone of many successful online platforms. The CSCW literature has long explored these systems, revealing a spectrum of models predicated on different forms of social organization and volunteer labor. At one end lies distributed, volunteer-led moderation, exemplified by platforms like Wikipedia and Reddit. Wikipedia stands as a monumental achievement in collaborative content creation, governed by a complex bureaucracy of volunteer editors and administrators who enforce a vast set of community-ratified policies [18]. Similarly, Reddit is composed of thousands of distinct, topic-based communities ('subreddits'), each managed by a team of volunteer moderators who establish and enforce highly contextualized local norms [5].

While effective at scaling governance, these models are sustained by immense amounts of often invisible and psychologically taxing labor [25]. Moderators frequently face burnout from the sheer volume of content and exposure to harmful material [21]. A critical insight from this research is that these governance structures are designed for communities of strangers, bound not by pre-existing social ties, but by a shared interest in a topic or a commitment to a collective project.

A different approach is seen in platform-intermediated community moderation. On Twitch, for instance, moderation is a hybrid model where the platform sets baseline rules, but individual streamers and their appointed moderators perform a visible, real-time 'moderation performance' in their specific channels [22]. This act is deeply social, often motivated by loyalty to the streamer and a desire to cultivate a particular community atmosphere. More recently, X/Twitter's Community Notes represents a shift towards a crowdsourced annotation model, empowering a broad base of users to add contextual notes to potentially misleading posts, rather than relying on a small cadre of moderators. The design of Community Notes explicitly attempts to surface consensus from contributors with diverse viewpoints. Across these varied models, the underlying context remains one of public or semi-public interaction, where governance mechanisms, whether formal rules or real-time interventions, are designed for interactions between relative strangers.

2.2 The "Wisdom of Crowds" as a Fact-Checking Mechanism

Our study focuses on crowdsourced fact-checking, a mechanism that operationalizes the "wisdom of crowds" phenomenon, which posits that aggregating the judgments of a diverse group of non-experts can yield remarkably accurate results. The application of this principle to misinformation has strong empirical support. A seminal study by Allen et al. demonstrated that the average accuracy rating of a small, politically balanced crowd of laypeople correlated as well with the judgments of professional fact-checkers as the fact-checkers did with each other [2]. This finding suggests that crowdsourcing is a highly promising and scalable approach to identifying false content, forming the theoretical bedrock for systems like Community Notes.

However, the real-world effectiveness of this mechanism is far from settled. Empirical evaluations of Community Notes have produced mixed results. While some research suggests that notes can reduce user engagement with and diffusion of false information [24], other large-scale studies found no significant evidence that their introduction

reduced engagement with misleading tweets [6]. The success of crowdsourcing appears to be highly contingent on its implementation and the context in which it is deployed.

A critical, and often overlooked, condition in the foundational work of Allen et al. [2] was the deliberate construction of politically heterogeneous crowds. By balancing groups with equal numbers of Democrats and Republicans, their methodology could effectively neutralize partisan bias through aggregation. This crucial design choice, however, highlights a significant limitation: the "wisdom of crowds" may be critically dependent on the social composition of the crowd itself. This raises the question of how such a mechanism would function in environments where the crowd is not artificially balanced, but is instead naturally homogenous.

2.3 Community Moderation on WhatsApp

The governance models and fact-checking mechanisms discussed above were designed for and studied on public-facing platforms. Their applicability to EMPs like WhatsApp is questionable, as WhatsApp is defined by a fundamentally different set of socio-technical properties. Research in HCI has begun to map this unique terrain, revealing a context that is private, relational, and high-context.

Unlike the communities of strangers on Reddit or X, WhatsApp groups are typically small, private circles composed of individuals with strong, pre-existing offline relationships, such as family, colleagues, or close friends. Within this setting, prior work has established that content moderation is not a formal, rule-based process but a delicate act of social maintenance. As Shahid et al. [23] compellingly argue, group administrators must constantly navigate a tension between care for their members' feelings and control over the conversation, adopting roles analogous to "parenting styles" (e.g., authoritative, permissive). An admin's authority is not derived from a platform policy, but is negotiated through their offline social standing, making direct confrontation a socially costly action.

Furthermore, observational studies of public WhatsApp groups confirm that explicit moderation is exceedingly rare [7]. When faced with problematic content, users prefer non-confrontational strategies or silence. The social fabric of these groups is often characterized by ideological homophily, making them fertile ground for echo chambers. Attempts to design interventions, such as conversational agents to facilitate deliberation, have surfaced deep-seated tensions between the desire for privacy, the preservation of group harmony, and the value of free expression [1].

2.4 Our Contributions

The literature, therefore, reveals a critical disconnect. On one hand, the "wisdom of crowds" has been validated as a powerful fact-checking mechanism under controlled, laboratory conditions that assume politically diverse groups of anonymous strangers [2]. On the other hand, the primary environment where encrypted misinformation spreads—private WhatsApp groups—is characterized by pre-existing social ties, relational complexities, and ideological homogeneity.

It remains a significant open question whether the core principles of crowdsourced fact-checking hold in this ecologically valid but "messy" context. Our work is the first to directly address this gap by empirically evaluating a community fact-checking model within this unique environment. We make the following contributions to the HCI literature:

- We conduct the first field study to test the accuracy of crowdsourced fact-checking within existing, socially embedded WhatsApp groups, moving the unit of analysis from anonymous crowd workers to members of a real community.

- We use ecologically valid content sourced from participants' own groups, providing a more realistic measure of performance than the standard headline-rating tasks.
- By analyzing the performance of these naturally formed, often homogenous groups, we empirically test the boundary conditions of the "wisdom of crowds" theory, providing critical insights for designing future moderation systems for high-context, relational online spaces.

