Generative AI and Political Discourse: Insights from WhatsApp in Rural India

Kiran Garimella

Abstract

This paper investigates the spread and nature of generative AI content in WhatsApp groups in rural India, focusing on its impact on political discourse. Utilizing a dataset from over 1,000 WhatsApp groups, encompassing more than one million pieces of content, we concentrate on 2,000+ items which were virally spreading on WhatsApp. Our analysis reveals a minimal presence of generative AI content, primarily images and videos. The AI content identified predominantly features deceptive material, propaganda, politically charged memes, erotic videos, and inspirational messages. Despite its limited prevalence, the study highlights the importance of understanding generative AI's role in political communication, especially given the Bharatiya Janata Party's (BJP) dominant influence on content dissemination via WhatsApp. Our findings argue against the perception of generative AI as a primary tool for electoral influence, emphasizing the greater impact of strategic content distribution. This research underscores the need for vigilant approaches to emerging technologies and warns against compromising encryption as a defense against AI-generated content, advocating for global and domestic strategies to mitigate potential misuses in the public sphere.

1 Introduction

This study examines the dissemination and characteristics of content generated through generative AI on WhatsApp in rural India. WhatsApp is the biggest end to end encrypted social network with over 2 billion monthly active users, including over 600 million in India. The data was collected from over 1,000 WhatsApp groups using an innovative, opt-in approach resulting in a diverse sample of 400 users from three villages in central and northern India. Our unique dataset includes over one million pieces of content, out of which we specifically focus on over 2,000 items flagged by WhatsApp as 'forwarded many times' between August and October 2023, indicative of their viral nature on the platform.

Our analysis of the content revealed that, despite the rarity of groups dedicated to political discussion, political matters predominated the viral content. A manual annotation was conducted to identify generative AI-generated content across text, images, audio, and video. Our findings suggest a minimal presence of generative AI creations, with fewer than a dozen instances, primarily images or videos, and no detectable generative AI-generated texts. The lack of generative AI text may be attributed to the nascent stage of Hindi-language AI content generation or the inherent difficulty in distinguishing AI-generated text.

The identified generative AI content fell into three main categories: deceptive material (e.g., futuristic infrastructure images misrepresented as current Indian projects), pro-Hindu propaganda (e.g. muscular Hindu guys, anti opposition memes), and other miscellaneous users (e.g. soft porn, and inspirational messages). Some examples of the different types of generative AI content found in our dataset can be found in Figures 1,2,3. Despite the seemingly low prevalence of generative AI content in our sample, the potential for wider prevalence and impact remains uncertain due to the non-representative nature of the dataset.

The study posits that generative AI is not a predominant tool for electoral influence in India. The Bharatiya Janata Party's (BJP, the right wing hindu majoritarian party currently in power) established infrastructure for content dissemination on WhatsApp overshadows the need for generative AI technologies. The true potency of generative AI lies not in content creation ease but in its amplification alongside existing political resources. The vast impact of social media is driven more by strategic dissemination than by straightforward content generation abetted by generative AI, as evidenced by the BJP's existing capabilities. Nonetheless, understanding generative AI's influence remains critical,

as it could provide other parties, especially new entrants, with a competitive edge. The role of encryption is particularly noteworthy, as it can obscure the spread of AI-generated content until potentially widespread dissemination.

This research underscores the necessity for a vigilant approach to emerging technologies in the information ecosystem, emphasizing the need for global and domestic strategies to address potential misuses in the public sphere. Our findings caution against suggestions, particularly by governments, to compromise or weaken encryption as a measure against the perceived threats posed by generative AI. They contribute to the dialogue on generative AI's role in shaping political narratives and the urgent need for preparedness against its misuse.

