Milking the Metaphors: Cultural Obfuscation in Wellness (COW) around Gomutra on YouTube using LLMs

Anamta Khan*[1], Ratna Kandala*[2], Deepti[3], Sheza Munir[1], Joyojeet Pal[1]

[1] University of Michigan, [2]University of Kansas, [3]IIT Jodhpur

Abstract

Social media platforms have become primary channels for health information in the Global South. Using *gomutra* (cow urine) discourse on YouTube in India as a case study, we present a post-facto Large Language Model (LLM)-assisted discourse analysis of 30 multilingual transcripts showing that promotional content blends sacred traditional language with pseudo-scientific claims in ways that sophisticated debunking content itself mirrors, creating a rhetorical register that LLMs, trained predominantly on Western corpora, are systematically ill-equipped to analyse. Varying prompt tone across three LLMs (GPT-4o, Gemini 2.5 Pro, DeepSeek-V3.1), we find that culturally embedded health misinformation does not look like ordinary misinformation, and this cultural obfuscation extends to gendered rhetoric and prompt design, compounding analytical unreliability. Our findings argue that cultural competency in LLM-assisted discourse analysis cannot be retrofitted through prompt engineering alone.

Introduction

Nearly sixty-four percent of the global population actively uses social media (5.2 billion), with 491 million in India alone (Kemp 2025a). Among these, YouTube use in India alone has increased by 29 million (+6.3%) between January 2024 and 2025 (Kemp 2025b), making it one of the primary channels for health information, and increasingly health misinformation (Li et al. 2022), posing a growing threat to public health (Vosoughi, Roy, and Aral 2018; Rodrigues et al. 2024), especially when it is embedded in cultural tradition (Madathil et al. 2015) and endorsed by figures of authority (Chaturvedi 2024; State Council Information Office of the People's Republic of China 2020; Kelland and Satter 2020).

LLMs have shown genuine promise as tools for detecting and moderating health misinformation (Chen and Shu 2024; Alarabid 2025), with 17% of the U.S. adults now consulting health chatbots monthly (Kaiser Family Foundation 2024). Recent work has extended this to accessing health content quality on YouTube directly (Khalil, Mohamed, and Shoufan 2025). Yet this promise is unevenly distributed: LLM accuracy on health claim verification is systematically higher in English and European languages than in non-European ones, with performance varying substantially by topic and cultural context (Garg and Fetzer 2025). Because LLMs generate text through next-token prediction over training corpora, health misinformation embedded upstream can be surfaced and amplified without any corrective mechanism. This risk is compounded by a systematic evaluation gap: 95% of 519 LLM health evaluations conducted between 2022 and 2024 were in English, rarely assessing bias or fairness across other languages (Bedi et al. 2025), and where multilingual performance has been tested, correctness and consistency dropped markedly outside English (Jin et al. 2024; Garg and Fetzer 2025).

This limitation is especially acute on YouTube, where the informal, conversational format allows speakers to draw on tone, facial expression, and personal testimony to build audience trust in ways text-based platforms do not. These rhetorical resources become particularly powerful when health claims are framed through shared cultural and religious identity, especially in India, creating a category of misinformation that is not straightforwardly false but persuades through cultural authority.

One such claim centers on *gomutra* (cow urine), a substance that holds deep religious significance in Hinduism, where cows are considered to be sacred, and their byproducts - including cow dung, urine, and milk are traditionally regarded as purifying and antiseptic agents (Notermans 2019; Essar et al. 2021). The widespread adoption of cow urine as a remedy has been fueled by misinformation endorsed by politicians, authorities, and government bodies through media interviews, speeches, and official documents. A notable example is Indian Minister of State for Health Ashwini Choubey's 2019 public endorsement of cow urine as a basis for developing cancer treatments (Gulf News 2019; India Today 2019; Daria and Islam 2021; Asian News International 2022). These endorsements have lent institutional legitimacy to health claims that lack scientific support, making them substantially harder to counter - whether through debunking, regulatory intervention, or platform moderation (Fig 1).

Yet despite this scale and urgency, discourse on *gomutra*-related health claims remains significantly underexplored computationally. No annotated datasets exist, and no systematic linguistic analyses have been conducted. Existing

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

work on Indian health misinformation has focused largely on COVID-19 claims on WhatsApp (Essar et al. 2021) or on general misinformation diffusion patterns, without analyzing rhetorical structure. This exploratory study is a first attempt to address that gap directly.

We investigate how promoting and debunking content on *gomutra* health claims differ in their linguistic and rhetorical strategies- specifically, how traditional metaphors and culturally resonant terminology function differently from scientific framing - using LLMs as analytical instruments to examine these patterns across a corpus of 30 YouTube transcripts. We employ three LLMs (GPT-4o, Gemini 2.5 Pro, and DeepSeek-V3.1) by systematically varying prompt tone (formal vs. informal/conversational) and vocabulary (traditional terms such as "Sanjivani" vs. scientific terminology) to evaluate how sensitive these models' outputs are to cultural framing. In doing so, we contribute to the broader literature on computational discourse analysis and cultural NLP by reflecting critically on using LLMs as viable tools in culturally specific health misinformation contexts - attending to both their analytical affordances and their blind spots.