3 Methodology

To test the viability of crowdsourced fact-checking on WhatsApp, we conducted a two-week field experiment with a mixed-methods design. The study involved a quantitative data collection phase where participants rated content daily, followed by a qualitative phase of semi-structured interviews to understand the reasoning behind their judgments.

3.1 Study Setting and Participant Recruitment

The study was conducted in Colombia, a setting where WhatsApp is the primary mode of digital communication for social, professional, and educational life. This deep integration provided an ecologically valid environment for our research. Participants were recruited between July 21 and August 1, 2025, using a snowball sampling strategy starting with personal network of one of the authors to achieve a heterogeneous sample of 38 users, with a concentration in the Valle del Cauca and Cesar regions (East and North of Colombia). A detailed demographic breakdown of the final participant group is available in Table 1 in the Appendix.

Each participant was onboarded via a short video call where they provided informed consent. They received a flat incentive of 20,000 COP (around 5 USD) for registering, with an additional performance-based incentive of 40,000-50,000 COP (10-12.5 USD) upon completion of the two-week study (August 11-22, 2025).

3.2 Content Corpus

Once the participants provided consent, we used WhatsApp Explorer [8] to enable users to donate the data from the WhatsApp groups they were a part of. WhatsApp Explorer allows customization to only download content marked 'forwarded many times' which lets us only download content which was virally spreading on WhatsApp [9]. Overall, the 38 users donated 243 unique groups and we collected all 807 pieces of content marked as viral shared in these groups during a two month window (July-August 2025).

The first author (native Spanish speaker from Colombia and a member of the community where the data was collected) manually went through all the viral content and identified misinformation which was fact checked by professional fact checkers as false. From the hand labeled content, we curated a corpus of 40 unique content items for the study. To ensure ecological validity, all items were sourced from material that had been circulating in the participants' own WhatsApp groups. The research team carefully screened this content to ensure it was fact-checkable and had a clear binary (true or false) verdict. The final set was diverse, covering national and regional news, health, and politics (with both left- and right-leaning claims), and was presented in various formats (text, images, videos) to mirror a realistic information diet.

3.3 Study Data Collection

Our data collection protocol contained two parts:

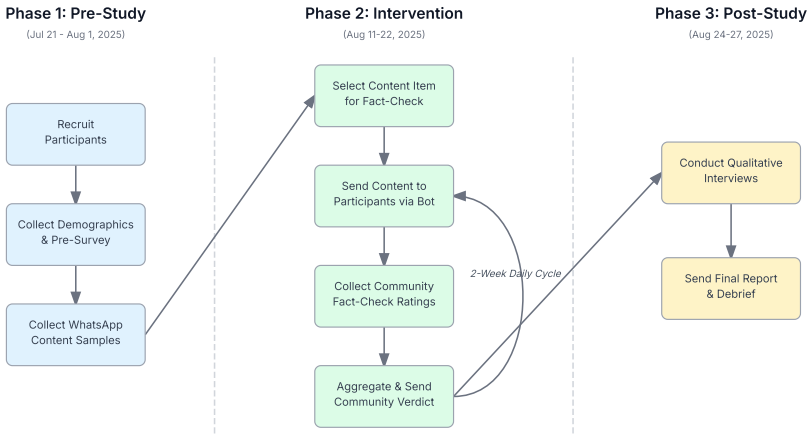

Fig. 1. A flowchart of the study protocol, detailing the Pre-Study (recruitment and content collection), Intervention (two-week daily fact-checking cycle), and Post-Study (qualitative interviews and debriefing) phases.

- **Quantitative Data Collection**. (Two-Week Intervention). Over two consecutive weeks (Monday-Friday, from Aug 11-22), participants received at least two content items per day via a WhatsApp-based research bot our team created. For each item, they were asked to make a simple binary classification: *"Please look carefully at the following video/image/text. How would you classify the content? A – True. B – False."* Participants submitted their responses using embedded buttons. After responding, they received feedback on how the rest of the "community" (i.e., other study participants) had voted on that item. The message which was sent after the person responded read: *"Here's how your community responded to this message: 94% → True 6% → False. Thank you for participating!"* For each piece of the 40 pieces of content, we made sure that at least 20 participants were sent the question.
- **Qualitative Follow-up Interviews**. Following the intervention, we conducted semi-structured remote interviews to provide rich, qualitative context for our quantitative findings. We randomly invited 20 participants, stratified by performance, and ultimately conducted 20-30 minute Zoom sessions with nine individuals ($N = 9$): three high-performers (accuracy > 0.7), three mid-performers (0.5–0.7), and three low-performers (< 0.5). The interviews explored four key areas: (1) motivations for participation, (2) heuristics used for content evaluation, (3) reactions to the community feedback loop, and (4) usability feedback and suggestions for a real-world version of the tool.

We nudged participants who were not responsive by contacting them personally. Overall, the study achieved a response rate of over 90%. At the conclusion of the study, all participants received a debriefing report clarifying the ground truth for all items. The entire data collection protocol, from initial recruitment to the final debriefing, is visually summarized in Figure 1. All the data collected during our study, including the stimuli used, demographics of the participants, participant responses, and interview transcripts are available for review.[1]

[1]Anonymized link: https://drive.google.com/drive/folders/1tY4fykTsFqZaTPXuVt1Gv4GZj4PZpT6A?usp=drive_link

4 Results

In this section, we present the results of our study on crowdsourced fact-checking. We begin by establishing the baseline for individual performance, analyzing the overall accuracy and identifying key biases in how participants evaluate true versus false information. Next, we examine dynamic factors by investigating how accuracy evolves over the course of the task to assess for learning or fatigue effects. Finally, we move from individual to collective performance, simulating the "wisdom of crowds" to determine how the accuracy of a group judgment scales with crowd size and the method used to aggregate opinions.