2 Background and Data

2.1 Background

WhatsApp, a widely used messaging platform, is characterized by its end-to-end encryption feature, ensuring that messages are only accessible to the sender and the recipient. This level of privacy, while beneficial for user security, also means that there is no content moderation in place, allowing for the unregulated spread of information, including misinformation and propaganda. A significant aspect of WhatsApp's impact in information dissemination is its extensive use by the BJP, one of India's major political parties. The BJP has established a vast and influential ecosystem within WhatsApp, utilizing it as a key tool for spreading political content and propaganda. This strategy includes the distribution of messages, images, and videos that align with the party's agenda and viewpoints. The use of WhatsApp by the BJP is not only extensive but also sophisticated. There is documented evidence of coordinated posting within this network, as highlighted by Jakesch et al. (2021) [JGEN21]. This coordination often extends to other social media platforms, such as Twitter. By synchronizing messages across a network of WhatsApp groups, the BJP can influence Twitter trends and set the agenda on a broader scale. This orchestrated approach allows for a concerted and impactful spread of the party's messages and propaganda, leveraging the private and unmoderated nature of WhatsApp to reach a wide audience with minimal external scrutiny.

In this context, the rapid development and increasing significance of generative AI add another layer of complexity. Generative AI, known for its ability to create realistic and convincing digital content, is evolving swiftly, becoming more accessible and sophisticated. The potential impact of this technology on electoral processes has been a topic of considerable discussion and concern. [SSB23] has highlighted the possible implications of generative AI in elections, underscoring the need for vigilance and regulatory measures. However, it's important to note that not all predictions about the impact of generative AI have materialized. [DJ21] pointed out instances where anticipated consequences did not occur as expected [DJ21]. This discrepancy between forecasted and actual outcomes emphasizes the dynamic and unpredictable nature of technological influence in political spheres. It suggests a need for ongoing analysis and adaptation in understanding and responding to the role of advanced technologies like generative AI in shaping political discourse and public opinion.

Academic consensus on the role of generative AI is still evolving, marked by divergent viewpoints. One perspective, as [Fer23] notes, is concerned with how generative AI democratizes content creation, potentially empowering bad actors to perpetrate a wide range of harms. This viewpoint emphasizes the risks posed by the ease of creating convincing misinformation and the variety of malicious uses that might emerge. Conversely, another viewpoint argues that the novelty of generative AI tools might not significantly amplify existing threats. Proponents of this view point to the capabilities of existing tools to inflict similar damage and stress the importance of platforms' distribution mechanisms in ensuring safety. Furthermore, they argue that the benefits of generative AI could substantially outweigh the potential harms, cautioning against a disproportionate focus on the downsides that could lead to biased policy outcomes [KN23, SAM23].

These contrasting perspectives highlight the ongoing debate and complexity surrounding generative AI. Both sides present valid arguments, and the reality of the situation may indeed be hard to discern. This ambiguity underscores the need for careful, continuous evaluation of generative AI's impact, particularly in sensitive domains like political communication and public discourse.

2.2 Dataset

The dataset underpinning this study is derived from WhatsApp groups, specifically focusing on content circulated in Uttar Pradesh in the north of India. The data was obtained through a process of data donation, where approximately 400 users voluntarily contributed their WhatsApp data. This sample, though a convenience sample, was selected with an eye towards demographic diversity to ensure a broad representation of content types and user backgrounds. In adhering to stringent privacy standards, the collection process was designed to be privacy-preserving. No personal information of the participants or their contacts was stored or analyzed. Instead, the focus was on the content of the messages, particularly identifying viral content—defined as content that was forwarded multiple times within the network. From this process, we compiled a dataset of approximately 2,000 pieces of content. A significant portion of this content was in Hindi, reflecting the predominant language of communication among the sampled users. The dataset was carefully curated to ensure comprehensive coverage. Each piece of content was manually examined, minimizing the likelihood of missing any relevant or widely circulated generative AI pieces.

In our study, the sampling methodology plays a crucial role in understanding the landscape of viral content on WhatsApp. By focusing on content flagged as 'forwarded many times', we ensured that our dataset captured messages that achieved widespread circulation within the network. This approach is pivotal in gauging the reach and impact of any particular type of content, including generative AI-generated material. The rationale behind this method is straightforward: if a piece of content has become viral, it is likely to appear in our dataset due to its extensive distribution among users. This sampling strategy allows us to confidently assert that widely spread content, especially those that have resonated significantly with users, is represented in our analysis. Consequently, this method provides a comprehensive view of the viral landscape on WhatsApp, ensuring that our findings are reflective of the content that has truly gained traction within the platform's vast user base.

