Evaluating Community-Based and Peer Fact-Checking on WhatsApp

SUDHAMSHU HOSAMANE, TANVI GOYAL, KRITI SHARMA, MOLLY OFFER-WESTORT, KIRAN GARIMELLA

Private messaging platforms hinder public oversight, making misinformation hard to counter. Meanwhile, platforms are pivoting to crowdsourced verification amid waning trust in institutional fact-checkers. This raises a question: how do peer corrections compare with local journalists or fact-checking tiplines? We tested this via a privacy-preserving randomized field study on participants' real WhatsApp group messages in India, complemented by interviews. All three fact-checking approaches outperformed the control, with peer-administered checks modestly more effective than local journalists or national tiplines. Yet these effects failed to generalize to uncorrected messages of the same themes, reflecting the nuance of Indian WhatsApp virals. Our contributions are threefold: (1) an empirical account of participants' sensemaking of corrections in closed messaging; (2) the first ecologically valid randomized test of peer-led fact-checking on WhatsApp, benchmarked against local journalists and tiplines; and (3) design implications for chat-native fact-checking, including recruiting and training high-social-capital peers as effective in-chat verifiers.

ACM Reference Format:
Sudhamshu Hosamane, Tanvi Goyal, Kriti Sharma, Molly Offer-Westort, Kiran Garimella. 2018. Evaluating Community-Based and Peer Fact-Checking on WhatsApp. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Misinformation on encrypted messaging platforms like WhatsApp represents one of the most pressing socio-technical challenges of our time, fueling real-world violence, distorting public health outcomes, and eroding democratic discourse [5, 24]. The problem is particularly acute in the Global South, where hundreds of millions of first-time internet users rely on the platform as their primary source of information, and where false content spreads rapidly through high-trust social networks [8, 44, 50]. Recent quantitative and qualitative analysis of WhatsApp data from multiple global south countries [18], confirms this threat: a high proportion of viral content consists of misinformation, much of it inflammatory propaganda. Given that this vulnerable population operates within a high-trust, low-moderation environment, there is a clear need for practical interventions that can be implemented at scale [23].

The very architecture of WhatsApp, however, makes this problem exceptionally hard. End-to-end encryption, while a vital privacy feature, renders traditional, top-down content moderation technically impossible. The platform-endorsed alternative, the fact-checking "tipline,"—a system wherein users voluntarily forward suspicious content to a dedicated number for review by a professional fact-checking organization— represents a naive approach that has proven largely ineffective [45]. This "pull" model relies on users to recognize a potential falsehood, and take the high-friction step of forwarding content to an impersonal, institutional entity [28]. Studies consistently show these tiplines suffer from

Author's Contact Information: Sudhamshu Hosamane, Tanvi Goyal, Kriti Sharma, Molly Offer-Westort, Kiran Garimella.

Manuscript submitted to ACM

extremely low public awareness, a lack of trust, and minimal uptake, failing to solve the critical "last-mile" problem of delivering corrections to the communities that need them most [27, 44, 45].

In response to the failure of platform-led efforts, a significant body of research has explored a different paradigm: bottom-up "social corrections" initiated by peers. Lab and field experiments have shown that corrections from other users can be highly effective, substantially reducing misperceptions even on sensitive topics and often avoiding the partisan backfire effects common in Western contexts [7, 11]. Yet, this promising approach has its own critical failure: it rarely happens in the wild. Observational studies of WhatsApp groups show that users almost never publicly challenge their peers, deterred by the powerful social friction and the desire to maintain group harmony [20, 33, 50].

This leaves the field at an impasse. The dominant top-down, institutional model (tiplines) fails due to a lack of trust and reach, while the promising bottom-up, social model (peer correction) fails due to a lack of incentive and high social costs. This reveals a crucial open research question: What does a viable hybrid model look like—one that leverages the structure and resources of an organized intervention to activate the persuasive power of socially-embedded messengers? Furthermore, while survey-based studies suggest that both strong social ties and expertise matter [37, 38], we lack real-world, behavioral evidence on who is the most effective messenger to operationalize such a model.

To address this gap, this paper presents one of the first field experiments to test a community-based, hybrid fact-checking model on WhatsApp. We conducted a four-arm randomized controlled trial in Southern India, collecting real-time viral WhatsApp data from 52 participants and delivering personalized fact-checks attributed to three sources along a spectrum of social proximity: an impersonal national fact-checker (T_1), a community-embedded local journalist (T_2), or a trusted close contact (T_3).

Our findings are four-fold: (i) quantitatively, corrections from a close contact were more effective at changing beliefs than those from an impersonal national fact-checker; (ii) this effect is explained not by a simple "peer vs. expert" dichotomy, but by a more nuanced interplay between a universal need for a clear explanation and the context-dependent trust of epistemic jurisdiction; (iii) the intervention's most profound impact was not in teaching generalizable skills—which failed—but in fostering a "deliberative pause," a metacognitive shift from passive consumption to active inquiry; and (iv) our study pilots and evaluates a novel, scalable model for misinformation interventions.

The primary contribution of this work is to offer an evidence-based path forward from the current impasse. We argue that the most effective solutions are not purely technical or pedagogical, but fundamentally socio-technical. Instead of pursuing failing top-down systems or hoping for organic peer correction, we demonstrate the efficacy of a hybrid model that leverages existing social structures. By identifying, empowering, and supporting trusted local messengers, we can build a more resilient information ecosystem from the community up, offering a concrete design vision for this promising new paradigm.

2 Related Work

Designing interventions for combating misinformation is a central challenge for the HCI community. While a vast body of work evaluates strategies to improve truth discernment through nudges, labels, and friction-based designs [15, 29, 39], our study addresses the distinct problem of designing interventions for high-trust, encrypted environments like WhatsApp. Our work is situated at the intersection of three key research areas: the unique socio-technical challenges of WhatsApp, the evolution of interventions from top-down to social models, and the nuanced role of trust and the source in the effectiveness of corrections.

2.1 The Socio-Technical Challenge of Misinformation on WhatsApp

Misinformation on private messaging platforms poses a unique threat because it spreads through networks of strong social ties, where false information can travel faster than the truth [8, 41]. The platform's end-to-end encryption, while essential for user privacy, renders traditional top-down content moderation technically infeasible. The dominant intervention model has therefore been the "fact-checking tipline," where users forward suspicious content to a dedicated number for verification [28, 45].

However, research consistently shows this model is insufficient. Tiplines suffer from bottlenecks in volume and distribution, failing to reach the very users who encountered the original rumor [27]. Crucially, they face a "last-mile" delivery problem, with extremely low public awareness, a lack of trust, and minimal uptake, particularly in rural contexts [44]. Furthermore, even automated solutions like fact-checking chatbots face social hurdles; users may appreciate them for protecting vulnerable ties but are often put off by their robotic tone and errors, highlighting the need for socially-attuned design [30]. The most significant barrier, however, is social: users are often unwilling to publicly challenge their peers for fear of disrupting social harmony, leading to widespread passive acceptance of misinformation [20, 33, 50].

2.2 Intervention Paradigms: From Top-Down Controls to Social Corrections

Given the constraints of encrypted spaces, the field has increasingly pivoted from top-down interventions toward models that leverage social and community dynamics.

2.2.1. The Limits of Top-Down and Educational Interventions

On public platforms like Reddit, platform-level actions such as quarantines and bans can curb the reach of toxic communities [14, 26]. However, these enforcement levers cannot be straightforwardly imported into private WhatsApp groups [19, 46]. The other major top-down approach, media literacy training, has also shown limited success in the Indian context. Field experiments have repeatedly found that such training has no significant effect on participants' ability to identify misinformation, particularly among populations with lower digital literacy [6, 22]. Even more concerning, these interventions can backfire due to motivated reasoning, with partisans becoming less able to identify false news that aligns with their political identity after receiving training [6].

2.2.2. The Promise and Perils of Social and Community Correction

In contrast, a growing body of work demonstrates the potential of social corrections delivered by ordinary users. Field experiments in the US have shown peer corrections to be as effective as algorithmic warnings [11, 49], and research in India has found that social corrections substantially reduce belief in misinformation, even on sensitive topics, without the partisan backfire effects common in Western contexts [7]. This aligns with a broader trend toward community-driven moderation, such as Twitter's Community Notes, where the perceived identity of the corrector plays a key role in the uptake of the correction [3]. However, the promise of social correction is tempered by its rarity in the wild [18] and the finding that its average effects can be small [4], underscoring the need to understand the specific conditions under which it is most effective.

2.3 Deconstructing Trust: The Role of the Source and Message

If social corrections are a promising path forward, a critical question for HCI is: who is the most effective corrector, and what makes a correction persuasive? Recent work has begun to unpack the complex role of the source and the message. Studies have found that source effects can differ significantly by context; for instance, rural participants in India may trust journalists more, while urban participants place more trust in family [47]. The nature of the peer relationship is also crucial, with factors like tie strength and political agreement heavily influencing a correction's impact [37].

Beyond the identity of the source, the content and framing of the correction matter. The perceived clarity of an explanation and the richness of the evidence are key drivers of a correction's success, particularly for topics like health misinformation [31, 48]. This echoes findings by Badrinathan and Chauchard [7] that the simple presence of a correction often matters more than its sophistication. Foundational survey work by Pasquetto et al. [38] on WhatsApp confirms that users perceive corrections as more credible when they come from sources with either strong social ties or recognized expertise. However, much of this work relies on hypothetical scenarios and stated preferences, leaving a gap in our understanding of how these factors influence behavior in a real-world setting.

2.4 Gaps and Our Contributions

Our study is designed to address three critical gaps—empirical, theoretical, and design-oriented—at the intersection of these research areas.

Empirical Gap. While prior work has relied on hypothetical exposures to gauge the perceived effectiveness of different sources [37, 38], our study provides, to our knowledge, one of the first real-world, behavioral tests of source effects on WhatsApp. By conducting a field experiment where participants' actual WhatsApp data was collected and their belief change was measured over time, we move beyond stated intentions to capture the causal impact of corrections delivered by a national fact-checker, a local journalist, and a trusted close contact.

Theoretical Gap. Our findings contribute a more nuanced theoretical understanding of trust and persuasion in social corrections. We complicate the simple "peer vs. expert" binary by showing that while corrections from a close contact were most effective, the primacy of the explanation was a universal principle that often superseded the source's identity. Furthermore, our qualitative analysis introduces the concept of epistemic jurisdiction, demonstrating that users maintain sophisticated mental maps of who is a credible authority on specific topics. This moves beyond a static view of trust to a dynamic, context-dependent model of sensemaking.

Design Gap. Finally, our work addresses a crucial design gap. The failure of impersonal tiplines and the social friction preventing organic peer correction leave a vacuum for effective intervention design. Our study pilots and evaluates a novel model of community-based fact-checking, where trusted, socially-embedded individuals are trained and activated to deliver corrections within their existing networks. This hybrid model, which blends the scalability of an organized intervention with the social trust of a peer network, offers a concrete design pathway for HCI researchers and practitioners seeking to build more effective, socially-aware systems for combating misinformation in high-trust, encrypted environments.

3 Data collection

3.1 Study Design and Intervention Arms

We employed a mixed-methods approach, combining a pilot Randomized Controlled Trial (RCT) with qualitative interviews to investigate the effectiveness of fact-checking interventions delivered by different sources. The central hypothesis was that fact-checks are more effective when delivered by a trusted, close contact compared to more socially distant sources, an approach commonly known as community-based fact-checking [40]. The primary objective was to test the differential impact of the fact-checking source on participants' receptiveness to informational corrections. To this end, we designed a field experiment with four arms: a control group and three different treatment groups.
The T_1 arm (a 'Wizard of Oz' style WhatsApp account purporting to be a Fact-checking tipline) attributed fact-checks to a national, institutional fact-checking organization. The T_2 (Local Fact-checker) arm attributed the same messages to a professional local fact-checker unknown to the user. The T_3 (Close Network Contact Fact-checker) arm delivered fact-checks from a trained surveyor whom participants knew as a personal contact. Finally, the Control Group received neutral, non-political news content to maintain equal engagement. The core content of the fact-checking messages remained identical across the three treatment arms, with only the source attribution changing. Our primary hypothesis posited a hierarchy of effectiveness based on social proximity: $T_3 > T_2 > T_1$. We received ethics approval for all study procedures from the Institutional Review Board (IRB) at [University name will be disclosed after peer review].

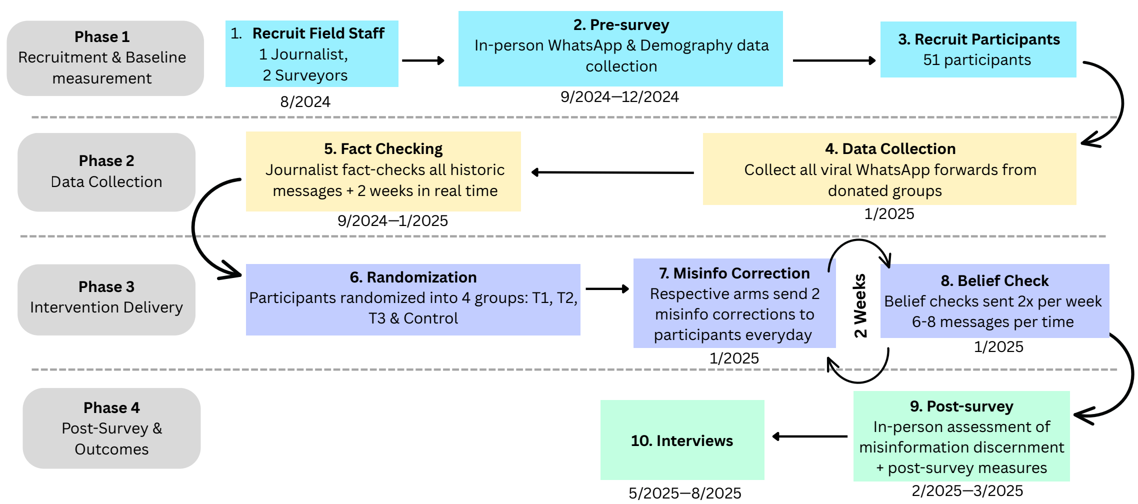

Fig. 1. Procedural workflow of the study along with timeline

3.2 Recruitment Strategy

Data collection was conducted in the state of Telangana, India. Telangana is a state in southern India with a population of 40 million people. The state is particularly interesting for this research as southern Indian states are typically understudied in the existing literature, despite their significant economic and social development in recent decades. The recruitment process involved two stages: first, the recruitment and training of surveyors and journalists; and second, the recruitment of study participants through these trained field workers.