4.1 Overall Fact-Checking Accuracy

First, we analyzed the overall accuracy of individual ratings across all participants and content items. As shown in Figure 2, the mean accuracy was 0.66. However, this overall figure masks a significant disparity in performance based on the ground truth of the content.

The most striking finding is a strong "pro-truth" bias in participant judgments. Participants were highly effective at identifying true content, achieving an accuracy of 0.78. Conversely, their ability to identify false content was significantly lower, with an accuracy of only 0.56 (i.e. , 44% of the time, a false item was rated as true). This suggests that while participants can reliably recognize valid information, they are far more susceptible to believing misinformation.

Our interviews illuminate *why* this asymmetry exists. Participants did not employ the rigorous, multi-step verification processes of professional fact-checkers [13]. Instead, they relied on a range of "good enough" heuristics that, while efficient, were ill-suited for detecting falsehoods. These strategies fell into two broad categories. Some participants relied on quick, informal external validation. This included using Google for unfamiliar topics (`P03`, `P05`, `P08`) or engaging in social consultation by discussing content with family or colleagues (`P01`, `P03`, `P08`). Others relied on internal, intuitive heuristics, making judgments based on a "hunch" (`P02`) or their pre-existing political knowledge (`P07`, `P09`). Some even developed specific mental shortcuts, such as an "anti-propaganda" heuristic that automatically flagged content as false if it appeared overtly political (`P04`), or using production quality as a proxy for credibility (`P05`).

Crucially, no single strategy guaranteed high accuracy. The reliance on these rapid, surface-level heuristics helps explain the poor performance on false content. While such shortcuts may be effective for recognizing legitimate information that aligns with one's existing knowledge and worldview, they lack the critical depth required to identify sophisticated or subtly fabricated falsehoods, thus explaining the significant performance gap observed in our quantitative data.

Further analysis across content modality (Image, Video, Text), topic (Political, Health) (see Figure 2), or across various demographic variables (like age, gender, education level, political leaning, and CRT) revealed no statistically significant differences in accuracy (see Figure 5 in the Appendix).

4.2 Performance Over Time: Investigating Learning and Fatigue

To investigate whether participant accuracy changed over the duration of the study, we analyzed performance for potential learning or fatigue effects. We divided the 20 content items rated by each participant into four sequential blocks of five questions and calculated the mean accuracy for each block across all participants.

As shown in Figure 3, we found no evidence of a learning effect. Mean accuracy was stable at 0.71 for the first two blocks (questions 1-10). In the second half of the task, performance slightly declined to a mean accuracy of 0.60 in the third block and 0.63 in the fourth. This trend suggests that participants did not improve with practice and feedback

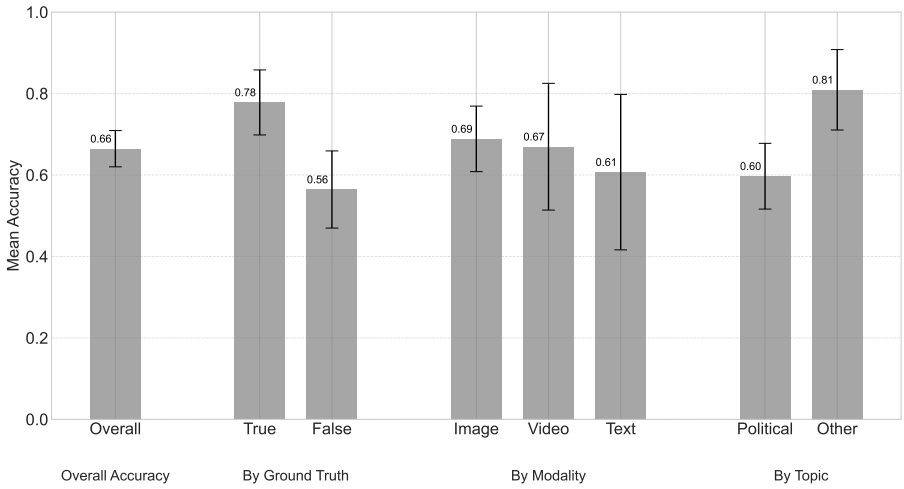

Fig. 2. Mean fact-checking accuracy broken down by ground truth, content modality, and topic. Error bars represent 95% confidence intervals. Participants were significantly more accurate at identifying true content than false content.

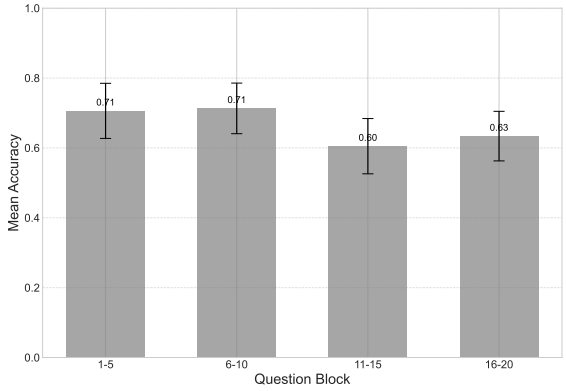

Fig. 3. Mean accuracy across four sequential blocks of questions. Error bars represent 95% confidence intervals. The results show no evidence of learning, with a slight decline in accuracy in the second half of the task.

about their community performance; rather, the slight decrease (not statistically significant) in accuracy indicates no learning (or onset of fatigue) as they progressed through the task.