3 Narratives

We categorized the content into specific narrative categories. We present a summary of the narratives below. All the raw videos and images can be found on our website.[1]

3.1 Misleading content

We identified instances of generative AI content to mislead audiences. A notable example includes the circulation of AI-generated images falsely claiming to depict a new train station in Ayodhya, as shown in Figure 1. These images, crafted with high fidelity to reality, exemplify the potential of generative AI to blur the boundaries between truth and fabrication, posing significant challenges for information veracity in the digital age. In a more interactive misuse of generative AI, an instance was recorded where such an image was incorporated into a video with added music.[2]

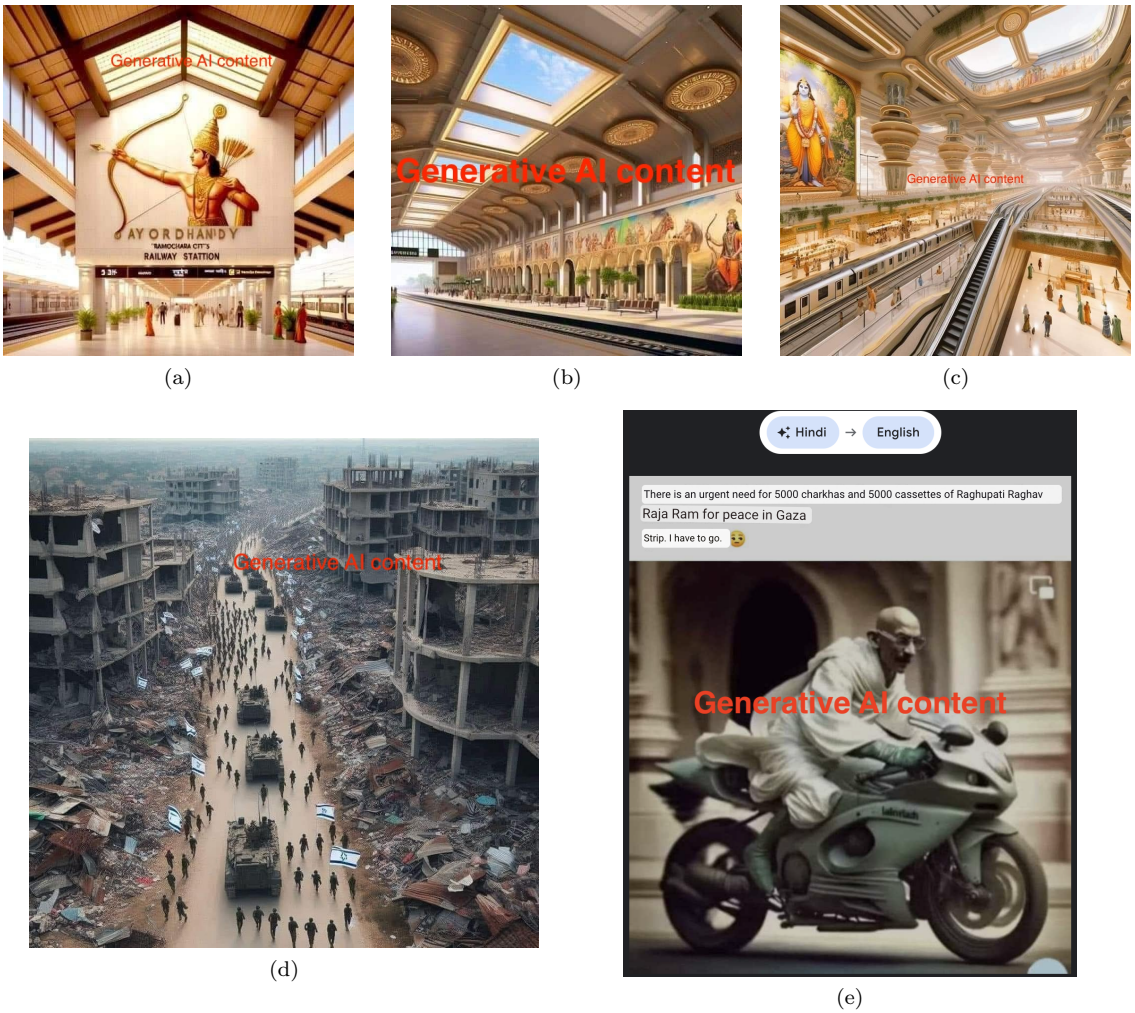
Further complicating the landscape of AI-generated misinformation is the sharing of images out of context, particularly for susceptible audiences. An instance involved a synthetic image of Indian Prime Minister Narendra Modi, erroneously represented as a natural phenomenon occurring in Fiji. This example highlights instances of generative AI which might be 'easy' to detect for digitally literate audiences, but may not be the case for many first time internet users, which makes up a majority of WhatsApp's user base in India. Similarly, Figure 1(d) the generation and dissemination of an image showing the Israeli military in Gaza, another product of AI manipulation, underscore the broader geopolitical ramifications. Even though a lot of content about Gaza was just content that was shared out of context (e.g. images from the Syrian war), we still see some content which could for instance be used to indicate specific aspects, such as the dominance of the Israeli military in the battlefield.