In this regard, we investigate the following research questions:

RQ1 How do linguistic and rhetorical strategies (i.e., deliberate choices in language use, framing, and discourse structure that serve communicative goals) differ between promotional and debunking *gomutra* content?

RQ2 How does prompt design, specifically tone, affect LLM sensitivity to linguistic markers of persuasion (i.e., lexical and rhetorical features associated with attitude change and influence) in *gomutra* content?

RQ3 What do LLM sensitivities to cultural framing, gender, and prompt design reveal about their reliability as instruments in culturally embedded health misinformation contexts?

Our findings reveal that promotional and debunking narratives deploy fundamentally different rhetorical strategies, and that LLMs' outputs are shaped by cultural framing, prompt design, and gendered rhetoric in ways that have direct implications for their use as discourse analytical instruments. This work sits at the intersection of NLP, health communication, and misinformation studies, and contributes from multiple angles: (a) empirically, a first attempt at an annotated video-transcript corpus of *gomutra* health claims; (b) analytically, a linguistic account of persuasion strategies in this domain; and (c) ethically, a set of methodological cautions for researchers using Western LLMs (Adilazuarda et al. 2024) as analytical instruments in ways that risk systematic misrepresentation of non-Western culturally embedded discourse, with direct implications for fairness in any downstream use of these tools.

Previous Work

Recent research on social media misinformation has increasingly focused on high-stakes domains such as health (Kong et al. 2021), where false claims pose a direct threat to societal wellbeing (Denniss and Lindberg 2025). Studies have also characterized the prevalence, spread, and psychological impact of health misinformation across these digital platforms (Van Der Linden 2022; Van der Linden et al. 2025). Existing literature reveals that health misinformation employs distinct persuasive strategies, such as fabricating narratives, politicizing public health issues, and misappropriating scientific evidence, to legitimize false claims (Peng, Lim, and Meng 2023).

While LLMs excel in standard evaluation frameworks, they face significant challenges when confronted with misinformation that is not straightforwardly false, but instead persuades through culturally resonant rhetoric, traditional authority, and shared identity. To deconstruct how such rhetoric operates, researchers increasingly rely on linguistic analysis. For example, topical and linguistic analyses of COVID-19 misinformation on YouTube have identified dominant themes centered around conspiracy theories and political dissemination (Thakur et al. 2024). As generative AI reshapes this landscape, recent evaluations of Chinese datasets have mapped specific linguistic features of misinformation, including distinct patterns in sentiment and cognitive framing, while simultaneously identifying the detection limits of standard LLMs (Ma et al. 2025). Despite these insights, the majority of LLM development and evaluation remains focused on Western contexts. Consequently, these models suffer from a pervasive lack of cultural awareness (Pawar et al. 2025), exhibiting inherent biases that severely limit their ability to parse culturally embedded rhetoric and implicit societal cues (Liu 2025).

This cultural limitation is particularly pronounced in the Indian context, where empirical evaluations demonstrate that LLMs consistently fail to accurately parse culture-specific traditions and regional dialects (Chhikara, Kumar, and Chakraborty 2025). Furthermore, these models frequently exhibit stereotypical biases and struggle to comprehend complex subcultures (Khandelwal et al. 2024). This specific cultural complexity in detecting misinformation is clearly visible in the health discourse surrounding *gomutra* (cow urine) in India. While recent literature highlights the spread of pseudoscience, narrative-based remedies, and the zoonotic risks associated with *gomutra* consumption during health crises like the COVID-19 pandemic (Essar et al. 2021; Hurford, Rana, and Sachan 2022), this is precisely the kind of culturally embedded, non-Western discourse that existing computational approaches are least equipped to handle, and yet it remains entirely uncharted without any datasets, linguistic analyses for this specific discourse. Furthermore, there are no existing works on LLM-based evaluations of the rhetorical strategies driving *gomutra*-related health misinformation in Indian languages, and no multilingual datasets exist in this regard.

This study addresses these critical gaps by introducing the first multilingual dataset on cow-urine discourse and providing a culturally aware, LLM-assisted discourse analysis of these traditional narratives, including sensitivity to prompting.