3.2.1. Recruitment of Field Staff

We recruited field staff through a structured process with distinct criteria for journalists and surveyors. The journalist was required to have valid certification in fact-checking (e.g., from IFCN or Fact-Shaala [51]), and their primary role was to verify the authenticity of WhatsApp messages collected during the study and train the field staff in fact-checking. Surveyors, also referred to as Field Monitors, were selected based on their prior experience in conducting field surveys and extensive social ties within local communities. Their responsibilities included participant recruitment, administering in-person surveys, delivering the T_3 treatment intervention and also fact-checking the WhatsApp forwards received from the participants during the study (independent of the journalist). We hired one journalist from a major Indian fact-checking organization and two experienced surveyors, one from northern Telangana and the other from southern Telangana, to ensure regional diversity. All hired staff were trained on the study protocol, privacy-preserving features of our data collection tools, and ethical research conduct.

The surveyors were instructed not to disclose the study's specific focus on misinformation, fact-checking, or data collection processes. In cases where potential participants requested further clarification, surveyors directed them to follow-up calls with the research team, who provided consistent information on study procedures and privacy safeguards. Prior to recruitment, surveyors were trained and shown a demo of the custom-built WhatsApp data-donation tool, which highlighted its privacy-preserving features. This tool only collected "viral" or "forwarded many times" messages from donated groups, with sensitive elements such as faces automatically blurred and all text related PII automatically obfuscated. The demo equipped surveyors to reassure participants about the protection of personal information.[1]

3.2.2. Participant Recruitment

We created a WhatsApp Business API [35] based chatbot to help with the participant recruitment. The surveyors shared a standardized WhatsApp message with a link to the chatbot which guaged the participant's interest in the study, explained the remuneration and whether a surveyor could call the participant to explain about the study. The recruitment messages were sent to a total of 125 participants of which *96* expressed to participate in the study.

To be eligible, potential participants had to be active WhatsApp users of at least 2 active groups created more than 2 months prior, with at least 5 people to prevent manipulation. After excluding those who did not meet this criteria or those who denied to donate sufficient group data, our final valid sample for the experiment consisted of 51 participants.

We observe that over 76% of these participants resided in rural Telangana[2]. A majority of the participants were male (Figure 5), educated (Figure 6), under 30 years (Figure 9) and reported to be from the 'Backward Class' caste (Figure 7).[3]. A complete list of users along with some of their demographic details can be found in Table 6.

Participants consented to donate (and were eligible for) 3 WhatsApp groups on an average (Figure 12 has the full distribution for each user)

3.3 Data Collection Procedures

The study was executed in five distinct phases: a pre-survey for baseline measurement, a period of passive data collection of messages forwarded many times on WhatsApp, the intervention, immediate knowledge measure of information that have been verified during the fact-check, and a final post-survey to measure general misinformation discernment ability and other outcomes.

[1]The specific instructions, training material provided to surveyors, stimulus presented to the participants during pre-survey, belief checks and post survey, the actual corrections and all other project relevant files can be found here: https://osf.io/nvgua/?view_only=2acca01bab354da3a3341fbc7666a421
[2]Based on the geographical and socioeconomic classification provided by the "Telangana Socio Economic Outlook Report 2024" [21]
[3]How caste is measured in India: urlhttps://www.pewresearch.org/decoded/2021/06/29/measuring-caste-in-india/

3.3.1. *Phase 1: Pre-Survey and Baseline Assessment*

Surveyors conducted in-person pre-surveys at locations convenient for the participants mostly by visiting their homes. To mitigate social desirability bias, the surveyor who conducted the interview had not personally recruited the participant (we had two surveyors recruiting and surveyor 2 recruited surveyor 1's contacts and vice versa). After explaining the project's objective and the user's duties in detail, and after obtaining informed consent, surveyors administered a comprehensive questionnaire collecting data on demographics, political and social attitudes, media trust, and digital literacy. A key component was a baseline belief check for misinformation discernability on 5 sample WhatsApp forwards, where participants rated their belief in a series of true and false WhatsApp messages on a four-point Likert scale. This provided a crucial baseline measure of their ability to discern misinformation.

3.3.2. *Phase 2: WhatsApp Data Collection and Fact-Checking*

At the conclusion of the pre-survey, participants were onboarded to a custom-built, privacy-preserving data donation tool, WhatsApp Explorer [19]. Participants provided explicit consent and selected which WhatsApp groups they wished to contribute. The tool was designed to protect user privacy by only collecting messages marked as "forwarded many times," (indicating content that spreads virally on WhatsApp) ignoring personal chats, and automatically blurring faces in and obfuscating and text-based PII. The collected viral messages were then systematically fact-checked by our trained journalist who not only used a standardized protocol of keyword searches, reverse image searches, and geolocation tools, reverse video search and metadata searches, but also found evidence to falsifiable parts of the claim when media debunks weren't readily available. This process yielded a set of verified misinformation items to be used in the intervention. The journalist fact-checked *1017* messages from about 3000 total messages received until mid January, when the journalist stopped fact-checking. Of the 1017 messages factchecked 656 were existing data in the groups and 361 were received between September 2024 and January 2025—the start of participant recruitment and the end of data collection. The journalist maintained a spreadsheet identifying each WhatsApp forward with its corresponding link from the 'WhatsApp Explorer' tool, well pre-defined categories for misinformation that the journalist carried over from his practice of correcting misinformation in similar contexts, information about whether a particular message was fake or not, the claims the message was making and the debunks.[4]. We want to highlight that while the journalist was able to view all the WhatsApp forwards on the dashboard of the data collection tool, they couldn't know which participant received it.

We disconnected the users' whatsapp groups from the WhatsApp explorer at the end of data collection phase and the participants were informed about it. Participants were also shown how to disconnect their WhatsApp groups from the data donation tool, and were allowed to disconnect their groups before the end of the study without any loss of compensation.

3.3.3. *Phase 3: Randomization and Intervention*

Following a two-week period of data collection, participants were randomized into one of the four study arms. Randomization was stratified based on their baseline misinformation discernment scores measured during pre-survey data collection, to ensure balance across groups. The intervention was delivered over 13 days via WhatsApp. All participants received fact-checks to two messages that were common to everyone, everyday, but sent by different sources. To mitigate bias arising from framing and verbiage, we provided the same fact checks to all the sources. Only on day 9, an additional

[4]We provide this spreadsheet along with the actual WhatsApp forwards received in the project data repository: https://osf.io/nvgua/?view_only=2acca01bab354da3a3341fbc7666a421

personalized fact check was sent to different misinformation received by 7 participants in their own WhatsApp groups.[5]. In order to send "fact checks" everyday, we follow previous research in misinformation intervention [42, 43] and had an all-false misinformation stimulus for the period of our interventions.

Participants were sent belief checks for viral misinformation from a range of themes that were prevalent in the data. These included 'health misinformation', 'partisan political propaganda', 'communal or religious misinformation', 'manipulated visuals', 'economic/legal claims', and 'neutral or inspirational content'—predefined themes borrowed by the journalist from his work.

While the intervention phase began with the delivery of corrections to misinformation, we also sent belief checks using an automated WhatsApp chatbot (the same chatbot we used for recruitment) for the same previously corrected misinformation twice a week at uniform intervals. This was measured on a binary scale ('Think it's fake', 'Think it's not fake', 'Don't know'). This helped us get the closest measure of misinformation discernment on the exact message that they received as fact-check. A detailed table of each stimulus sent along with the theme and the truth discernment is provided in table 10.

3.3.4. Phase 4: Post-Survey and Outcome Measurement

After the intervention, surveyors conducted in-person surveys to measure the impact of corrections on uncorrected misinformation, but on similar themes as the misinformation sent for belief checks. We also measured whether any other key measures: (1) Their perception of trusting in news sources—Government officials, local journalist, national news website, news channel, politician, peers, (2) Their trust in news on WhatsApp, (3) Their favorability towards two major but ideologically opposite political parties—BJP and INC (Indian National Congress) changed from what they had answered in the presurvey, (4) Favorability for Hindus, Muslims and Sikhs, changed by any amount compared to the measures during the pre-survey.

We also operationalised a novel *Willingness to pay (WTP)* measure for prioritized fact checking service. WTP is the highest amount of money a person says they would spend to get a clearly described service. We use it as a behavioral measure of that person's *private value*—i.e., the benefit to *them personally*. To elicit WTP in a way that encourages honest answers, we use the Becker–DeGroot–Marschak (BDM) procedure [10, 32]. [6] The exact operationalization is described in section A.1. We had two hypothesis regarding WTP:

(H1) Participants exposed to fact-checking during the study (T1, T2, T3) will show higher mean WTP than Control: $\mathbb{E}[WTP_{T1,T2,T3}] > \mathbb{E}[WTP_{\text{Control}}]$.

(H2, exploratory) Because perceived credibility and speed likely matter, we expect $WTP_{T2\ \text{(journalist)}} \geq WTP_{T1\ \text{(tipline)}} \geq WTP_{T3\ \text{(peer)}}$.

3.4 Interviews

To add nuance and depth to our quantitative findings, we conducted semi-structured qualitative interviews with 15 participants (P6-P20 in table 6). These interviews explored their sensemaking processes, information verification habits, and their experience with the fact-checking intervention. The interview protocol is uploaded to the project's data repository for reference.

[5]While the initial goal was to send personalized fact checks for everyone, everyday, we could not find enough misinformation from a majority of our participants to make personalized fact-checking feasible

[6]In BDM, a participant writes down the *highest price* they would pay. Then a price is drawn at random. If the drawn price is at or below their number, they receive the service and pay the *drawn* price; otherwise, they do not buy and pay nothing. Intuitively, overstating can make you pay more than the service is worth to you, and understating can make you miss getting the service when you would have liked to—so the best move is to state your true maximum.

We employed reflexive thematic analysis in line with Braun and Clarke [13]. All the interviews were conducted in Telugu—the local language of Telangana, by the first author. The first author transcribed the audio recordings manually, verified a subset of the transcripts with the surveyors (who were based out of Telangana) for correctness, and conducted an exhaustive, line-by-line open-coding [16]. This open-coding generated 172 preliminary codes. The first author, with the consulation of the last author, further conducted axial coding—iteratively clustering relating codes, refining boundary categories and articulating higher-level patterns. Through successive rounds of discussion, we converged on six overarching themes— *'Calibrated Trust and Source Evaluation'* (Trust was not purely relational – it depended on who spoke, what topic they spoke on, and how they explained it), *'Evidence-seeking and proof orientation'* (Participants sought evidence through various channels and emphasised that explanations must accompany claims), *Collective sensemaking and social gatekeeping* (Participants shared messages selectively and engaged others in verification, acting as gatekeepers within their networks), *Perceived bias and motive* (Participants were acutely aware of bias in both mainstream media and individuals), *Behavioural shifts and Meta-learning* (After participating in the study, many respondents described changes in their digital habits and awareness of misinformation). We used Nvivo to code the transcripts and spreadsheets to organize open codes.

3.5 Researcher Positionality

The research team brought a diverse set of skills and backgrounds to this project. Four of the five authors are based in India; two are from South India and are fluent in Telugu. Two authors have extensive fieldwork experience in India, and four have led or contributed to misinformation research in India and other developing countries. At the same time, we acknowledge that we were all born and raised in urban settings and most are university-affiliated researchers. These backgrounds—and the privileges and assumptions that accompany them—may shape how we interpret rural participants' behaviors and sensemaking around fact-checking.

We approached the study reflexively. The first author (Telugu-speaking) conducted all interviews in Telugu. Quotations presented in this paper are close translations intended to preserve participants' wording and intention; translations were reviewed by Telugu-fluent co-authors, and paraphrasing was avoided except to clarify untranslatable idioms (marked in brackets when used). Throughout data collection and analysis, we maintained reflexive memos, discussed positionality and potential confirmation bias in team meetings, and sought disconfirming evidence when developing themes. We also validated descriptive details and interpretation with field surveyors based in Telangana. These practices help us center participants' voices while tempering our urban, academic vantage point.

4 Findings

4.1 The Content and Context of Forwarded Messages

Intervention studies on misinformation often test generic (usually popular, already fact checked) falsehoods on samples of users. However, to design effective interventions, one must first understand the specific information ecosystem in which users operate. We begin by characterizing the nature of the content circulating within participants' WhatsApp networks and their existing practices for navigating it. Our analysis reveals an environment dominated by community-oriented groups where emotionally resonant, narrative-driven propaganda often overshadows easily verifiable misinformation, posing a significant challenge for traditional fact-checking.

4.1.1. *The Primacy of Community-Anchored Groups*

Our analysis of WhatsApp groups from which participants donated data reveals that misinformation and propaganda primarily flow through high-trust, non-political community networks. These were not niche political forums but groups deeply integrated into the social fabric of users' lives. Village or town-based groups formed the largest single category (27%), followed by school and alumni networks (11%), and groups centered on caste, family, religion, and work (around 10% each). Explicitly partisan political groups constituted a small minority of the total (around 6%). This context is critical, as information received through these trusted community channels is likely to be met with less initial skepticism [23]. Unlike the topic based themes for misinformation mentioned in section 3.3.3, which the journalist had ready at-hand and dealt with everyday, coding the themes for WhatsApp groups wasn't derivable. The journalist took a pragmatic, exploratory approach. Knowing Telugu made the journalists task a bit easier—they first analyzed group names and descriptions and started with a small set of broad initial groups based on geography or topic when it was explicit in the name or description. For groups that didn't have a clear identifier, the journalist inferred themes from message and media contents in the group and the language and social cues in those forwarded messages. They iteratively coded the groups into more granular codes until they couldn't break it down further.

4.1.2. *Verifiable Falsehoods vs. Unverifiable Narratives*

Within these groups, we analyzed a corpus of over 650 messages marked as 'forwarded many times'—the same messages that were pre-existing in all the participants' WhatsApp groups, which were collected before September 2024. Our journalist collaborator coded these messages into thematic categories described in previous sections, including generic misinformation, health misinformation, religious propaganda, and political propaganda (similar to a rubric developed by previous work [20], on a similar WhatsApp dataset). A key distinction emerged not just in the topic of the content but in its fundamental verifiability, which profoundly impacts the feasibility of fact-checking.