We also found an interesting example generative AI in meme creation, where its implications are both culturally significant and complex. A notable instance is a meme that satirizes Mahatma Gandhi's advocacy for non-violence (see Figure 1 (e)), specifically regarding the Palestinian struggle. This meme utilizes an AI-generated image depicting Gandhi in an anachronistic and incongruous scenario, where

[1] https://kiran-research2.comminfo.rutgers.edu/generativeAI/

[2] See full video here http://analysis-backend.whats-viral.me/videos/fe4c47d7-025e-4864-8c18-bca630bc5dd0.mp4?platform=whatsapp

he is humorously portrayed rushing to provide peaceful music and cotton spinning machines to the people of Gaza. The use of generative AI in this context matters profoundly because it enables the creation of visually compelling and culturally resonant imagery that can quickly resonate with and influence the perceptions of a wide audience. While memes are often viewed as vehicles of humor or satire, their ability to shape discourse and influence public opinion, especially when amplified by the realism and accessibility of AI-generated content, is significant. This example reflects the broader impact of generative AI in shaping cultural narratives and highlights the need for critical engagement with such content in understanding its broader societal implications.



(a) (b) (c)

(d) (e)

Figure 1: Examples of misleading generative AI content found in our dataset. (a–c) Images claiming to be the new train station in India. (d) Misleading: AI generated image of Israeli military moving through Gaza. (e) This is a meme mocking the non-violence struggle Mahatma Gandhi suggested for Palestine. It uses an AI generated image where Gandhi is supposed to be racing to provide peaceful music and cotton spinning machines to the people of Gaza. Note that the text has been translated from Hindi.

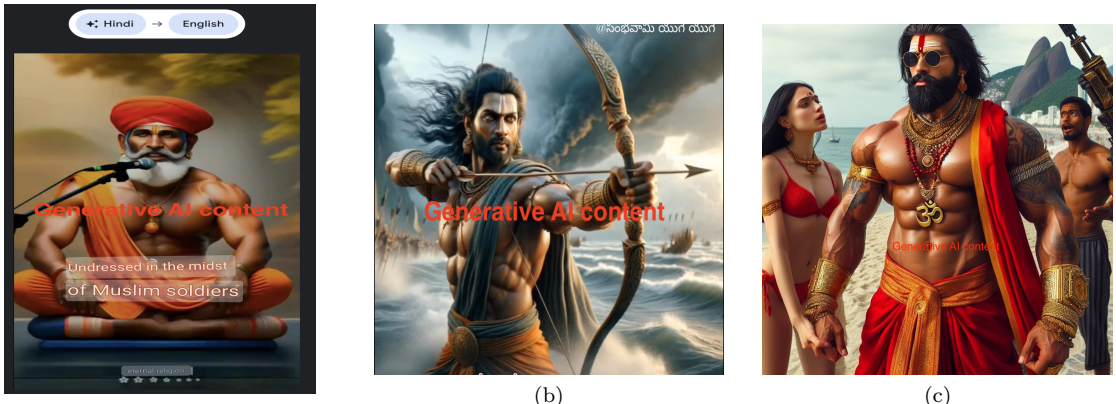
3.2 Projecting Hindu supremacy

The use of generative AI in projecting Hindu supremacy manifests notably through the depiction of muscular men and gods, symbolizing Hindu power (Figure 2). This trend involves creating AI-generated videos that often include hate speech. These videos typically feature imagery such as muscular monks, with lip-syncing AI technology animating these figures. Tools like `https://gooey.ai/Lipsync` facilitate this process by requiring only an image and an audio file. However, the creation