The qualitative data reveals the mechanism behind this stability: a consistent, multifactorial motivation structure. While the financial incentive was a primary and candidly stated driver for continued participation by over half of the participants, it was consistently coupled with intrinsic factors that fostered genuine engagement. The key factor was the ecological validity of the content. Because we used content sourced from their own communities, participants found it highly relevant. They noted it reflected what was happening "in day-to-day life" (P05) and showed "what is really happening in the community" (P03). This connection to their lived experience transformed the task from a simple chore into a compelling activity. As one high-performing participant (P01) noted, the local relevance of the material was a

key factor in keeping him engaged. This consistent blend of extrinsic reward and intrinsic interest likely explains the lack of performance degradation over the two-week period, as participants' engagement was sustained by more than just the financial reward.

4.3 Wisdom of Crowds: Collective Accuracy in Fact-Checking

To evaluate the "wisdom of crowds" for fact-checking, we simulated how collective accuracy evolves with increasing group size and different methods of aggregating judgments. We simulated crowds of sizes 1 to 38 by sampling participants with replacement over 100 runs. For each simulated crowd, we calculated the majority verdict on 40 content items and compared three aggregation methods: (1) an unweighted majority vote, (2) votes weighted by each user's historical accuracy, and (3) votes weighted by their Cognitive Reflection Test (CRT) score.

The results, presented in Figure 4, show that collective accuracy improves as crowd size increases for all aggregation methods. The sharpest gains occur with small crowds (up to 10 participants), after which the improvements demonstrate diminishing returns.

Weighting votes by expertise significantly enhances performance. The accuracy-weighted method is the top performer, consistently achieving higher accuracy than the unweighted baseline across all crowd sizes. Notably, an accuracy-weighted crowd of just seven participants achieves an accuracy (0.72) comparable to that of a much larger unweighted crowd of 15-20 participants. This expert-weighted method reaches a peak accuracy of approximately 0.80, a substantial improvement over the baseline's plateau of 0.75. Weighting by CRT score provides a marginal benefit over the unweighted method for smaller crowds but converges with the baseline as the crowd size grows.

The qualitative interviews provide a compelling behavioral explanation for the dramatic success of this accuracy-weighted model. A participant's performance was strongly correlated with their confidence and their reaction to the community feedback loop. High-performing participants (`P01`, `P09`) exhibited high confidence and were largely unswayed by dissenting community feedback. When their answer differed from the majority, they did not question their own judgment. Instead, they rationalized the disagreement by questioning the community's knowledge or political bias. As `P09` concluded, the community was simply "more conservative" and did not represent his viewpoint. In essence, they acted as stable, confident anchors whose correct judgments were not perturbed by social influence.

Conversely, mid- and low-performing participants were more likely to interpret disagreement as a signal of their own error, prompting self-doubt and re-evaluation. They described feeling like they had "lost a game" (`P04`) or were motivated to "analyze the next item more carefully" (`P06`).

This behavioral divergence is precisely why accuracy-weighting works so effectively. The algorithm algorithmically achieves what the social dynamics could not: it amplifies the signal from the confident, stable, and correct "anchors" while reducing the influence of less accurate participants who are more easily swayed by the (potentially incorrect) majority. The weighting system leverages the underlying psychological trait of confidence that correlates with performance.

5 Discussion

Our study was designed to test the viability of crowdsourced fact-checking in one of the most challenging yet critical environments for misinformation intervention: the private, high-context, and socially-embedded groups of WhatsApp. Our mixed-methods results present a nuanced, and at times contradictory, picture. While the "wisdom of crowds" does not fail outright, its effectiveness is deeply contingent on the social context and the method of aggregation. In this section, we discuss the implications of our key findings, situating them within the HCI literature. We first explore why the crowd's performance deviated from foundational studies, then discuss the profound implications of identifying

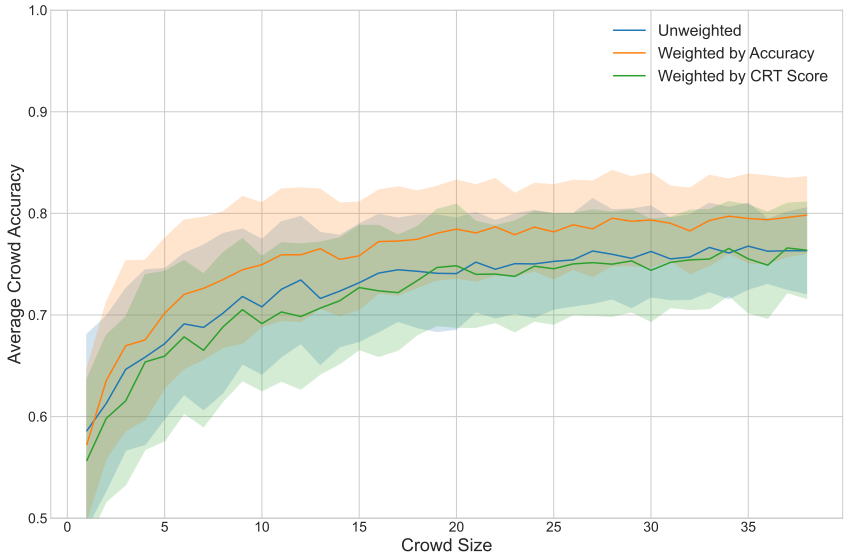

Fig. 4. The relationship between crowd size and average fact-checking accuracy for three aggregation methods. Shaded regions represent 95% confidence intervals across 100 simulations. Weighting votes by historical accuracy consistently yields the highest collective accuracy.

"expert" performers within these small communities, and finally, translate these insights into actionable design principles for future socio-technical systems in encrypted spaces.