A significant portion of the misleading content consisted of verifiable falsehoods. This included "generic misinformation" with concrete, factual claims that could be readily debunked using online search tools. Nearly 90% of messages in this category had a corresponding media fact-check available online. These forwards often employed common tactics like decontextualization, where authentic media was repurposed with false captions to fit a new narrative. Health misinformation also relied on factual claims but was less consistently covered by existing fact-checks (only $\approx$ 46% had an available debunk), likely because the claims were hyper-local, mutated rapidly, or required specialized medical knowledge to verify. While these messages were often false, their claim-based structure made them amenable, in principle, to evidence-based correction.

In stark contrast, a larger and more insidious category of content consisted of unverifiable narratives, primarily in the form of political and religious propaganda. This content was engineered to resist simple fact-checking by relying on emotional resonance and identity-based appeals rather than disprovable facts. For instance, propaganda targeting Hindus frequently paired authentic media with inflammatory text overlays to suggest a threat to the community. Because the underlying media was real, a traditional fact-check was often irrelevant; the harm was in the interpretive frame. Our data reflects this challenge: only 5% of Hindu-centric propaganda (95/650) messages had a corresponding media fact-check, as most were not straightforwardly debunkable. Similarly, political propaganda, primarily targeting the opposition Congress party, focused on building a negative narrative of corruption and anti-Hindu sentiment. The goal was not to convey factual information but to reinforce in-group identity and mobilize political anger. Such narrative-based content falls outside the scope of binary true/false verification, rendering traditional fact-checking an ineffective tool. Although the share of total participants inclined towards BJP wasn't drastically different from INC (Table 3), all the political propaganda found in our dataset was against the INC.

4.1.3. Participants' Information Verification Practices

The distinction between verifiable falsehoods and unverifiable narratives is crucial because it maps directly onto how participants decide whether to scrutinize a piece of information. Our interviews reveal that existing verification practices are not applied uniformly to all content. Instead, they are highly situational, triggered almost exclusively by messages that resemble the fact-based, verifiable falsehoods described above, particularly those with direct utilitarian consequences. As we explore below, this leaves users vulnerable to narrative-based propaganda, which fails to activate their established methods of scrutiny.

Participants were not passive consumers of information but engaged in a range of existing verification practices, often blending digital tools with social consultation. Their decision to verify a message was not random but a calculated act, typically motivated by utilitarian concerns and constrained by practical barriers.

The most common impetus for verification was utilitarian value and perceived risk. Participants were most likely to invest effort in checking information with a direct bearing on their financial or physical well-being, such as potential online scams or government schemes. This vigilance was often born from negative personal experience. Participant P27, for instance, recounted losing money in an online scam, which made her deeply cautious about financial links: she now relies on her husband to verify every online transaction. This protective instinct extended to family, with more digitally literate participants acting as gatekeepers. P2 described her role in protecting her father from fraud: *"My father brings everything to me first... I tell him not to share the OTP [One Time Password (for two factor authentication)] with anyone..."*.

When participants chose to verify a claim, they employed two primary strategies: digital self-checks and social outsourcing. Younger, more tech-savvy individuals turned to Google or AI assistants. However, many others treated verification as a social act. P16 preferred *"asking people I know who are more aware"* rather than searching alone. This practice of social consultation often complemented digital searches. P22 would first consult knowledgeable friends before turning to Google for a final verdict, explaining, *"they (friends) are usually up-to-date with the current affairs... has become a habit relying on [them] to keep updated..."*.

However, these verification practices were situational, not constant. The decision to check information was gated by significant practical constraints. Time was the most frequently cited barrier; as P26 noted, her job left little opportunity for extensive verification. Beyond time, personal interest was a key factor. As P10 candidly admitted, he only bothers to google a claim if he finds the topic personally interesting. These barriers underscore that even for those with the skills and intent to verify, fact-checking is a high-friction, occasional activity rather than a reflexive habit.

4.2 Intervention Outcomes and Sensemaking Across Arms

Our analysis of the intervention reveals a clear, statistically significant advantage for fact-checks delivered by a close social contact. Participants in this arm (T_3) demonstrated the greatest improvement in accurately identifying misinformation. However, this quantitative effect is nuanced by complex social dynamics, including lower engagement rates and a sophisticated, context-dependent model of trust. In this section, we first present the quantitative outcomes of the intervention and then draw on our interviews to explain the mechanisms driving these results.

4.2.1. Quantitative Efficacy: The Advantage of Close-Contact Corrections

The primary outcome of our experiment was the change in participants' ability to correctly identify the veracity of news items. We measured this as participant-level accuracy, comparing the share of correct answers in each treatment arm to the control group. The results, presented in Table 1, show a clear hierarchy of effectiveness consistent with our hypothesis ($T_3 > T_2 > T_1$).

The most substantial impact was observed in the T_3 (Close Network Contact) arm. Participants receiving corrections from their known surveyor contact exhibited a 28.1 percentage point increase in accuracy over the control group. This effect was statistically significant even after correcting for multiple comparisons (p=0.010, p_{Holm}=0.030). The T_2 (Local Journalist) arm also showed a notable positive effect, with a 24.2 percentage point increase in accuracy, though this result was borderline significant after correction (p=0.040, p_{Holm}=0.079). The T_1 (National Fact-Checker) arm produced a smaller, non-significant increase of 11.1 percentage points. This quantitative evidence provides strong directional support for the idea that social proximity is a key mediator of a fact-check's persuasive power.

Table 1. Participant-level accuracy: pairwise Welch tests versus Control (two-sided). Means are proportions; ATE reported in percentage points. Covariance type = HC3 for small sample efficiency

Arm	n_{treat}	n_{ctrl}	$\text{Mean}_{\text{treat}}$	$\text{Mean}_{\text{ctrl}}$	ATE (pp)	SE	95% CI	p	p_{Holm} / p_{FDR}
T1	12	11	0.319	0.208	0.111	0.099	[−0.088, 0.310]	0.260	0.260 / 0.260
T2	11	11	0.450	0.208	0.242	0.117	[0.005, 0.479]	0.040	0.079 / 0.059
T3	12	11	0.489	0.208	0.281	0.109	[0.061, 0.500]	0.010	0.030 / 0.030

However, the powerful effect of the T_3 arm is made more remarkable by its engagement patterns. As shown in Figure 2, the belief-check completion rate in the T_3 arm exhibited a notably lower median and wider variance compared to the other arms, which clustered near 100% completion. This suggests that the strong average treatment effect in T_3 was achieved despite some participants disengaging from the daily belief-check prompts. This paradox—higher efficacy coupled with lower completion—points to complex social dynamics at play.

While the numbers demonstrate a clear hierarchy of effectiveness, they do not explain the underlying mechanisms. Our qualitative interviews provide critical insight into why corrections from a close contact proved so persuasive, why journalists were also effective, and why the social nature of the T3 intervention simultaneously increased its impact and created barriers to engagement.

4.2.2. Interpretive Mechanisms: Deconstructing Trust and Persuasion

Our qualitative findings reveal that the quantitative effects are driven by two primary factors: a universal demand for clear explanations and a highly contextual, subject-specific hierarchy of trust that privileges different sources for different topics.

The Primacy of Explanation. Across all treatment arms, the single most important factor for a fact-check's acceptance was the presence of a reasoned explanation. A simple verdict of 'true' or 'false' was insufficient; participants required justification. As P2 insisted, *"without an explanation, how can we believe it directly?... I won't."*. This evidence-first stance was the foundation of persuasion. P14 articulated a common heuristic: the sheer effort of crafting an explanation signals credibility, as he questioned, *"why would someone go out of their way to create an explainer?"*. This underlying principle explains why all three treatment arms outperformed the control group: they provided reasoned arguments that satisfied this fundamental cognitive need.

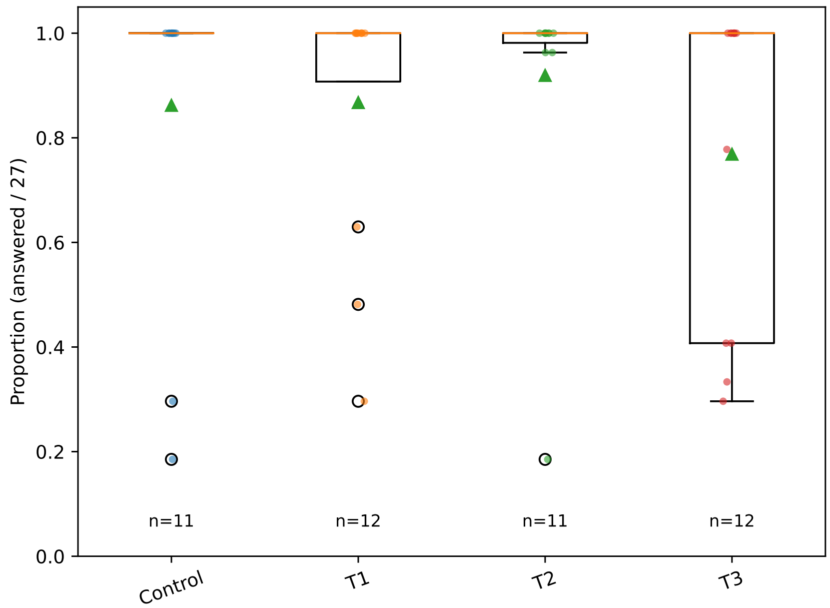

Fig. 2. Belief-check completion rate per intervention arm

Table 2. Pairwise comparisons of pre-study trust by source (4-point scale; lower = more trustworthy).

Comparison (A vs B)	n	Mean(A)	Mean(B)	Δ	95% CI for Δ	p_{Holm}
Panel A: Holm-significant at α = 0.05 (bolded)						
Peer vs Politician	**50**	**1.620**	**3.120**	**1.500**	**[1.248, 1.752]**	<0.001
Peer vs National news website	**51**	**1.628**	**2.059**	**0.431**	**[0.185, 0.678]**	**0.014**
Peer vs Local journalist	**51**	**1.627**	**2.098**	**0.471**	**[0.223, 0.718]**	**0.0006**
Local journalist vs Politician	**50**	**2.100**	**3.120**	**1.020**	**[0.736, 1.304]**	<0.001
News channel vs Politician	**50**	**1.900**	**3.120**	**1.220**	**[0.988, 1.452]**	<0.001
Government official vs Politician	**50**	**1.960**	**3.120**	**1.160**	**[0.823, 1.497]**	<0.001
National news website vs Politician	**50**	**2.080**	**3.120**	**1.040**	**[0.715, 1.365]**	<0.001
Panel B: Not Holm-significant († = BH-FDR < 0.05)						
Peer vs Government official†	51	1.628	1.961	0.333	[0.078, 0.589]	0.141
Local journalist vs News channel	51	2.098	1.882	-0.216	[-0.405, -0.027]	0.215
National news website vs News channel	51	2.059	1.882	-0.176	[-0.385, 0.032]	0.620
Local journalist vs Government official	51	2.098	1.961	-0.137	[-0.395, 0.121]	0.738
Government official vs National news website	51	1.961	2.059	0.098	[-0.149, 0.345]	1.000
Government official vs News channel	51	1.961	1.882	-0.078	[-0.341, 0.185]	1.000
Local journalist vs National news website	51	2.098	2.059	-0.039	[-0.271, 0.192]	1.000
Peer vs News channel	51	1.628	1.882	0.255	[0.011, 0.499]	0.264

Notes. Trust measured on a 4-point scale *before* the study. Δ = mean(B) − mean(A); positive values indicate source B rated higher than A. p_{Holm} is the Holm–Bonferroni adjusted two-sided p for the Wilcoxon signed-rank test; values $<$ .05 are bolded. †Also significant under Benjamini–Hochberg FDR at q = .05 (here, Peer vs Government official; p_{FDR} = 0.033) but not under Holm. Unadjusted p-values are omitted for brevity.

Hierarchies of Trust and the Source as a Tiebreaker. While explanations were a necessary condition, the source of the explanation acted as a powerful tiebreaker, explaining the variance between the T_1, T_2, and T_3 arms. Participants did not have a uniform model of trust; instead, they deployed what we term epistemic jurisdiction—assigning credibility based on the topic at hand. This is corroborated by our pre-study quantitative data in Table 2, which shows that a Peer was rated as significantly more trustworthy than a Local journalist (p <0.001) and a National news website (p=0.014) even before the intervention began.

Our interviews illuminate this hierarchy. For local news, peers were supreme. For national issues, participants deferred to professional journalists, as articulated by P10: *"For my village, I trust the villagers. For national-level issues, I have to trust the news channels."* Journalists (T_2) were seen as credible due to their professional mandate. P26 explained she would trust a journalist over a peer on a contentious political topic because *"Journalism is their work... they research thoroughly."* This perceived rigor explains the strong, albeit not fully significant, effect of the T2 arm.

However, for many, this professional mandate was viewed not as a sign of rigor but of bias. A deep skepticism of media institutions led some participants, like P23, to trust their friends and family far more for political information, viewing them as more honest brokers. It is this segment of the population for whom the T_3 intervention was likely most powerful, leveraging the high baseline trust evident in Table 2. This pre-existing social bond gave the corrections an authenticity that an unknown journalist or an anonymous tipline (T_1) could not replicate.

The Social Friction of Close-Contact Intervention. Finally, our qualitative data directly explains the completion rate paradox seen in Figure 2. The very social tie that made the T_3 intervention effective also introduced social friction that inhibited engagement for some. Participants in T_1 and T_2 were interacting with an impersonal source, making it a low-stakes activity. In contrast, T_3 participants were being asked to state their beliefs to a known member of their community. [Laxmanna], a T_3 non-responder, explicitly stated that because the surveyor was his neighbor, he preferred not to share opinions that could strain their relationship. He confirmed he would have answered if his responses were anonymous. This reveals a critical trade-off: close-contact interventions can be more persuasive, but they may also suppress participation due to the social risks involved in expressing potentially controversial opinions to a peer.