of these videos still demands considerable effort in generating the images, audio, and music. This form of content frequently aligns with fear speech narratives, a prevalent form of hate speech identified in recent studies [SGK+23]. An illustrative example is a video screenshot showing a muscular Hindu monk uttering hateful remarks against Muslims, referencing historical injustices by Muslim rulers. Another instance within this genre is a video of a woman expressing grievances over perceived preferential treatment towards Muslims can be found here.

The portrayal of Hindu gods and figures in a highly muscular, heroic form is another facet of this phenomenon (Figures 2 (b,c). These images are not only easy to generate using AI but have been previously used by organizations like Sangh Parivar for propaganda purposes, showcasing Hindu supremacy [Kis90, Kau02]. While not inherently nefarious, these representations often serve as powerful tools for propagating certain ideological narratives, reinforcing the perception of Hindu dominance and strength.

This particular use case highlights the dual use of generative AI in promoting religious and cultural supremacy, emphasizing the need to understand the broader socio-political implications of such content. These AI-generated representations, ranging from hate speech to heroic imagery, may play a significant role in shaping and reinforcing cultural and religious narratives, underscoring the importance of critically engaging with and understanding the impact of AI-generated content in socio-political contexts.



(a) (b) (c)

Figure 2: Examples of Generative AI content found in our dataset. (a) Hate speech: A screenshot of a video showing a muscular Hindu monk saying hateful stuff against Muslims invoking historic injustices done to Hindus by Muslim rulers. (b,c) Depictions of deities and Hindu men with robust physiques symbolizing heroism and assertiveness.

3.3 Other examples

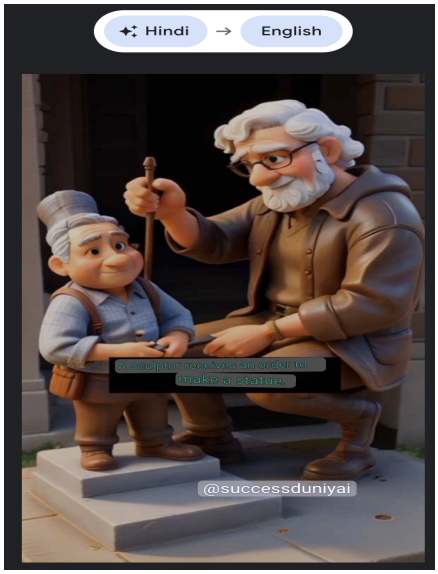
We also see examples of uses in softer domains, such as soft pornography and inspirational videos. An example includes a screenshot from an AI-generated video featuring a woman uttering erotic statements (Figure 3 (a)). The use of AI in creating such content reflects its versatility and the ease with which it can cater to diverse, even controversial, content demands. Additionally, AI is utilized in crafting inspirational short videos, demonstrating the technology's potential in creating engaging and motivational content, as evidenced by a full-length video generated entirely through AI (Figure 3 (b)).

4 Discussion

The results above showed examples of creative uses of generative AI on WhatsApp. There is clear evidence that generative AI is not a huge threat vector, at least in terms of its volume and reach, which seems to be much smaller compared to other forms of viral, misleading and hateful content.

In discussing the intricacies of generative AI within the context of WhatsApp in rural India, it's imperative to dissect the production, consumption, distribution, and persuasion aspects of such content. The role of technology in each of these facets varies significantly, with production being relatively

(a)



(b)

Figure 3: Examples of generative AI content used for creative purposes, such as (a) soft porn, and (b) inspirational videos.

straightforward due to the advancements in AI capabilities. However, the distribution and persuasion aspects present more significant challenges.