5.1 The Fragile Wisdom of Homogeneous Crowds

A central finding of our study is that an untrained, socially-embedded crowd performs significantly worse than the crowds in foundational lab-based studies, particularly when identifying false content (56% accuracy). This stands in stark contrast to the high performance reported by Allen et al. [2], whose work underpins systems like Community Notes on Twitter. Our findings do not refute the "wisdom of crowds" principle but rather reveal its critical boundary conditions. The key difference is context. The success of prior work hinged on using anonymous, artificially constructed, and politically balanced crowds, where partisan biases could be statistically canceled out.

Our field experiment, by design, embraced the "messiness" of a real-world context. Participants evaluated ecologically valid content sourced from their own often ideologically homogeneous groups. In such an environment, the "wisdom of crowds" is fragile. Rather than averaging out bias, a simple majority vote risks amplifying the dominant viewpoint of an echo chamber, potentially validating rather than correcting misinformation. The strong "pro-truth" bias we observed (78% accuracy on true content) further complicates this picture. It suggests that participants' heuristics are optimized for confirmation of what they already believe to be plausible, but are poorly equipped for the more cognitively demanding task of identifying sophisticated falsehoods. This presents a significant practical challenge: a simplistic deployment of crowdsourcing in these environments could backfire, lending a false sense of community consensus to misinformation.

5.2 Utilizing the Potential of Confident Anchors

While the baseline performance was modest, our simulation of an accuracy-weighted crowd provides a powerful counter-narrative. The finding that a small, expert-weighted crowd of just seven participants could outperform a much larger unweighted crowd is the most promising result of our study. The qualitative interviews reveal the psychological mechanism behind this success: a strong correlation between performance, confidence, and resilience to social influence.

Our high-performing participants acted as confident "anchors." They trusted their judgments and were not swayed when the community disagreed, instead attributing the discrepancy to the community's own biases. Conversely, lower-performing participants were more susceptible to self-doubt, interpreting disagreement as a sign of their own error. The accuracy-weighting algorithm succeeds precisely because it identifies and amplifies the signal from these stable, reliable anchors while down-weighting the noise from less confident, less accurate members.

This has profound implications for design. It suggests that the primary challenge for community fact-checking on WhatsApp is not merely one of aggregating opinions, but of identifying and elevating expertise within a trusted social network. A system that simply reports a majority vote is flawed. A more sophisticated system would focus on identifying these "expert" peers and leveraging their insights, a direction we explore in our design implications. This also explains the lack of a social learning effect over time; the feedback from an anonymous "community" was not salient enough to override the strong internal heuristics of the high-performers or provide clear corrective guidance to the low-performers.

5.3 Designing for Trust and Purpose in High-Context Spaces

Moving from our prototype to a real-world, scalable system requires addressing the significant social and psychological barriers identified by our participants. Their insights form the basis for three core design principles for future interventions in encrypted, high-context environments like WhatsApp.

(1) Design for Trust and Legitimacy, Not Just Usability. While the bot interface was rated as highly usable by our participants, the primary barrier to adoption is social, not technical. Participants universally cited the fear of scams and distrust of unknown or politically biased actors. Any practical system must be deployed by a highly reputable entity (e.g., as `P01` says, "WhatsApp itself" or a trusted non-profit (`P07`)) to be seen as legitimate. Design must prioritize transparency and user control to build this trust.

(2) Frame Participation as Purposeful Contribution, Not Free Labor. For users to remain engaged without financial compensation, the system must offer a clear sense of purpose. As one participant noted, she would not contribute to a system that just collects opinions—"words that the wind carries away." (`P03`). The system's output must be tangible and action-oriented. For example, instead of just a verdict, the system could provide users with well-reasoned, contextual "notes" from high-performing peers that they can easily share within their groups, empowering them to act on the collective judgment.

(3) Create Mechanisms for Peer-Based Expertise Curation. The success of accuracy-weighting points to a system that moves beyond simple democracy. A future system could allow users to create a private, trusted "fact-checking circle" by nominating peers whose judgment they respect. The system could then privately weigh the opinions of these nominated "experts" more heavily when providing a verdict to the user. This approach leverages existing social trust to bootstrap an expert network, respecting privacy while still elevating more reliable voices. It transforms the "wisdom of crowds" into the "wisdom of *my* trusted crowd," a model far better suited to the relational dynamics of WhatsApp [1].

5.4 Limitations

Our study has several limitations that bound the interpretation of our findings. First, our content selection, while ecologically valid, was intentionally curated to have clear, binary true-or-false verdicts. Real-world misinformation is often more ambiguous, including content that is misleading, out-of-context, or satirical. Our task design simplified this complexity, and the accuracy scores may not reflect performance on more nuanced content.

Second, our participants were aware they were part of a study and were financially compensated, which may have influenced their behavior. They may have been more diligent or analytical than they would be in their everyday use of WhatsApp, and their motivations do not fully reflect those of unpaid volunteers. Finally, our study was conducted in a specific cultural context and with a limited sample size, which means the findings may not be directly generalizable to other populations without further research.

5.5 Future Work

Our study opens several critical avenues for future research. First, there is a pressing need to explore privacy-preserving mechanisms for identifying expert users. While our simulation demonstrates the power of accuracy-weighting, building a real-world system requires methods to infer user reliability without compromising privacy or requiring extensive tracking. Future work could investigate lightweight proxies for expertise or federated learning models that allow for reputation scoring without centralizing data.

Second, a key area for exploration is the design of more effective feedback and learning mechanisms. Our simple community verdict feedback did not lead to user learning and was sometimes dismissed by high-confidence users. Future systems could experiment with providing more detailed rationales, links to external fact-checks, or even "expert" scores to see if these interventions can foster self-correction and improve the overall accuracy of the ecosystem over time.