4.3 Post-Intervention Outcomes: Behavioral Shifts and Their Limits

While our analysis of the in-study belief checks demonstrated that the intervention could successfully correct specific pieces of misinformation, a critical question remains: did this experience lead to broader, more durable changes in skills, attitudes, and behaviors? Our post-survey quantitative analysis reveals a complex picture. We find limited evidence that the intervention imparted generalizable fact-checking skills or altered deeply entrenched political and religious attitudes. However, we find significant evidence that the intervention successfully shifted participants' trust in specific information sources in a manner consistent with their treatment arm. It also fostered a greater perceived value for fact-checking services, though this did not reach statistical significance.

Table 3. Participant-level accuracy per arm (BINARY). Columns show sample size (n), mean accuracy, standard deviation (sd), standard error (se), and 95% CI bounds.

Arm	n	Mean	sd	se	ci_lo	ci_hi
Control	12	0.535	0.126	0.037	0.455	0.616
T1	13	0.495	0.140	0.039	0.430	0.560
T2	12	0.495	0.149	0.043	0.401	0.590
T3	13	0.511	0.101	0.028	0.450	0.572

4.3.1. Quantitative Findings: A Narrow Path of Influence

Our post-survey analysis began by testing whether participants in the treatment arms had developed a generalized skill in identifying misinformation. Participants were shown a new set of ten true and false news items they had not seen during the intervention. As shown in Table 3, we find no evidence of such skill transfer. The mean accuracy scores across all three treatment arms were statistically indistinguishable from the control group, with all arms performing at a level close to chance (around 50%). This null result strongly suggests that exposure to specific fact-checks, even when effective in the moment, does not automatically equip individuals with the abstract principles needed to identify novel falsehoods. The skills learned appear to be context-bound and do not readily generalize.

Moving from cognitive skills to social attitudes, we also assessed whether the intervention altered participants' favorability toward major political parties and religious groups. Here too, our data reveals a powerful inertia. We found no consistent, significant pre-post changes in these deeply held views across the treatment arms (see Appendix, Tables 7-9). While we noted an isolated significant change in the T_3 arm, the overwhelming pattern indicates that a short-term informational intervention is insufficient to alter beliefs rooted in long-standing personal and community identity. We did however see qualitatively that many participants participants tempered party extremeties to move closer to the ideological center. We see that for both the congress and BJP, post-survey results show that many pro-INC supporters became more accepting of BJP (Table 14) many INC supporters lessened their support to INC (Table 17), and many BJP supporters, increased their support for the INC (Table 16).

However, where the intervention failed to change generalized skills or broad attitudes, it succeeded in a more targeted manner: shifting whom participants trust for information. The results in Table 4 show a clear, source-specific impact. Participants in the T_2 arm, who received corrections from a local journalist, reported a large and statistically significant increase in their trust for both the Local journalist (Δ=0.846, p=0.021) and News channels (Δ=0.846, p=0.017). This indicates that a positive, helpful interaction with a media professional can directly improve trust in that professional category. Conversely, we see a fascinating, statistically significant decrease in trust for Peers in the T_1 arm (Δ=0.308, p=0.046), suggesting that receiving impersonal, institutional fact-checks may lead individuals to view their own social networks as comparatively less reliable. The T_3 arm showed a directional, though not significant, increase in trust for peers. This demonstrates that the intervention's most potent and lasting quantitative impact was not on what participants believed in general, but on who they believed.

Finally, to gauge the perceived value of the intervention, we conducted a willingness-to-pay (WTP) experiment. As shown in Table 5, participants in the treatment arms were, on average, willing to pay more for a fact-checking service than those in the control group (INR 73.4 vs. INR 62.5). While this provides directional evidence that the service was valued, the difference was not statistically significant (p=0.267). The highest WTP was observed in the T_2 (journalist) arm, suggesting that participants may associate fact-checking most strongly with a professional service (H2). We could however, find any evidence for H1 ($\mathbb{E}[WTP_{T1,T2,T3}] > \mathbb{E}[WTP_{\text{Control}}]$.)

4.3.2. *Qualitative Insights*

While the quantitative data points to narrow and specific impacts, our qualitative interviews reveal a profound and consistent change in some of the participants' cognitive process for engaging with information. The most significant outcome reported was the development of a 'deliberative pause'—a newfound habit of stopping to think before believing or sharing. P14 captured this metacognitive shift:

> *"Previously, it was just blind trust... whatever came, we would just accept it. Now there is a bit of observation, of cross-checking. There's more patience to see if it's real or not."*

Table 4. Per-source changes with paired means, 95% CIs, Wilcoxon p, and rank–biserial r (post and pre-study measured on a 4-point scale; lower = more trustworthy)

		Arms			
Source	Metric	T1	T2	T3	Control
Government official	n_{pairs}	13	13	13	12
	Mean (pre)	1.846	1.769	2.077	2.167
	Mean (post)	2.000	1.846	2.000	2.000
	Δ (post−pre)	0.154	0.077	-0.077	-0.167
	95% CI for Δ	[-0.262, 0.570]	[-0.550, 0.704]	[-0.598, 0.444]	[-0.533, 0.200]
	Wilcoxon p / r	0.414 / 0.333	0.455 / 0.071	0.511 / -0.143	0.317 / -0.500
Local journalist	n_{pairs}	13	13	13	12
	Mean (pre)	2.154	2.154	2.077	2.000
	Mean (post)	2.154	1.308	2.077	2.250
	Δ (post−pre)	0.000	**-0.846**	0.000	0.250
	95% CI for Δ	[-0.349, 0.349]	**[-1.443, -0.250]**	[-0.427, 0.427]	[-0.229, 0.729]
	Wilcoxon p / r	1.000 / 0.000	**0.021** / -0.714	0.631 / 0.000	0.296 / 0.600
National news website	n_{pairs}	13	13	13	12
	Mean (pre)	2.000	2.154	2.154	1.917
	Mean (post)	1.923	1.846	1.538	1.833
	Δ (post−pre)	-0.077	-0.308	-0.615	-0.083
	95% CI for Δ	[-0.598, 0.444]	[-0.931, 0.316]	[-1.246, 0.015]	[-0.508, 0.341]
	Wilcoxon p / r	0.908 / -0.143	0.360 / -0.417	0.066 / -0.689	0.655 / -0.200
News channel	n_{pairs}	12	13	13	12
	Mean (pre)	1.750	2.077	2.077	1.667
	Mean (post)	1.667	1.231	1.923	1.750
	Δ (post−pre)	-0.083	**-0.846**	-0.154	0.083
	95% CI for Δ	[-0.587, 0.420]	**[-1.389, -0.303]**	[-0.570, 0.262]	[-0.341, 0.508]
	Wilcoxon p / r	0.705 / -0.143	**0.017** / -0.736	0.414 / -0.333	0.655 / 0.200
Politician	n_{pairs}	12	11	11	9
	Mean (pre)	2.917	3.091	2.909	3.333
	Mean (post)	3.333	2.909	3.545	3.333
	Δ (post−pre)	0.417	-0.182	0.636	0.000
	95% CI for Δ	[-0.087, 0.920]	[-0.686, 0.323]	[-0.116, 1.389]	[-0.384, 0.384]
	Wilcoxon p / r	0.096 / 0.714	0.526 / -0.500	0.100 / 0.667	1.000 / 0.000
Peer	n_{pairs}	13	13	13	12
	Mean (pre)	1.308	1.615	1.923	1.667
	Mean (post)	1.615	1.615	1.462	1.667
	Δ (post−pre)	**0.308**	0.000	-0.462	0.000
	95% CI for Δ	**[0.017, 0.598]**	[-0.493, 0.493]	[-1.046, 0.123]	[-0.383, 0.383]
	Wilcoxon p / r	**0.046** / 1.000	0.746 / 0.000	0.096 / -0.714	1.000 / 0.000

Notes. Δ is the paired mean difference (post − pre). Scale is 1–4; lower values indicate higher perceived trustworthiness. Decrease in mean values during post-study measurements indicate increased trust in a news source compared to pre-study. Bold indicates $p < .05$ (two-sided Wilcoxon signed-rank).

Table 5. Post-survey WTP (in INR): group summaries and Welch comparison (Any treated vs Control)

Panel A: Group summaries (means ± 95% CI)

Group	n	Mean	SD	SE	95% CI (lo)	95% CI (hi)
Control	12	62.500	30.189	8.715	43.319	81.681
Any treated	38	73.421	22.931	3.720	65.884	80.958

Panel B: Welch comparison (Any treated − Control)

Comparison	Diff (INR)	95% CI (lo)	95% CI (hi)	t	df	p	$n_{\text{treated}}/n_{\text{control}}$
Any treated − Control	10.921	-9.250	31.092	1.153	15.223	0.267	38 / 12

Notes. WTP3 is the amount (INR) participants would pay for one week of priority fact-checking. Panel A CIs are mean ±1.96×SE (normal approximation). Panel B reports Welch two-sample t-test (unequal variances) comparing Any treated to Control on WTP3; Diff = $\text{Mean}_{\text{treated}} - \text{Mean}_{\text{control}}$. The Welch test is not significant at $\alpha = 0.05$.

This pause explains the disconnect between the process and the outcome. While participants did not become expert fact-checkers overnight (as seen in the null generalization results in Table 3), they overwhelmingly adopted a more skeptical and cautious disposition. They moved from being passive recipients to active questioners.

This new disposition manifested as a sense of civic duty. Participants began to see themselves as responsible nodes in their information network. Some, like P23, moved from non-sharers to active curators of what they deemed to be helpful health and political information, motivated by a desire to create "awareness" for others. More strikingly, some became public correctors. P34 described actively debunking a false story in a group by searching for evidence on Google and pasting the results for others to see. These actions represent a significant behavioral shift fostered by the intervention—not just a change in personal belief, but the adoption of a community-oriented role in maintaining information hygiene.

Finally, our interviews shed light on why participants do not see fact-checking as an exclusive, expert service, which helps contextualize the WTP results. Most viewed it as a skill anyone could learn, primarily involving "Google search," and were unaware of professional certifications or more advanced techniques. This perception—that fact-checking is an accessible, democratized skill—is both an opportunity and a challenge. It empowers individuals like P34 to take action but may also lead to an underestimation of the expertise required to debunk complex misinformation, and thus a lower willingness to pay for it as a professional service.

5 Discussion

This study set out to conduct one of the first field experiments testing a community-based, "bottom-up" fact-checking model against traditional institutional approaches within an encrypted messaging environment. The results paint a complex but coherent picture: while corrections from a trusted close contact were most effective at shifting beliefs about specific falsehoods, the intervention's most profound impact was not in teaching generalizable skills, but in fostering a more deliberative mindset. These findings compel us to re-evaluate the core assumptions behind misinformation interventions. We argue that effective design for encrypted spaces must pivot from a futile quest for scalable, universal truth-adjudication and toward a more nuanced, socio-technically grounded goal: scaffolding community resilience.

5.1 Deconstructing the 'Trusted Messenger'

A primary contribution of our work is to complicate the simplistic "peer vs. expert" dichotomy that dominates much of the discourse on source credibility (Section 4.2). Our quantitative results clearly show a hierarchy of effectiveness ($T_3 > T_2 > T_1$), demonstrating that the identity of the messenger is a powerful mediator of an intervention's success. However, our qualitative data reveals that this effect is not driven by a static perception of "trust," but by a dynamic, context-dependent social process we term epistemic jurisdiction.

Participants did not hold a universal belief that peers are more trustworthy than journalists. Instead, they maintained sophisticated mental maps of who holds the legitimate authority to speak on specific topics. A doctor held unquestioned jurisdiction over health; a fellow villager was the ultimate authority on local affairs; a journalist was a credible source for distant public events. The T_3 (close contact) arm was so effective not merely because the source was a "peer," but because much of the misinformation in circulation—local scams, community rumors, fabricated government schemes—fell squarely within the jurisdiction where social proximity is the primary credential. The moderate success of the T_2 (journalist) arm likewise reflects the instances where the content (e.g., a national political claim) fell within the journalist's jurisdiction.

This finding is further nuanced by the primacy of the explanation. Across all arms, a well-reasoned argument was a prerequisite for acceptance. The source's jurisdiction, therefore, can be understood as granting the license to be heard, while the explanation provides the substance to be believed. This interplay has critical implications. It suggests that interventions are most potent when the right messenger is equipped with the right message for the right topic. It also reveals a significant vulnerability: sophisticated propaganda that mimics an explanatory format can succeed if it addresses a topic where users feel their peers hold jurisdiction, effectively borrowing the legitimacy of the social tie.

5.2 Designing for Efficacy Amidst Social Friction

While leveraging close social ties is powerful, our findings reveal it is also fraught with peril. A key paradoxical result of our study (Section 4.2.1) was that the most effective arm (T_3) also exhibited the lowest and most variable rates of engagement with our daily belief-check prompts. This points to the critical, and often overlooked, role of social friction in computer-mediated correction.

Our qualitative data provides a clear explanation: publicly stating a belief or correcting a peer is a socially risky and costly act. Participants in the T_1 and T_2 arms were interacting with an impersonal entity, making their responses low-stakes. In contrast, T_3 participants were being asked to express potentially controversial opinions to a known member of their community. As one non-responder articulated, he was unwilling to share views that could strain his relationship with his neighbor.

This paradox presents a central design challenge. Any system that aims to facilitate peer-to-peer correction cannot simply assume that users will act like disembodied rational agents. It must be designed to actively mitigate the social costs of dissent. The goal of such a system is not just to make correction possible, but to make it safe and practical. This moves beyond simply connecting users to information and toward designing affordances that manage social relationships and preserve social harmony—a far more complex, but essential, design goal.

5.3 A New Goal for Intervention Design: From Skill Transfer to Mindset Shift

Perhaps the most sobering, yet insightful, finding of our study is the utter failure of the intervention to produce generalizable fact-checking skills (Section 4.3). The flat post-survey accuracy scores across all arms are a stark reminder that short-term exposure to corrected facts does not turn a novice into an expert. While popular media literacy strategies like prebunking are premised on the acquisition of transferable skills, our findings challenge this entire paradigm by demonstrating a clear failure of such generalization [9].