Generative AI's role in production is clear – it enables the creation of diverse content types, from text to images and videos. This production is streamlined by AI technologies, making it increasingly accessible. On the other hand, consumption is directly influenced by the platforms that host this content. In the case of WhatsApp, the absence of content moderation uniquely intertwines production and consumption. Users can easily receive and propagate AI-generated content, creating a seamless link between these two processes. This lack of moderation, coupled with the platform's encrypted nature, makes it difficult to disrupt the flow of AI-generated content, potentially leading to widespread dissemination without oversight.

The concept of generating content for political influence is not novel, especially in the context of the Bharatiya Janata Party (BJP) in India. The BJP has long been recognized for its extensive network of IT cells across the country, adeptly producing content aimed at shaping public opinion and political narratives. This well-oiled machinery has been operational for over a decade, effectively utilizing human-driven efforts to create viral content. The introduction of generative AI tools into this already potent mix could mark a significant shift in content creation dynamics. While generative AI might simplify the process of content generation, its real impact lies in the potential for personalization and targeted messaging. This ability to tailor content to specific audiences or individuals could amplify the efficacy and reach of such campaigns, making the political use of generative AI a force multiplier for existing content generation strategies. In this context, the advent of generative AI represents not just a technological evolution, but a strategic enhancement to the already sophisticated content dissemination practices employed by political entities like the BJP.

Distribution and persuasion are where the challenges intensify. While generative AI facilitates the creation of content, distributing this content effectively to influence opinions or behaviors – particularly in a politically charged environment – is a complex task. This complexity is partly because the effectiveness of persuasion depends on numerous factors, including the content's credibility, the audience's receptiveness, and the context in which it is shared. Generative AI may not inherently aid in persuasion, as the persuasive power of content often hinges more on how it resonates with the audience's pre-existing beliefs and values rather than its mode of creation.

In the realm of generative AI, the dynamics of distribution and consumption on encrypted platforms like WhatsApp present unique challenges and potential harms. The lack of moderation inherent in these platforms enables a seamless convergence of production and consumption, creating an unregulated space where content can be distributed without oversight. This scenario is particularly concerning when

considering the persuasive power of generative AI. While the technology itself may not inherently aid in persuasion, the context in which AI-generated content is distributed and consumed can significantly amplify its persuasive impact. In an encrypted environment, the source and authenticity of content often remain obscured, allowing misleading or harmful AI-generated content to spread unchecked. This unchecked dissemination can lead to the reinforcement of false narratives, the deepening of social and political divides, and the manipulation of public opinion. The unbroken link between production and consumption on platforms like WhatsApp, devoid of moderating influences, thus raises serious concerns about the potential for misuse of generative AI in shaping perceptions and influencing behaviors in subtle yet profound ways.

However, generative AI can play a significant role in creative and positive ways. It can aid in generating educational content, personalized messaging for health awareness, or culturally relevant entertainment, thereby enhancing engagement and potentially leading to positive societal impacts. For instance, AI-generated content tailored to local languages and cultural contexts in rural India could enhance information accessibility and relevance. For instance, we already see a huge increase in generative AI content for devotional purposes [Sah23].

Moreover, understanding the dynamics of generative AI in such a unique environment as rural India's WhatsApp networks can provide valuable insights for developing AI governance and ethical guidelines. It's crucial to consider how these technologies can be harnessed for societal good while mitigating risks like misinformation or the erosion of public trust.

This discussion also opens avenues for future research, particularly in exploring how generative AI could evolve to become more effective in persuasion and distribution while ensuring ethical usage. As generative AI continues to develop, its impact on various facets of digital communication, including political discourse, will likely become more pronounced, necessitating ongoing scrutiny and adaptive strategies to maximize its benefits while minimizing potential harms.

The observation that generative AI content is not predominantly present in the viral WhatsApp content analyzed in our study does not diminish the potential threat it poses. History has shown that the harms of technology often manifest in ways that are not initially anticipated. Just as previous technological advancements have brought unforeseen challenges, the relatively nascent field of generative AI harbors similar potential for unintended consequences. The fact that it is currently not a significant element in viral content may be reflective of its early developmental stage, or a current lack of adaptation to local languages and contexts. However, as the technology evolves and becomes more accessible and sophisticated, its ability to influence and potentially disrupt social and political discourse could increase significantly. This possibility underscores the importance of proactive vigilance and the development of strategies to mitigate potential negative impacts. The current absence of generative AI content in viral messages should not be a reason for complacency, but rather a call to prepare for its possible future implications and to understand its capabilities more deeply.