Finally, long-term sustainability requires moving beyond the paid, two-week model of our study. Future research should investigate models for intrinsic motivation and long-term volunteer engagement. Drawing on our qualitative findings, this involves designing systems that are not only trustworthy but also instill a strong sense of purpose and collective efficacy, transforming the act of fact-checking from a transactional task into a meaningful form of community stewardship.

6 Conclusion

This paper presented the first field experiment to empirically evaluate the feasibility of crowdsourced fact-checking on WhatsApp, a private, high-context environment where traditional content moderation is infeasible. By deploying a bespoke fact-checking tool within existing community groups in Colombia and using ecologically valid content, we moved the study of the “wisdom of crowds” from the lab to the wild.

Our work contributes a crucial, empirically grounded data point to the HCI and CSCW discourse on governing encrypted spaces. We show that while community-based fact-checking is a promising direction, its success is not a given. It is a socio-technical challenge that requires careful design: systems must be architected to overcome inherent cognitive biases and to intelligently leverage the distributed expertise that exists within any community.

References

[1] Dhruv Agarwal, Farhana Shahid, and Aditya Vashistha. 2024. Conversational agents to facilitate deliberation on harmful content in whatsapp groups. *Proceedings of the ACM on human-computer interaction* 8, CSCW2 (2024), 1–32.

[2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.

[3] Chinmayi Arun. 2019. On WhatsApp, rumours, lynchings, and the Indian Government. *Economic & Political Weekly* 54, 6 (2019).

[4] Shakuntala Banaji, Ramnath Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. (2019).

[5] 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 (2018), 1–25.

[6] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024. Did the roll-out of community notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on human-computer interaction* 8, CSCW2 (2024), 1–52.

[7] Kiran Garimella. 2025. Community-Driven Fact-Checking on WhatsApp: Who Fact-Checks Whom, Why, and With What Effect? *Computational Approaches to Content Moderation and Platform Governance workshop at ICWSM* (2025).

[8] Kiran Garimella and Simon Chauchard. 2024. WhatsApp explorer: A data donation tool to facilitate research on WhatsApp. *Mobile Media & Communication* (2024), 20501579251326809.

[9] Kiran Garimella, Princessa Cintaqia, Juan José Rojas-Constain, Bharat Kumar Nayak, and Aditya Vashistha. 2025. Global Patterns of Viral Content on WhatsApp. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 586–601.

[10] Kiran Garimella and Dean Eckles. 2020. Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review* (2020).

[11] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety* 1, 1 (2021).

[12] Jacob Gursky, Martin J Riedl, Katie Joseff, and Samuel Woolley. 2022. Chat apps and cascade logic: A multi-platform perspective on India, Mexico, and the United States. *Social Media+ Society* 8, 2 (2022), 20563051221094773.

[13] Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–36.

[14] Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A Hale. 2022. Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp. *Harvard Kennedy School Misinformation Review* 3, 1 (2022).

[15] Neta Kligler-Vilenchik. 2022. Collective social correction: addressing misinformation through group practices of information verification on WhatsApp. *Digital Journalism* 10, 2 (2022), 300–318.

[16] Pranav Malhotra. 2024. Misinformation in WhatsApp family groups: Generational perceptions and correction considerations in a meso-news space. *Digital journalism* 12, 5 (2024), 594–612.

[17] Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G Rand. 2024. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science* 19, 2 (2024), 477–488.

[18] Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[19] Nic Newman, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Digital news report 2025.* Reuters Institute for the study of Journalism.

[20] Sheryl Wei Ting Ng and Taberez Ahmed Neyazi. 2022. Self-and social corrections on instant messaging platforms. *International Journal of Communication* 17 (2022), 21.

[21] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* 26, 10 (2024), 5677–5701.

[22] Joseph Seering and Sanjay R Kairam. 2023. Who moderates on Twitch and what do they do? Quantifying practices in community moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (2023), 1–18.

[23] Farhana Shahid, Dhruv Agarwal, and Aditya Vashistha. 2025. One Style Does Not Regulate All: Moderation Practices in Public and Private WhatsApp Groups. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30.

[24] Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. 2025. Community notes moderate engagement with and diffusion of false information online. *arXiv preprint arXiv:2502.13322* (2025).

[25] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[26] Inga K Trauthig, Katlyn Glover, Zelly Martin, Anastasia Goodwin, and Samuel Woolley. 2023. Polarized Information Ecosystems and Encrypted Messaging Upps. *Center for Media Engagement at the University of Texas* (2023).

[27] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. 2022. Accost, accede, or amplify: attitudes towards COVID-19 misinformation on WhatsApp in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

A User sample details

The complete demographic details of our sample are shown in Table 1.