However, where the intervention failed to transfer skills, it succeeded in catalyzing a profound mindset shift. The consistent emergence of a "deliberative pause" across our interviews represents a fundamental change in behavior—a move from reflexive, passive consumption to active, conscious inquiry. This shift toward deliberation is the core goal of other interventions like accuracy nudges or "think before you share" prompts [39]. Yet, where those interventions act as lightweight, external triggers to prompt a moment of reflection, our study suggests that a sustained, content-based intervention can help internalize this pause, transforming it into a more durable metacognitive habit. This shift is arguably a more valuable and durable outcome than the ability to debunk a specific category of fake image. It transformed participants from mere consumers into active agents who began to feel a sense of responsibility for the health of their community's information ecosystem, some even taking on the role of public correctors.

This finding aligns with and extends a significant body of HCI research focused on promoting user reflection through the intentional design of system interactions. Scholars have long explored how "design frictions" can create microboundaries to encourage more mindful engagement [17] and how interface nudges can be used to foster greater deliberativeness in online discourse [34, 53]. Our observation of an emergent "deliberative pause" resonates with this tradition. However, where much of this work focuses on using lightweight, external interface elements to prompt a momentary pause, our study offers a complementary perspective. We show that a sustained, socially-embedded informational intervention can help internalize this deliberative mindset, potentially transforming a prompted action into a more durable cognitive habit.

5.4 Implications for Design: Scaffolding Community Resilience

Acknowledging these complexities requires moving beyond generic recommendations like "invest in local fact-checking." While necessary, such investment is insufficient if it doesn't address the core socio-technical challenges of distribution, trust, and user labor. We propose a design agenda focused on scaffolding community resilience.

1. Design and Fund a "Community Caregiver" Program for Trusted Messengers. The failure of impersonal tiplines and the rarity of organic peer correction are not user failings; they are outcomes of a system that ignores social context and expects uncompensated labor. Our findings argue for a new model that professionalizes and supports the work of trusted community members, akin to the highly successful community health worker model in public health.

This requires platforms to design and fund a "Community Caregiver" program. Rather than building yet another tool for individual end-users, this approach focuses on empowering the trusted opinion leaders and information gateways—the doctors, teachers, and local leaders who already hold epistemic jurisdiction [2, 25]. The centerpiece would be a specialized "caregiver dashboard." Imagine a local doctor receiving a feed of professionally verified health fact-checks from institutional partners. The tool would allow her to not just forward it, but to easily translate it into the local dialect and, crucially, add a personal 15-second voice note ("I've reviewed this and it's important for our village"). She could then disseminate this augmented message to her existing patient groups with a single tap. This design pattern combines the rigor of institutional fact-checking with the trusted license of a community expert, creating a far more potent intervention than either could achieve alone. Importantly, such community centric interventions have been shown to work in prior work [12].

Critically, this must be treated as compensated labor. A platform-sponsored program could provide stipends, data packages, and ongoing training, transforming informal community support into a sustainable, professionalized role. It is only through such institutional investment that this model can scale beyond isolated pilots and become a meaningful, systemic solution.

2. Build an Intelligent Verification Assistant to Mitigate Friction and Scaffold Learning. While the "Community Caregiver" program addresses structural issues, there is a parallel need to support individual users in the moments they experience doubt. Our study shows that the intervention fostered a deliberative pause, but it failed to transfer generalizable skills, leaving a critical gap between intent and action. The act of verification is currently a high-friction, multi-app chore, and the act of correction is socially risky.

To address this, we propose an Intelligent Verification Assistant, an AI-powered feature integrated directly into the messaging client [1]. A key technical enabler for such a feature is the recent advancement in small language models [52]. To balance advanced capabilities with the non-negotiable need for privacy, this assistant would be designed to operate primarily on-device. Simple heuristic checks, text analysis, and pattern matching for known misinformation could run locally, fully respecting end-to-end encryption. As on-device models continue to improve in efficiency and power, more sophisticated analyses could progressively transition from privacy-preserving server-side checks to running completely on the user's device.

This assistant would function not merely as a "suite of tools" like a button for reverse image search [36]. Instead, it would act as a contextual tutor to scaffold learning at the moment of need. For example, upon receiving a message with an AI-generated image, the assistant could privately prompt the user: "This image has features common in AI-generated content. Would you like a 30-second guide on how to spot them?" This approach directly addresses the skill-generalization failure by providing targeted micro-learning that is immediately relevant.

Furthermore, this assistant must be designed to mitigate social friction. When a user identifies a message as false, instead of only offering a "forward" button, the assistant could provide options to reduce the social cost of correction. It could offer a pre-written, diplomatic template for a private reply to the sender, or facilitate an "Anonymous Corroboration" feature, where sharing a fact-check attaches a system-generated wrapper ("A verified source has noted this claim is inaccurate") that decouples the corrective act from the user's personal identity. By making the tools easier, the social process safer, and the interaction educational, such an assistant can transform the high-friction, high-risk act of verification into a seamless, safe, and empowering learning moment.

5.5 Limitations

This study has several limitations that offer clear avenues for future research. First, our sample size (N=51) is small and drawn from a specific region in India, which limits the generalizability of our quantitative claims. We must acknowledge that operationalizing a field experiment with multiple stakeholders (journalists, surveyors, participants) and live, real-world data is inherently difficult to conduct at a large scale. This study, therefore, serves as a crucial pilot, providing a strong proof of concept for the community-based model.

Second, our intervention was short-term. While we observed a promising and significant shift toward a "deliberative pause," longitudinal studies are needed to determine if this metacognitive change is durable over time. It remains an open question whether this nascent habit of critical inquiry would persist without the continued presence of the intervention, or if it would fade over time.

Finally, the operationalization of our T_3 (close contact) arm contains an important nuance. The "peers" delivering corrections were trained surveyors leveraging their existing social networks, not just any friend or family member of the participant. While this was a necessary design choice for logistical feasibility and to ensure consistent intervention delivery, it introduces a potential confound. The observed effectiveness of the T_3 arm may be a result not only of the pre-existing social tie, but also of the professional and articulate nature of the surveyors themselves. Future work could disentangle these factors by comparing the effects of a trained surveyor to those of an untrained but equally close social contact.

6 Conclusion

Combating misinformation in high-trust, encrypted spaces like WhatsApp is not primarily a technical problem of content moderation, nor is it a simple pedagogical problem of teaching individuals abstract skills. Our findings demonstrate that it is a profoundly socio-technical challenge that requires interventions to be woven into the existing social fabric. We found that while exposure to fact-checks did not create expert fact-checkers, it did foster a more cautious and deliberative mindset. The most effective path to changing beliefs was through a known, trusted contact who could deliver a clear, reasoned explanation. The central contribution of this work is to provide empirical evidence for a new model—one that moves away from impersonal, top-down systems and toward empowering trusted community members to build informational resilience from the ground up.

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] Leah Hope Ajmani, Jasmine C Foriest, Jordan Taylor, Kyle Pittman, Sarah Gilbert, and Michael Ann Devito. 2024. Whose Knowledge is Valued? Epistemic Injustice in CSCW Applications. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–28.

[3] Jennifer Allen, Cameron Martel, and David G. Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3502040

[4] Sacha Altay, Simge Andı, Sumitra Badrinathan, Camila Mont'Alverne, Benjamin Toff, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. The small effects of short user corrections on misinformation in Brazil, India, and the United Kingdom. *Harvard Kennedy School (HKS) Misinformation Review* 6, 3 (2025). https://doi.org/10.37016/mr-2020-179

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

[6] Sumitra Badrinathan. 2021. Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review* 115, 4 (2021), 1325–1341.

[7] Sumitra Badrinathan and Simon Chauchard. 2023. "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics* 29, 2 (2023), 394 – 416. https://doi.org/10.1177/19401612231158770

[8] 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).

[9] Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Ünal, William P. McClanahan, and Sander van der Linden. 2021. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data Soc.* 8, 1 (2021), 205395172110138. https://doi.org/10.1177/20539517211013868

[10] Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak. 1964. Measuring Utility by a Single-Response Sequential Method. *Behavioral Science* 9, 3 (1964), 226–232. https://doi.org/10.1002/bs.3830090304

[11] Leticia Bode and Emily K. Vraga. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication* 33, 9 (2018), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

[12] Jeremy Bowles, Horacio Larreguy, and Shelley Liu. 2020. Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe. *PloS one* 15, 10 (2020), e0240005.

[13] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22. https://doi.org/10.1145/3134666

[15] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* 42, 4 (2020), 1073–1095.

[16] Juliet M. Corbin and Anselm L. Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4 ed.). SAGE Publications, Thousand Oaks, CA. Chapter 8 provides a step-by-step guide to open coding..

[17] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397.

[18] 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).

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

[20] 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.

[21] Government of Telangana, Planning Department; Directorate of Economics & Statistics. 2024. *Telangana Socio Economic Outlook 2024*. Technical Report. Government of Telangana. https://www.des.telangana.gov.in/publications/Socio%20Economic%20Outlook-2024.pdf Placed in the State Legislature during the budget session..

[22] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545.

[23] 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.

[24] Yasna Haghdoost. 2020. Alcohol Poisoning Kills 100 Iranians Seeking Virus Protection. *Bloomberg Markets* (2020). https://www.bloomberg.com/news/articles/2020-03-18/alcohol-poisoning-kills-100-iranians-seeking-virus-protection

[25] Farnaz Jahanbakhsh, Amy X Zhang, and David R Karger. 2022. Leveraging structured trusted-peer assessments to combat misinformation. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW2 (2022), 1–40.

[26] Shagun Jhaver, Eshwar Chandrasekharan, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Transactions on Computer-Human Interaction* 29, 4 (2022), 1–26. https://doi.org/10.1145/3490499

[27] Pranay Juneja and Tanushree Mitra. 2022. Human and Technological Infrastructures of Fact-Checking: Challenges Faced by Fact-Checkers in India. In *Proceedings of the ACM on Human-Computer Interaction (CSCW2)*, Vol. 6. 1–36. https://doi.org/10.1145/3555128

[28] 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).

[29] Loukas Konstantinou and Evangelos Karapanos. 2025. Behavior Change Interventions Combating Online Misinformation: A Scoping Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 336, 19 pages. https://doi.org/10.1145/3706598.3713127

[30] Tuan-He Lee and Susan R. Fussell. 2025. Countering Misinformation in Private Messaging Groups: Insights From a Fact-checking Chatbot. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP10 (Jan. 2025), 30 pages. https://doi.org/10.1145/3701189

[31] Xingyu Liu, Li Qi, Laurent Wang, and Miriam J Metzger. 2025. Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research* 52, 6 (2025), 719–746.

[32] Jayson L. Lusk and Jason F. Shogren. 2007. *Experimental Auctions: Methods and Applications in Economic and Marketing Research*. Cambridge University Press, Cambridge.

[33] 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. https://doi.org/10.1080/21670811.2023.2213731 arXiv:https://doi.org/10.1080/21670811.2023.2213731

[34] Sanju Menon, Weiyu Zhang, and Simon T Perrault. 2020. Nudge for deliberativeness: How interface features influence online discourse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[35] Meta Platforms, Inc. 2025. *WhatsApp Business API*. https://developers.facebook.com/docs/whatsapp Official documentation.

[36] Times of India. 2024. *WhatsApp Web to introduce reverse image search feature: Here's what it is*. https://timesofindia.indiatimes.com/technology/social/whatsapp-web-to-introduce-reverse-image-search-feature-heres-what-it-is/articleshow/116799311.cms Describes WhatsApp's integration of a reverse image search button to help users identify fake photos and combat misinformation..

[37] Saumya Pareek and Jorge Goncalves. 2024. Peer-supplied credibility labels as an online misinformation intervention. *International Journal of Human-Computer Studies* 188 (2024), 103276.

[38] Irene V Pasquetto, Eaman Jahani, Shubham Atreja, and Matthew Baum. 2022. Social debunking of misinformation on WhatsApp: the case for strong and in-group ties. *Proceedings of the ACM on human-computer interaction* 6, CSCW1 (2022), 1–35.

[39] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[40] Nicolas Pröllochs. 2022. Community-Based Fact-Checking on Twitter's Birdwatch Platform. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, Vol. 16. 794–805. https://doi.org/10.1609/icwsm.v16i1.19335

[41] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*. 818–828.

[42] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5 (2019), 65. https://doi.org/10.1057/s41599-019-0279-9

[43] Jon Roozenbeek and Sander van der Linden. 2020. Breaking Harmony Square: A game that "inoculates" against political misinformation. *Harvard Kennedy School (HKS) Misinformation Review* 1, 8 (2020). https://doi.org/10.37016/mr-2020-47

[44] Ananya Seelam, Arnab Paul Choudhury, Connie Liu, Miyuki Goay, Kalika Bali, and Aditya Vashistha. 2024. "Fact-checks are for the Top 0.1%": Examining Reach, Awareness, and Relevance of Fact-Checking in Rural India. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 56 (April 2024), 34 pages. https://doi.org/10.1145/3637333

[45] Gautam Kishore Shahi and Scott A Hale. 2025. WhatsApp tiplines and multilingual claims in the 2021 Indian assembly elections. *Online Social Networks and Media* 49 (2025), 100323.

[46] 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.

[47] Farhana Shahid, Shrirang Mare, and Aditya Vashistha. 2022. Examining Source Effects on Perceptions of Fake News in Rural India. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 89 (April 2022), 29 pages. https://doi.org/10.1145/3512936

[48] Huiyun Tang, Gabriele Lenzini, Samuel Greiff, Björn Rohles, and Anastasia Sergeeva. 2024. "Who Knows? Maybe it Really Works": Analysing Users' Perceptions of Health Misinformation on Social Media. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1499–1517. https://doi.org/10.1145/3643834.3661510

[49] Toni G. L. A. van der Meer and Yan Jin. 2020. Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source. *Health Communication* 35, 5 (2020), 560–575. https://doi.org/10.1080/10410236.2019.1573295

[50] 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.