Limitations. The methodology of our study, while comprehensive, acknowledges the possibility that some generative AI content may have been missed in our analysis. We only manually looked for generative AI content in images and videos. Important modalities of text and images have not been analyzed. As of early 2024, most generative AI content is relatively easy to detect due to certain identifiable characteristics. However, the landscape of AI technology is rapidly evolving, and there is a significant possibility that, in the future, generative AI content could become increasingly sophisticated and harder to distinguish from human-generated content. This advancement poses a challenge in maintaining the efficacy of content analysis techniques and underscores the need for continuous development of detection methods. As AI-generated content becomes more refined and indistinguishable, it could seamlessly blend into the digital discourse, making it more difficult to assess its prevalence and impact. This potential shift in the detectability of generative AI content is a critical consideration for future research and for the ongoing monitoring of AI's role in information dissemination, especially in contexts where the authenticity of content can significantly influence public opinion and societal dynamics.

The potential of generative AI to facilitate personalization in content creation is a significant aspect to consider, particularly in the context of our study focusing on viral and popular content on WhatsApp. Generative AI has the capability to tailor content to individual preferences, beliefs, and behaviors, making it a powerful tool for creating highly personalized messages. However, this aspect of generative AI might not be fully captured in our dataset, which centers on content that has achieved widespread circulation.

Our focus on viral content inherently filters out those AI-generated messages that are personalized for specific individuals or small groups, as such content is less likely to be widely forwarded and therefore less likely to appear in our dataset. This limitation raises the possibility that we might be missing a segment of AI-generated content that is more nuanced and targeted, yet less visible in the broader context of viral messaging.

While we acknowledge this gap, it's also arguable that the landscape of personalized AI-generated content is not yet fully developed, particularly in the context of rural India and the language and cultural nuances therein. As of early 2024, the technology for creating highly personalized, context-specific AI content at scale may still be in its infancy, especially in languages other than English. Therefore, while our dataset might miss some personalized AI-generated content, it's plausible that this type of content is not yet prevalent enough to significantly alter our study's findings.

while generative AI holds the potential for personalization, our study's focus on viral content means that we may be overlooking this aspect of AI-generated messages. However, given the current stage of technological development, it's likely that such personalized content is not yet a dominant factor in the digital communication landscape we are examining. Future studies, with methodologies designed to capture more personalized content, would be necessary to fully understand this dimension of generative AI's impact.

References

[DJ21] Nicholas Diakopoulos and Deborah Johnson. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7):2072–2098, 2021.

[Fer23] Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *arXiv preprint arXiv:2310.00737*, 2023.

[JGEN21] Maurice Jakesch, Kiran Garimella, Dean Eckles, and Mor Naaman. Trend alert: A cross-platform organization manipulated twitter trends in the indian general election. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW2):1–19, 2021.

[Kau02] Raminder Kaur. Martial imagery in western india: The changing face of ganapati since the 1890s. *South Asia: Journal of South Asian Studies*, 25(1):69–96, 2002.

[Kis90] Madhu Kishwar. In defence of our dharma. *Manushi*, 60(4), 1990.

[KN23] Sayash Kapoor and Arvind Narayanan. How to prepare for the deluge of generative ai on social media, 2023.

[Sah23] Deepak Sahi. The role of Generative AI in expanding a modern-age religion — linkedin.com. https://www.linkedin.com/pulse/role-generative-ai-expanding-modern-age-religion-deepak-sahi/, 2023. [Accessed 01-03-2024].

[SAM23] Felix M Simon, Sacha Altay, and Hugo Mercier. Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, 4(5), 2023.

[SGK+23] Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120, 2023.

[SSB23] Kaylyn Jackson Schiff, Daniel S Schiff, and Natalia Bueno. The liar's dividend: The impact of deepfakes and fake news on trust in political discourse. 2023.