Table 1. Demographic information of participants

ID	Gender	Age	Pol. Leaning	Yrs of Edu.	Yrs since Last Study	Internet Use (1–5)	Religiosity	CRT
P01	Female	52	Right	10	No info	5	Yes	0
P02	Female	21	Left	9	1–5 years	4	Yes	2
P03	Female	47	Right	11	> 10 years	5	Yes	2
P04	Female	21	Right	18	< 1 year	5	No	2
P05	Male	26	Left	19	< 1 year	4	No	2
P06	Male	21	Right	12	< 1 year	5	Prefer not answering	2
P07	Female	27	Center	17	< 1 year	5	Yes	1
P08	Female	23	Left	15	< 1 year	4	Yes	2
P09	Male	27	Left	21	< 1 year	2	No	3
P10	Male	57	Left	15	> 10 years	3	Yes	2
P11	Male	18	Left	13	< 1 year	5	Yes	1
P12	Male	53	Right	11	5–10 years	5	Yes	1
P13	Male	19	Center	12	< 1 year	5	No	2
P14	Female	21	Right	2	1–5 years	5	Yes	2
P15	Male	22	Left	13	< 1 year	4	Yes	1
P16	Female	36	Right	5	> 10 years	5	No	2
P17	Male	62	Left	15	> 10 years	5	No	3
P18	Male	22	Center	14	< 1 year	1	Yes	2
P19	Female	36	Left	11	> 10 years	5	Yes	1
P20	Female	24	Center	17	< 1 year	5	No	2
P21	Female	64	Right	11	> 10 years	5	Yes	1
P22	Male	36	Right	11	> 10 years	4	Yes	2
P23	Female	64	Right	0	No info	4	No	2
P24	Male	20	Left	16	< 1 year	3	Yes	3
P25	Female	46	Right	11	1–5 years	5	Yes	1
P26	Male	21	Left	2	< 1 year	3	Yes	2
P27	Female	52	Right	20	> 10 years	5	Yes	1
P28	Female	59	Left	14	> 10 years	3	Prefer not answering	2
P29	Female	45	Right	13	< 1 year	5	Yes	1
P30	Female	18	Left	13	< 1 year	5	Yes	3
P31	Female	18	Left	13	< 1 year	5	Yes	2
P32	Female	49	Right	11	> 10 years	5	Yes	2
P33	Female	18	Left	12	< 1 year	5	Prefer not answering	3
P34	Male	19	Center	12	< 1 year	4	Yes	2
P35	Male	24	Left	5	1–5 years	3	Yes	2
P36	Female	43	Left	15	> 10 years	3	Yes	2
P37	Female	25	Right	24	< 1 year	4	Yes	1
P38	No info	No info	No info	No info	No info	No info	No info	No info

B Dataset details

Here we include the information on the content we used in the intervention. All the content we used is available here:

Table 2 contains the details of the content.

Table 2: Content description of the dataset

ID	Media Name	Truth Value	Topic	Modality	Short description
1	c1_G1W1_real	Real	National News	Image	Screenshot of a RRSS post claiming Colombia is the country with the most tolls in Latin America.
2	c2_G1W1_fake	Fake	Fear mongering	Image	Screenshot of a WhatsApp story showing the face of 4 men that allegedly are kidnapping kids from schools and houses.

Continued on next page

Table 2: Content description of the dataset (Continued)

ID	Media Name	Truth Value	Topic	Modality	Short description
3	c3_G1W1_fake	Fake	Conspiracy	Image	Conspiracy infographic about "the real 2030 agenda", which changes each ODS: "1. Reducing population, ... 3. Abortion and contraception..."
4	c4_G1W1_fake	Fake	Health	Video	A doctor claiming that dengue fever is caused by a lack of calcium and should be treated with calcium and vitamin D supplements.
5	c5_G1W1_real	Real	Nature	Text	Text chain warning about the impacts of installing hummingbird feeders and recommending flowers instead.
6	c6_G1W1_real	Real	National News	Image	Information table showing the different minimum wage rates in Colombia for 2025.
7	c7_G1W1_real	Real	Regional News	Video	Journalistic format about the new traffic barriers that the Department of Mobility in Cali, Colombia, would use.
8	c8_G1W1_fake	Fake	Political (left-leaning)	Image	Screenshot of an alleged tweet from former Colombian President Ivan Duque, in which he acknowledged an alleged drug trafficker as his campaign financier.
9	c9_G1W1_real	Real	Regional/National News	Text + Image	News article about the capture of a military-linked gang that provided support to illegal armed groups.
10	c10_G1W1_fake	Fake	Fear mongering	Text	A message about an alleged new telephone crime, in which if the person answers a call or text, their SIM card will be automatically duplicated.
11	c1_G2W1_real	Real	Informative	Image	Infographic about what to do during a flood if in a vehicle.
12	c2_G2W1_fake	Fake	Health	Image	An infographic-style post purportedly contains information on the number of deaths and adverse effects from COVID-19 vaccines, claiming that this is not a respiratory virus but poisoning.
13	c3_G2W1_fake	Fake	Fear mongering	Image	Screenshot of a WhatsApp status with supposed information "direct from prison," in which a prisoner warns about an order to steal children.
14	c4_G2W1_fake	Fake	Regional/National News	Video	A citizen at a toll booth claims that a supposed "Law 769 of 2002" dictates that after waiting five minutes, you don't have to pay a toll. This law doesn't exist, but in the video, they let him go without paying.
15	c5_G2W1_real	Real	Health	Text	A chain message with recommendations that frames good health practices such as physical exercise or eating fruits and vegetables as "medicine."
16	c6_G2W1_real	Real	National News	Video	Official informational video about a line of associative credit offered by a government agency.
17	c7_G2W1_fake	Fake	Political (left-leaning)	Image	Bar chart with inaccurate and/or false information showing the number of on-duty deaths of members of the public force during three presidential terms in Colombia.
18	c8_G2W1_fake	Fake	Informative	Image	Screenshot of a post stating that when withdrawing funds from an ATM, pressing "cancel" twice prevents data theft.
19	c9_G2W1_real	Real	Health	Video	A doctor reports on the symptoms of a cerebral stroke to learn how to recognize and treat them.
20	c10_G2W1_real	Real	Nature	Text + Image	Message informing about opossums, their ecosystem benefits, and recommending respecting their lives.
21	c1_G1W2_real	Real	National News (Information)	Image	Image containing Article 199 of the Colombian Political Constitution.
22	c2_G1W2_fake	Fake	Fear mongering	Video	Video of a young man putting on a mask, with a message warning that these are the masks that "Venezuelans (immigrants) are using to rob."