[51] Kim Verhoeven, Steve Paulussen, and Gert-Jan de Bruijn. 2024. Genre Conventions of Fact-Checks: Topics, Style, and Form of Fact-Checking in Two IFCN-Certified News Media. *Journalism Practice* (2024), 1–17.
[52] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088* (2024).
[53] ShunYi Yeo, Zhuoqun Jiang, Anthony Tang, and Simon Tangi Perrault. 2025. Enhancing Deliberativeness: Evaluating the Impact of Multimodal Reflection Nudges. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

A Appendix: Supplementary Material

Table 6. Demographics of participants

ID	Age	Gender	Education	Occupation	Religion	Arm
P1	26	Male	Senior secondary	Self Employee	Hindu	Control
P2	22	Female	Bachelor (university 1st)	Unemployed	Hindu	T1
P3	23	Female	Bachelor (university 1st)	Unemployed	Hindu	Control
P4	26	Male	Bachelor (university 1st)	Unemployed	Hindu	T2
P5	33	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	Control
P6	38	Male	Upper primary	Company Employee	Hindu	T1
P7	25	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	Control
P8	22	Male	Bachelor (university 1st)	Self Employee	Hindu	T3
P9	20	Male	Senior secondary	Self Employee	Hindu	T3
P10	30	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	T1
P11	32	Male	Bachelor (university 1st)	Company Employee	Hindu	T2
P12	25	Male	High school level	Skilled Labor	Hindu	T3
P13	23	Male	Senior secondary	Company Employee	Hindu	Control
P14	34	Male	Nursing, general nursing and midwifery (GNM)	Company Employee	Hindu	T2
P15	31	Male	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	T1
P16	21	Female	Bachelor (university 1st)	Self Employee	Hindu	Control
P17	20	Male	Senior secondary	Unemployed	Hindu	T3
P18	55	Male	Bachelor (university 1st)	Entrepreneur or Business Owner	Hindu	T3
P19	21	Male	Senior secondary	Self Employee	Hindu	T2
P20	37	Female	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	T3
P21	31	Male	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	Control
P22	33	Male	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	Control
P23	23	Male	Technical education training	Self Employee	Hindu	Control
P24	35	Male	Bachelor (university 1st)	Self Employee	Hindu	T3
P25	30	Male	Senior secondary	Self Employee	Muslim	T2
P26	39	Female	Master (university 2nd) / Post-graduate diploma	Unemployed	Hindu	T3
P27	29	Female	Master (university 2nd) / Post-graduate diploma	Self Employee	Hindu	T2
P28	27	Male	Master (university 2nd) / Post-graduate diploma	Unemployed	Hindu	T2
P29	26	Male	Bachelor (university 1st)	Unemployed	Hindu	T3
P30	35	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	T2
P31	30	Male	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	T1
P32	29	Male	Bachelor (university 1st)	Company Employee	Hindu	T3
P33	30	Female	Master (university 2nd) / Post-graduate diploma	Unemployed	Hindu	T3
P34	42	Male	Bachelor (university 1st)	Company Employee	Hindu	T2
P35	30	Female	Bachelor (university 1st)	Self Employee	Hindu	T2
P36	25	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	Control
P37	32	Male	Senior secondary	Farmer or Agricultural Worker	Hindu	T2
P38	28	Male	Bachelor (university 1st)	Farmer or Agricultural Worker	Hindu	T3
P39	34	Male	Master (university 2nd) / Post-graduate diploma	Self Employee	Hindu	Control

Continued on next page

ID	Age	Gender	Education	Occupation	Religion	Arm
P40	28	Male	Technical education training	Company Employee	Hindu	T1
P41	22	Female	Bachelor (university 1st)	Self Employee	Hindu	T1
P42	36	Male	Master (university 2nd) / Post-graduate diploma	Company Employee	Hindu	T2
P43	28	Male	Technical education training	Company Employee	Hindu	T1
P44	26	Male	Bachelor (university 1st)	Unemployed	Hindu	T3
P45	32	Male	Bachelor (university 1st)	Company Employee	Hindu	T1
P46	45	Male	Bachelor (university 1st)	Company Employee	Hindu	T2
P47	22	Male	Senior secondary	Company Employee	Hindu	T1
P48	24	Male	High school level	Self Employee	Hindu	T1
P49	48	Male	Bachelor (university 1st)	Self Employee	Hindu	Control
P50	20	Male	Senior secondary	Unemployed	Hindu	T1
P51	34	Female	Bachelor (university 1st)	Company Employee	Hindu	T1

Table 7. BJP favorability (post − pre) by arm

arm	n_pairs	mean_pre	mean_post	mean_diff (post - pre)	ci95_lo	ci95_hi	p_value
Control	10	1.600	1.400	-0.200	-0.764	0.364	0.443
T1	13	1.385	1.615	0.231	-0.207	0.669	0.273
T2	13	2.000	1.462	-0.538	-1.123	0.046	0.068
T3	13	1.462	1.154	-0.308	-0.598	-0.017	0.040

Table 8. Congress favorability (post − pre) by arm

arm	n_pairs	mean_pre	mean_post	mean_diff (post - pre)	ci95_lo	ci95_hi	p_value
Control	9	1.778	1.889	0.111	-0.351	0.573	0.594
T1	13	2.154	2.231	0.077	-0.721	0.875	0.837
T2	13	2.000	1.692	-0.308	-0.824	0.209	0.219
T3	13	2.077	2.077	0.000	-0.552	0.552	1.000

Table 9. Change in favorability toward religious groups (post − pre) by arm (4-point scale; lower = more favorable)

Religion	Arm	n_{pairs}	Mean(pre)	Mean(post)	Δ	95% CI for Δ	Wilcoxon p	r
Hindus	Control	11	1.000	1.000	0.000	[0.000, 0.000]	—	0.000
	T1	13	1.077	1.154	0.077	[-0.221, 0.375]	0.564	0.333
	T2	12	1.083	1.083	0.000	[0.000, 0.000]	—	0.000
	T3	13	1.385	1.077	-0.308	[-0.762, 0.146]	0.169	-0.667
Muslims	Control	9	2.222	2.000	-0.222	[-0.863, 0.418]	0.518	-0.500
	T1	11	2.182	2.091	-0.091	[-0.725, 0.543]	0.886	-0.143
	T2	13	1.846	1.846	0.000	[-0.427, 0.427]	1.000	0.000
	T3	13	2.615	2.308	-0.308	[-0.880, 0.265]	0.272	-0.400
Sikhs	Control	5	2.200	1.800	-0.400	[-1.511, 0.711]	0.317	-0.500
	T1	8	1.750	2.000	0.250	[-0.491, 0.991]	0.512	0.500
	T2	7	2.000	2.000	0.000	[-0.534, 0.534]	1.000	0.000
	T3	13	2.000	2.077	0.077	[-0.091, 0.245]	0.317	1.000

Notes. Favorability is on a 4-point scale (1=Very favorable, 2=Somewhat favorable, 3=Somewhat unfavorable, 4=Very unfavorable); lower values are more favorable. Δ denotes the paired mean difference (post − pre). 95% CIs are for Δ. Wilcoxon p from signed-rank test; r is rank–biserial. "—" indicates the test was undefined due to all-zero differences.

A.1 Operationalization of WTP

We measured WTP for a *one-week, priority fact-checking service* that was identical and excludable for others delivered directly to the participant. Each person received an endowment of INR 100, stated their maximum price on a simple grid $\{0, 10, \ldots, 100\}$, and an enumerator drew one price uniformly at random from the same grid (operationallized using a bag of paper chits containing these grid values) in view of the participant. If the drawn price $p \leq b$, they purchased the service and paid p from the endowment (keeping $100 - p$); otherwise they kept the full INR 100. Enumerators used a short script and two quick comprehension questions before running the binding draw.

Table 10. Accuracy by belief-check stimulus (all items, including letter-suffixed IDs) with arm-level coverage and mistakes

QID	Theme(s)	*n*	Acc. (mean [95% CI])	Err.	Correct	Mistakes	T1 (ans/mis)	T2 (ans/mis)	T3 (ans/mis)	Ctrl (ans/mis)
1	religion	46	0.500 [0.361, 0.639]	0.500	23	23	12/8	11/6	12/2	11/7
2	health related	46	0.348 [0.227, 0.492]	0.652	16	30	12/11	11/7	12/5	11/7
3	health related	46	0.543 [0.402, 0.678]	0.457	25	21	12/5	11/4	12/3	11/9
4	generic	46	0.413 [0.283, 0.557]	0.587	19	27	12/8	11/4	12/6	11/9
5	generic	46	0.261 [0.156, 0.403]	0.739	12	34	12/10	11/8	12/7	11/9
6	health	44	0.341 [0.219, 0.489]	0.659	15	29	12/9	10/5	12/6	10/9
7	generic	44	0.477 [0.338, 0.621]	0.523	21	23	12/8	10/4	12/5	10/6
8	anti inc, anti muslim, religion	44	0.432 [0.297, 0.578]	0.568	19	25	12/9	10/5	12/5	10/6
9a	health related	1	1.000 [0.207, 1.000]	0.000	1	0	0/0	0/0	1/0	0/0
9b	health related	1	1.000 [0.207, 1.000]	0.000	1	0	0/0	0/0	1/0	0/0
9c	religion	1	0.000 [0.000, 0.793]	1.000	0	1	0/0	1/1	0/0	0/0
9d	generic	1	1.000 [0.207, 1.000]	0.000	1	0	0/0	0/0	1/0	0/0
9e	generic	1	0.000 [0.000, 0.793]	1.000	0	1	1/1	0/0	0/0	0/0
9f	health related	1	1.000 [0.207, 1.000]	0.000	1	0	1/0	0/0	0/0	0/0
10	generic	41	0.268 [0.157, 0.419]	0.732	11	30	11/7	10/8	11/7	9/8
11	religion	40	0.350 [0.221, 0.505]	0.650	14	26	11/8	10/5	10/7	9/6
12	health related	40	0.425 [0.285, 0.578]	0.575	17	23	11/6	10/4	10/7	9/6
13	generic, health related	38	0.289 [0.170, 0.448]	0.711	11	27	11/8	10/5	8/5	9/9
14	health related	38	0.316 [0.191, 0.475]	0.684	12	26	11/9	10/5	8/4	9/8
15	nationalist	37	0.432 [0.287, 0.591]	0.568	16	21	10/7	10/5	8/3	9/6
16	anti inc	37	0.351 [0.218, 0.512]	0.649	13	24	10/6	10/4	8/5	9/9
17	anti muslim, religion	37	0.378 [0.241, 0.539]	0.622	14	23	10/5	10/6	8/4	9/8
18	pro bjp	37	0.432 [0.287, 0.591]	0.568	16	21	10/6	10/5	8/5	9/5
19	nationalist	36	0.333 [0.202, 0.497]	0.667	12	24	9/7	10/5	8/4	9/8
20	anti muslim	36	0.361 [0.225, 0.524]	0.639	13	23	9/6	10/5	8/4	9/8
21	anti muslim	36	0.361 [0.225, 0.524]	0.639	13	23	9/7	10/6	8/5	9/5
22	anti inc, anti muslim	36	0.278 [0.158, 0.440]	0.722	10	26	9/8	10/5	8/4	9/9
23	generic	35	0.343 [0.208, 0.508]	0.657	12	23	9/6	10/6	7/3	9/8
24	health related, pro hindu	35	0.400 [0.256, 0.564]	0.600	14	21	9/5	10/5	7/4	9/7
25	anti inc, pro hindu	35	0.343 [0.208, 0.508]	0.657	12	23	9/6	10/4	7/5	9/8
26	anti ycp	35	0.457 [0.305, 0.618]	0.543	16	19	9/4	10/7	7/3	9/5
27	generic	35	0.429 [0.280, 0.591]	0.571	15	20	9/5	10/4	7/4	9/7

Notes. Items are the actual belief-check stimuli; seven participants received personalized stimuli (letter-suffixed IDs such as 9a–9f), which are included here. Not all participants answered each item; n is the number of responses for that row. Accuracy is the fraction of correct answers; Error rate = 1 − accuracy. 95% CIs are for mean accuracy (normal approximation). Arm columns show answered/mistakes per arm (T1, T2, T3, Control). Kruskal–Wallis test across questions on correct_bin: $k = 33$ groups, $N = 1065$, $H = 30.957$x, $p = 0.519$ (not significant). Abbreviations: INC - Indian National Congress (Party), BJP - Bharatiya Janata Party, YCP - YSR Congress Party

Notes. Mean accuracy is the fraction correct on each item (0–1). Error rate = 1 − mean accuracy (also shown as a percentage). 95% CIs are for the mean accuracy (normal approximation); bounds may slightly exceed [0,1] due to approximation. Themes were supplied by the researcher for each item index.

Table 11. Theme-level accuracy and mistakes (sorted by highest error rate)

Theme	***n***	**Accuracy (mean [95% CI])**	**Error rate**	**Correct**	**Mistakes**	**T1 (ans/mis)**	**T2 (ans/mis)**	**T3 (ans/mis)**	**Ctrl (ans/mis)**
anti inc	152	0.355 [0.284, 0.434]	0.645	54	98	40/29	40/18	35/19	37/32
civil awareness	287	0.355 [0.302, 0.412]	0.645	102	185	77/53	72/39	70/37	68/56
anti muslim	189	0.365 [0.300, 0.436]	0.635	69	120	49/35	50/27	44/22	46/36
pro hindu	70	0.371 [0.268, 0.489]	0.629	26	44	18/11	20/9	14/9	18/15
nationalist	73	0.384 [0.281, 0.498]	0.616	28	45	19/14	20/10	16/7	18/14
health related	290	0.390 [0.335, 0.447]	0.602	113	177	79/53	72/35	71/34	68/55
religion	168	0.417 [0.345, 0.492]	0.583	70	98	45/30	42/23	42/18	39/27
pro bjp	37	0.432 [0.287, 0.591]	0.568	16	21	10/6	10/5	8/5	9/5
anti ycp	35	0.457 [0.305, 0.618]	0.543	16	19	9/4	10/7	7/3	9/5

Notes. Accuracy is the fraction correct (0–1) for each theme; Error rate = 1 − accuracy. 95% CIs are for the mean accuracy (normal approximation). Arm columns show answered/mistakes per arm to indicate coverage and error distribution. Kruskal–Wallis test across themes on `correct_bin`: $k = 11$ groups, $N = 1301$, $H = 4.276$, $p = 0.934045$ (not significant).

Table 12. Participant-level accuracy per arm for post-survey misinformation discernment (GRADED scores)

Arm	n	Mean	sd	se	ci_lo	ci_hi
Control	12	0.532	0.089	0.026	0.475	0.589
T1	13	0.491	0.105	0.029	0.451	0.534
T2	12	0.512	0.105	0.030	0.445	0.579
T3	13	0.514	0.078	0.022	0.467	0.561

Table 13. Welch tests for comparison of Treatment arms vs Control for post-survey misinformation discernment (GRADED scores)

Arm	Control	t	*p_raw*	*p_holm*	*p_fdr*
T1	Control	-1.2311	0.2321	0.6962	0.6250
T2	Control	-0.4960	0.6250	1.0000	0.6250
T3	Control	0.5480	0.5892	1.0000	0.6250

Any-treated vs Control (GRADED): $t = -0.8965$, $p = 0.3822$.

Table 14. Top movers in BJP favorability measured before and after the study(Post − Pre), 4-point scale (lower = more favorable)

Participant ID	**Arm**	**Baseline party affiliation**	**Religion**	**Pre**	**Post**	**Δ**	**Direction**
P30	T2	Congress Party / Indian National Congress (INC)	Hindu	3.0	1.0	-2.0	more_favorable
P28	T2	Congress Party / Indian National Congress (INC)	Hindu	4.0	2.0	-2.0	more_favorable
P19	T2	Bharatiya Janata Party (BJP)	Hindu	3.0	1.0	-2.0	more_favorable
P1	Control	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P49	Control	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P40	T1	Telangana Rashtra Samithi (TRS)	Hindu	2.0	1.0	-1.0	more_favorable
P39	Control	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P37	T2	Bharatiya Janata Party (BJP)	Hindu	2.0	1.0	-1.0	more_favorable
P33	T3	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P24	T3	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P17	T3	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P14	T2	Bharatiya Janata Party (BJP)	Hindu	2.0	1.0	-1.0	more_favorable
P51	T1	Congress Party / Indian National Congress (INC)	Hindu	3.0	2.0	-1.0	more_favorable
P7	Control	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P32	T3	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable

Notes. Listed participants have the largest absolute change in BJP favorability (top 15). $\Delta < 0$ indicates movement toward *more favorable* (since lower values mean more favorable on a 1–4 scale).

Table 15. Top movers toward *less* BJP favorability measured before and after the study (Post − Pre), 4-point scale (lower = more favorable)

Participant ID	Arm	Baseline party affiliation	Religion	Pre	Post	Δ	Direction
P15	T1	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P41	T1	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P21	Control	Telangana Rashtra Samithi (TRS)	Hindu	1.0	2.0	1.0	less_favorable
P25	T2	Congress Party / Indian National Congress (INC)	Muslim	2.0	3.0	1.0	less_favorable
P50	T1	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P43	T1	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P10	T1	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P23	Control	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P13	T2	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P26	T3	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P44	T3	Bharatiya Janata Party (BJP)	Hindu	1.0	2.0	1.0	less_favorable
P29	T3	Bharatiya Janata Party (BJP)	Hindu	1.0	1.0	0.0	no_change

Notes. Listed participants show the largest absolute change in BJP favorability in the *less favorable* direction. $\Delta > 0$ indicates movement toward less favorable (since lower values are more favorable on the 1–4 scale).

Table 16. Top movers in INC favorability measured before and after the study (Post − Pre), 4-point scale (lower = more favorable)

Participant ID	Arm	Baseline party affiliation	Religion	Pre	Post	Δ	Direction
P4	T2	Bharatiya Janata Party (BJP)	Hindu	4.0	2.0	-2.0	more_favorable
P18	T3	Bharatiya Janata Party (BJP)	Hindu	4.0	2.0	-2.0	more_favorable
P40	T1	Telangana Rashtra Samithi (TRS)	Hindu	3.0	1.0	-2.0	more_favorable
P41	T1	Bharatiya Janata Party (BJP)	Hindu	4.0	2.0	-2.0	more_favorable
P14	T2	Bharatiya Janata Party (BJP)	Hindu	3.0	2.0	-1.0	more_favorable
P34	T2	Congress Party / Indian National Congress (INC)	Hindu	3.0	2.0	-1.0	more_favorable
P19	T2	Bharatiya Janata Party (BJP)	Hindu	3.0	2.0	-1.0	more_favorable
P17	T3	Congress Party / Indian National Congress (INC)	Hindu	2.0	1.0	-1.0	more_favorable
P11	T2	Bharatiya Janata Party (BJP)	Hindu	2.0	1.0	-1.0	more_favorable
P50	T1	Bharatiya Janata Party (BJP)	Hindu	4.0	3.0	-1.0	more_favorable
P49	Control	Congress Party / Indian National Congress (INC)	Hindu	3.0	2.0	-1.0	more_favorable
P2	T1	Bharatiya Janata Party (BJP)	Hindu	3.0	2.0	-1.0	more_favorable
P44	T3	Bharatiya Janata Party (BJP)	Hindu	3.0	2.0	-1.0	more_favorable
P42	T2	Bharatiya Janata Party (BJP)	Hindu	2.0	2.0	0.0	no_change
P29	T3	Bharatiya Janata Party (BJP)	Hindu	2.0	2.0	0.0	no_change

Notes. Participants are ordered by the absolute change $|\Delta|$ in INC favorability, where Δ = Post − Pre. On the 4-point scale (1=very favorable, 4=very unfavorable), lower values indicate greater favorability; hence $\Delta < 0$ denotes movement toward *more favorable* to INC.

Table 17. Top movers toward *less* INC favorability measured before and after the study (Post − Pre), 4-point scale (lower = more favorable)

Participant ID	Arm	Baseline party affiliation	Religion	Pre	Post	Δ	Direction
P10	T1	Bharatiya Janata Party (BJP)	Hindu	2.0	4.0	2.0	less_favorable
P6	T1	Bharatiya Janata Party (BJP)	Hindu	3.0	4.0	1.0	less_favorable
P24	T3	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P32	T3	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P17	T3	Bharatiya Janata Party (BJP)	Hindu	2.0	3.0	1.0	less_favorable
P39	Control	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P32	T3	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P30	T2	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P46	T2	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P7	Control	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P47	T1	Congress Party / Indian National Congress (INC)	Hindu	1.0	2.0	1.0	less_favorable
P44	T3	Bharatiya Janata Party (BJP)	Hindu	2.0	2.0	0.0	less_favorable
P42	T2	Bharatiya Janata Party (BJP)	Hindu	2.0	2.0	0.0	less_favorable
P29	T3	Bharatiya Janata Party (BJP)	Hindu	2.0	2.0	0.0	less_favorable
P33	T3	Congress Party / Indian National Congress (INC)	Hindu	1.0	1.0	0.0	no_change

Notes. Listed participants exhibit the largest movements away from INC (positive Δ indicates becoming *less* favorable to INC). Ordering is by Δ (magnitude first, then as shown). Scale: 1=very favorable, 4=very unfavorable (lower = more favorable).

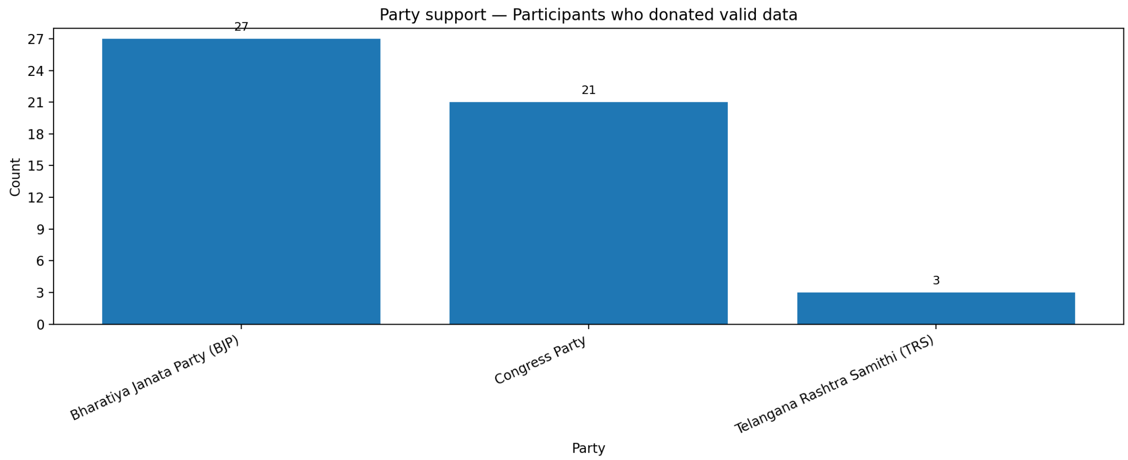

Fig. 3. Baseline party affiliation among **treated participants** (N=51). Bars show counts by self-reported party; y-axis is integer counts.

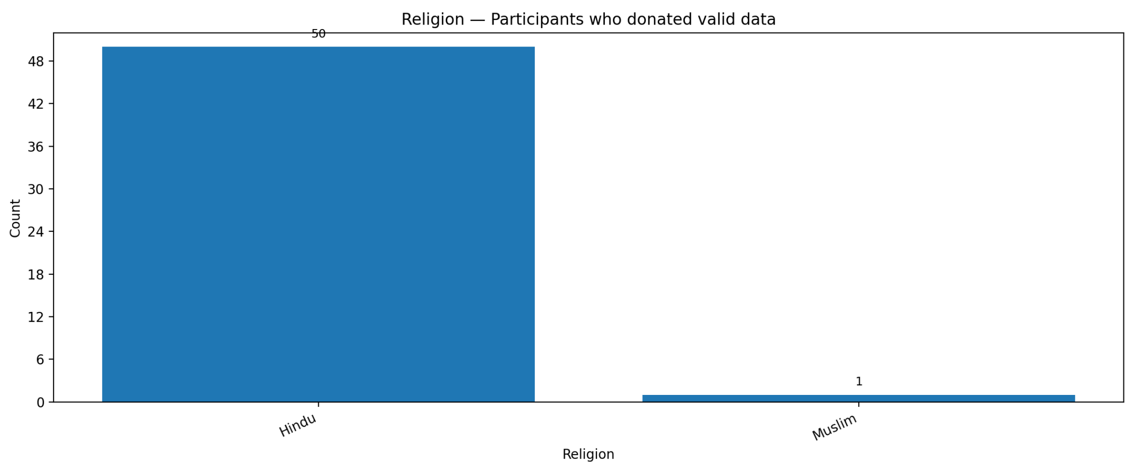

Fig. 4. Self-reported religion among **treated participants** (N=51). Bars show counts by religion.

Table 18. Post-Survey Question-level accuracy and mistakes by theme (sorted by highest error rate)

Question No.	Themes	True Veracity	n	Accuracy (mean [95% CI])	SD	Correct	Mistakes	Error rate
9	political	False	49	0.204 [0.087, 0.321]	0.407	10	39	0.796 (79.6%)
4	religion; entertainment	False	52	0.212 [0.097, 0.326]	0.412	11	41	0.788 (78.8%)
7	religious; ai generated	False	52	0.269 [0.145, 0.394]	0.448	14	38	0.731 (73.1%)
6	religious	False	52	0.308 [0.178, 0.437]	0.466	16	36	0.692 (69.2%)
8	ai generated; generic	False	51	0.333 [0.199, 0.467]	0.476	17	34	0.667 (66.7%)
5	sports; nationalist	False	52	0.404 [0.266, 0.542]	0.495	21	31	0.596 (59.6%)
10	generic	True	52	0.596 [0.458, 0.734]	0.495	31	21	0.404 (40.4%)
3	sports	True	51	0.863 [0.765, 0.960]	0.348	44	7	0.137 (13.7%)
2	sports; nationalist	True	52	0.885 [0.795, 0.974]	0.323	46	6	0.115 (11.5%)
1	health	True	52	0.981 [0.942, 1.019]	0.139	51	1	0.019 (1.9%)

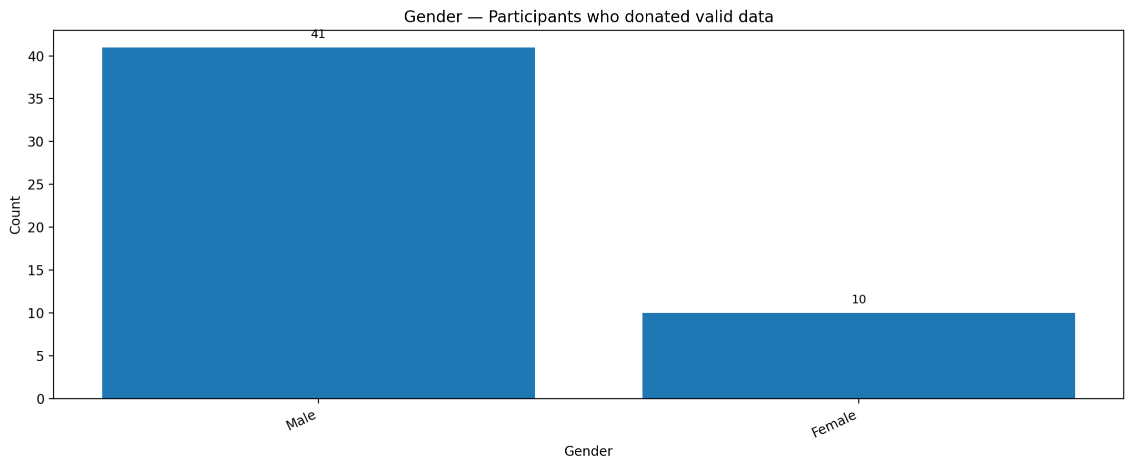

Fig. 5. Gender composition among **treated participants**. Bars show counts by reported gender.

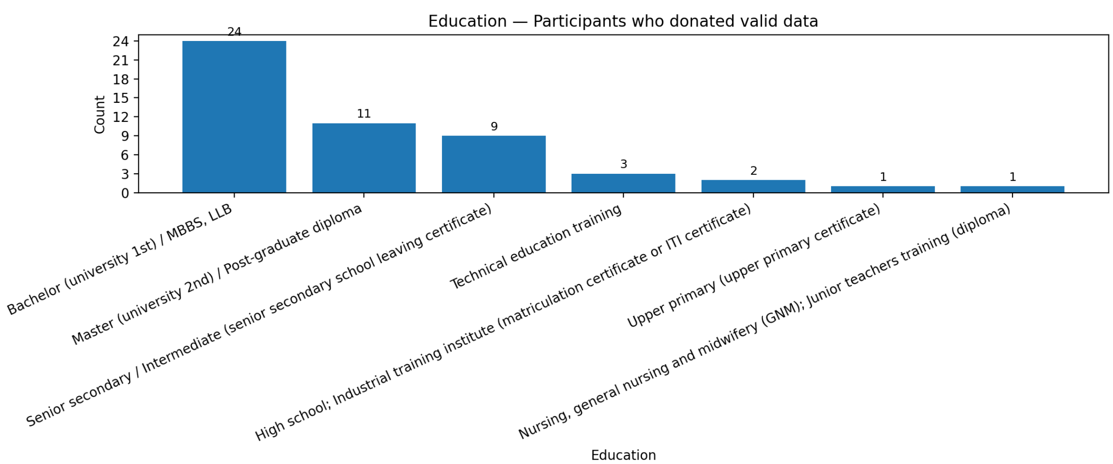

Fig. 6. Education levels among **treated participants**. Bars show counts by highest reported education.