Continued on next page

Table 2: Content description of the dataset (Continued)

ID	Media Name	Truth Value	Topic	Modality	Short description
23	c3_G1W2_real	Real	National News	Video	An influencer reporting on a government technology training program based on virtual courses.
24	c4_G1W2_fake	Fake	Political (left-leaning)	Video	A series of fake or hard-to-confirm anti-right political propaganda images, accompanied by a background speech by Frédéric Bastiat and delivered by an unknown host.
25	c5_G1W2_real	Real	Nature	Image	Post recommending cutting plastic rings before disposing of them to prevent birds and other animals from getting caught in them.
26	c6_G1W2_fake	Fake	Political (right-leaning)	Image	A political propaganda image containing a series of supposed promises made but not kept by Colombian President Gustavo Petro. The first four are real, but the following are false or unverifiable.
27	c7_G1W2_fake	Fake	Health	Text	Long post mixing fake information about COVID-19. Claims that it is not a virus but a bacteria, that it is caused by blood coagulation, etc.
28	c8_G1W1_real	Real	Regional News	Video	Video of an inspection visit by the Health Superintendency to a drug dispensing establishment in Bucaramanga, Colombia.
29	c9_G1W1_fake	Fake (Propaganda)	Political (no-leaning)	Video	A political propaganda video with ambiguous ideological leaning, which overlays an anti-corruption speech by El Salvador's President Nayib Bukele onto imagery associated with former Colombian President Álvaro Uribe and the Colombian right. The juxtaposition creates the misleading impression that Bukele's remarks are directed at Colombian political figures.
30	c10_G1W1_real	Real	Political News (right-leaning)	Image	Screenshot of a real headline from a mainstream media outlet about how Colombian President Petro authorized escorts for members of armed groups (in the context of a pro-peace meeting).
31	c1_G2W2_fake	Fake	Fear mongering	Image	Photo of a fake letter, allegedly issued by a security company, warning about a group of people posing as employees of the "Ministry of the Interior" and robbing homes.
32	c2_G2W2_real	Real	National News	Video	A pro-government influencer reporting on a railway project initiated by President Gustavo Petro's government.
33	c3_G2W2_fake	Fake	Political / National News	Image	Screenshot of a social media post denouncing an alleged corruption scandal involving mishandling of seized assets. Although this is suspected, it mixes up information and is inaccurate.
34	c4_G2W2_real	Real	National News	Video	Excerpt from a renowned traditional newscast about the capture of a criminal in a town in Valle del Cauca, Colombia.
35	c5_G2W2_fake	Fake	Political (left-leaning)	Image	Screenshot of an alleged quote from Congresswoman Maria Fernanda Cabal, threatening to commit a massacre against a teachers' union in Colombia.
36	c6_G2W2_real	Real	National News	Image	Screenshot of a social media post stating that The Economist magazine named Colombia the fifth best economy in 2024.
37	c7_G2W2_fake	Fake	Fear mongering	Text	Chain message warning not to answer calls from certain numbers, claiming they are criminals from other countries who can copy your entire contact list in 3 seconds.
38	c8_G2W2_real	Real	Regional News	Video	Official report on how a municipal government met with merchants in its central area to limit noise and encroachment into public spaces.
39	c9_G2W2_fake	Fake (Propaganda)	Political (left-leaning)	Video	Propaganda video edited for social media alleging that mainstream media is the true opposition to the government of Colombian President Gustavo Petro.

Continued on next page

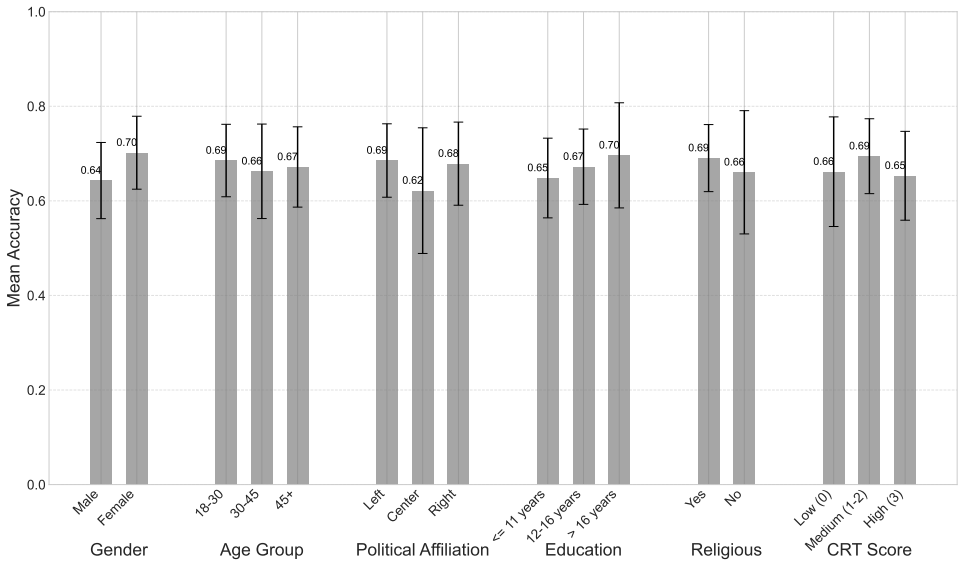

Fig. 5. Accuracy by various demographic groups

Table 2: Content description of the dataset (Continued)

ID	Media Name	Truth Value	Topic	Modality	Short description
40	c10_G2W2_real	Real	Informative	Video	Informational video about necessary precautions when riding motorcycles at intersections with trucks.

C Accuracy by demographics

Figure 5 shows the accuracy by gender, age group, political affiliation, etc.

D Supplementary Materials

All supplementary materials, including the contents used in the study, raw data, interview transcripts and notes, are available in the following online repository: https://drive.google.com/drive/folders/1tY4fykTsFqZaTPXuVt1Gv4GZj4PZpT6A?usp=drive_link.