Table 19. Post-survey Misinformation Discernment Theme-level accuracy and mistakes (sorted by highest error rate)

Theme	n	Accuracy (mean [95% CI])	SD	Correct	Mistakes	Error rate
political	49	0.204 [0.087, 0.321]	0.407	10	39	0.796 (79.6%)
entertainment	52	0.212 [0.097, 0.326]	0.412	11	41	0.788 (78.8%)
religion	156	0.263 [0.193, 0.333]	0.442	41	115	0.737 (73.7%)
ai generated	103	0.301 [0.211, 0.391]	0.461	31	72	0.699 (69.9%)
generic	103	0.466 [0.368, 0.564]	0.501	48	55	0.534 (53.4%)
nationalist	104	0.644 [0.551, 0.738]	0.481	67	37	0.356 (35.6%)
sports	155	0.716 [0.644, 0.788]	0.452	111	44	0.284 (28.4%)
health	52	0.981 [0.942, 1.019]	0.139	51	1	0.019 (1.9%)

Notes. Mean accuracy is the fraction correct per theme. Error rate = 1 − mean accuracy (also shown as a percentage). 95% CIs use a normal approximation and may slightly exceed [0,1] at the upper bound. Kruskal–Wallis across themes on `correct_bin`: $H = 170.677$, $p < 10^{-6}$.

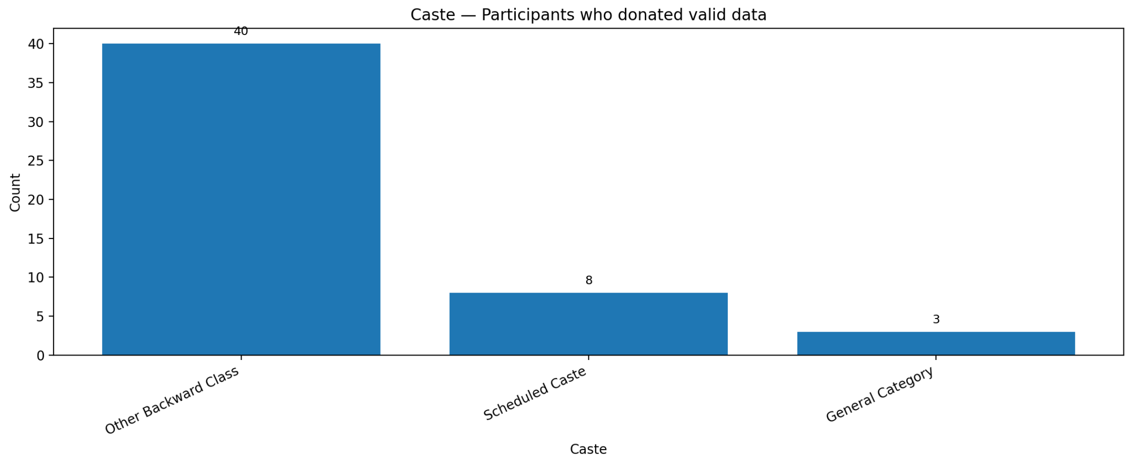

Fig. 7. Caste categories among **treated participants**. Bars show counts by self-reported caste category.

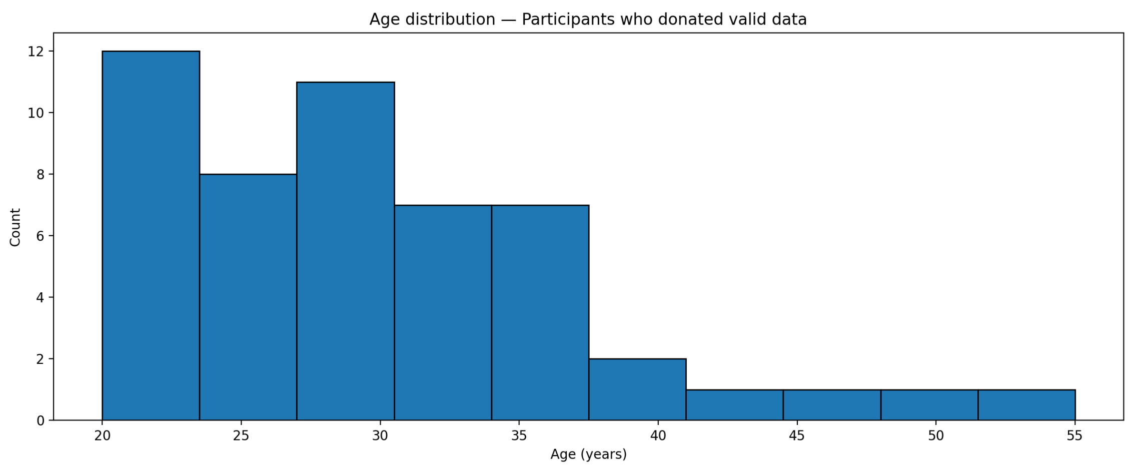

Fig. 8. Age distribution among **treated participants**. Histogram shows counts by age bins.

Table 20. Welch tests comparing each arm to Control on post-survey WTP measures.

Arm	n_{trt}	n_{ctrl}	Mean(trt)	Mean(ctrl)	Diff	t	p_{raw}	p_{Holm}	p_{FDR}	Cohen's d (unpooled)
T1	13	12	66.923	62.500	4.423	0.395	0.697	0.765	0.697	0.159
T2	12	12	81.667	62.500	19.167	1.903	0.074	0.221	0.221	0.804
T3	13	12	72.308	62.500	9.808	0.892	0.383	0.765	0.574	0.361

Notes. Welch two-sample t-tests (unequal variances) comparing each treatment arm to Control. Diff = Mean(trt) − Mean(ctrl). p_{Holm} = Holm–Bonferroni; p_{FDR} = Benjamini–Hochberg. No comparison is significant after Holm correction at $\alpha = 0.05$.

Fig. 9. Occupation distribution among **treated participants**. Histogram shows counts by age bins.

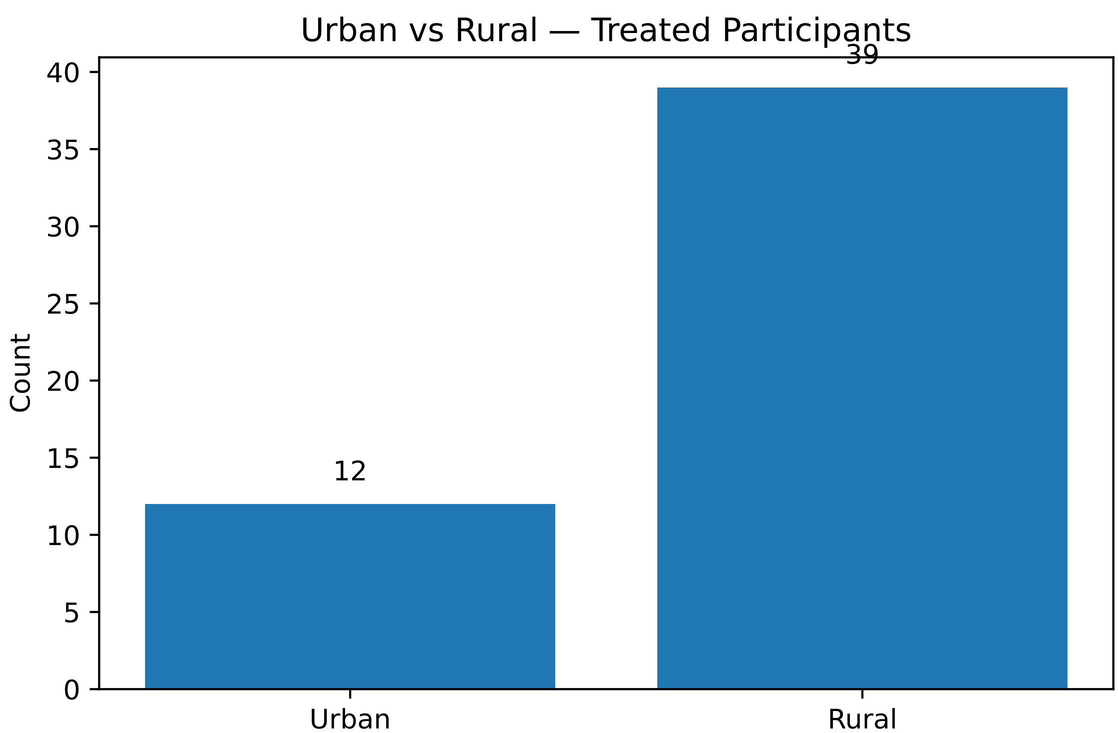

Fig. 10. Urban vs. Rural composition of the inhabitation of our participants.

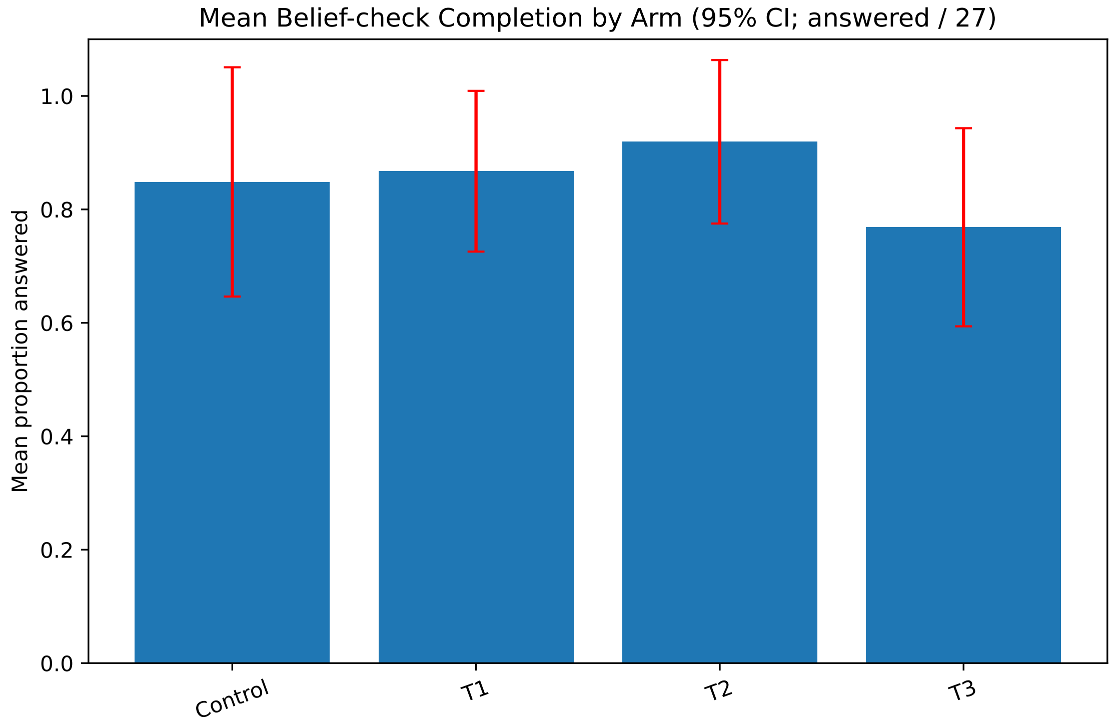

Fig. 11. Outcome completion by treatment arm: mean completion with 95% confidence intervals.

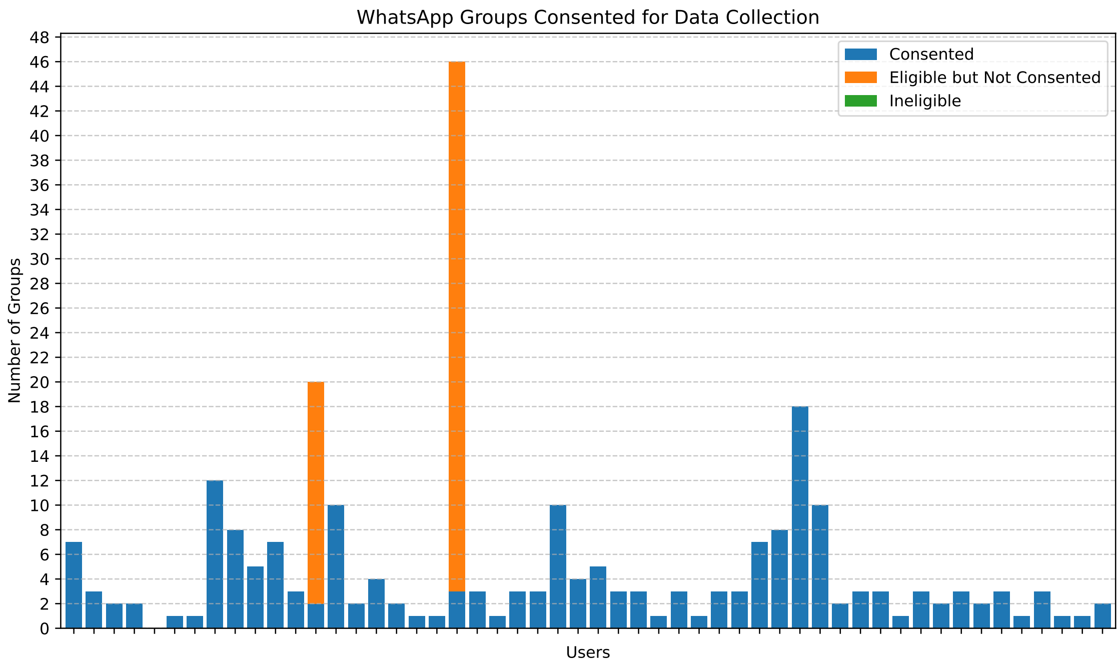

Fig. 12. Histogram of user consented and donated groups for each user