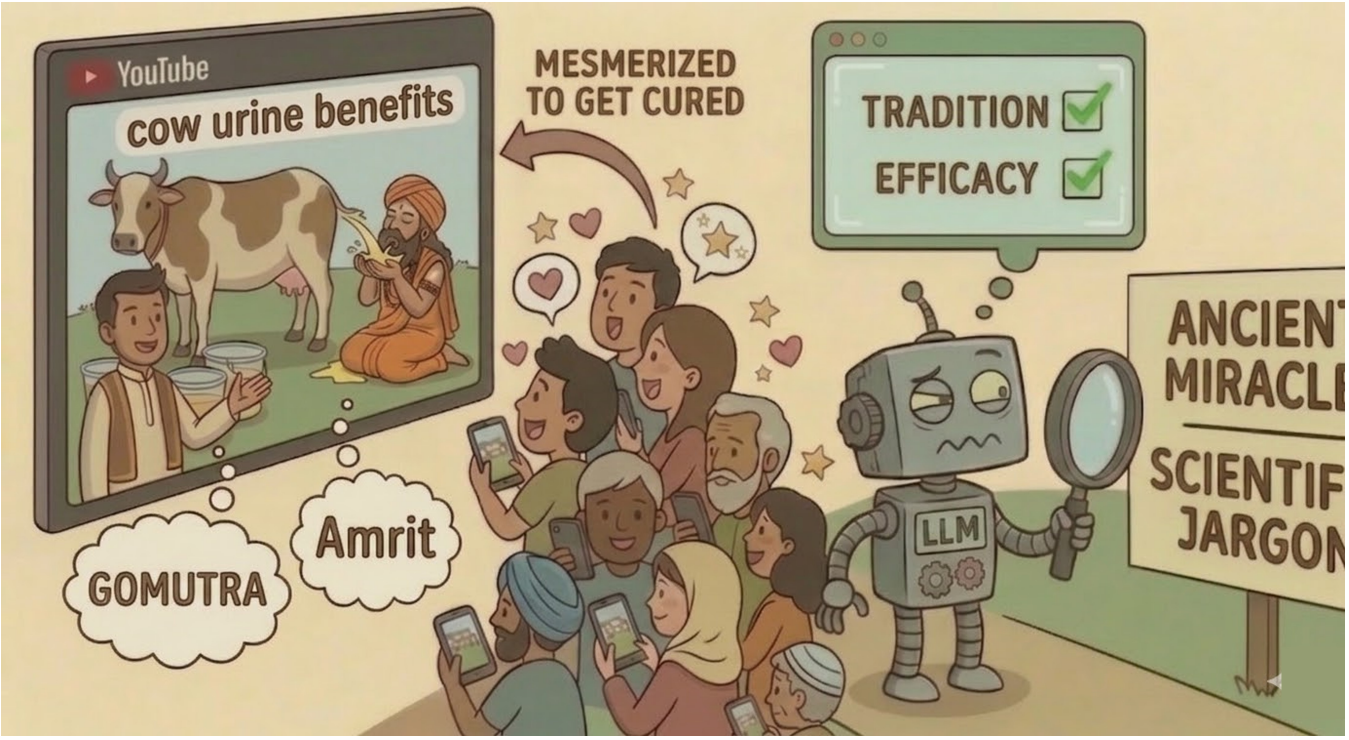

Figure 1: An infographic visualizing the 'Digital Proliferation of Traditional Health Beliefs' through cow urine (*gomutra*) remedies. The diagram contrasts the viral social media uptake against the challenge for Large Language Models (LLMs) to critically parse blended narratives that conflate ancient claims with pseudo-scientific terminology.

Methodology and Evaluation

Below, we present the methodology for data collection for our preliminary experiments and evaluation of the LLMs.

Dataset Collection and Curation

We collected YouTube videos related to *gomutra* discourse using the YouTube Data API v3. To capture relevance and linguistic variation, we queried the API using a combination of keywords and hashtags, including *gomutra*, cow urine, cowurine, *gaumutra*, *gaumutra* health, *gomutra* health, *#gomutra*, #gomata, and #cowurine. The search query returned videos in multiple languages (e.g., English, Hindi, Urdu, Kannada, Tamil, and Gujarati), which were curated via a two-step process. First, titles, descriptions, and audio were manually reviewed to ensure topic relevance, retaining only videos with primary spoken languages of English, Hindi, or Urdu. Second, after removing duplicates and derivative content, we relied on the API's default search relevance to finalize our curated subset of 30 videos for this preliminary exploratory study.

For each curated video, we extracted the audio track using the open-source yt-dlp library. To ensure strict compliance with the platform's terms of service, we do not store or redistribute raw audio files; all downstream processing relies exclusively on publicly accessible content for non-commercial research. Audio was transcribed using OpenAI's Whisper model (the large checkpoint) (Radford et al. 2022), selected for its state-of-the-art accuracy across multiple languages.

To evaluate transcription accuracy, we calculated the Word Error Rate (WER) on a randomly selected subset comprising 16% of the video corpus. A single author manually reviewed the machine-generated text against the human-generated reference transcript to compute the WER using the following standard formulation:

$$\text{WER} = \frac{S + D + I}{N}$$

where S represents the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript. The average WER across this sampled dataset was 7.04%, demonstrating a high level of transcription fidelity for our analysis. These non-English transcripts were translated into English using GPT-4 (OpenAI 2023) via the OpenAI API. Full details regarding transcription hyperparameters and translation prompt design are provided in the Appendix.

Finally, utilizing predefined written guidelines, a single author manually annotated each video to determine both the speaker's gender and the overall stance toward *gomutra* (promoting or debunking), thereby establishing the ground truth for our comparative analysis. We structured our gender annotations to reflect the varied formats of these videos, clearly differentiating between monologues with a single primary speaker and interactive, discussion-style videos featuring multiple speakers. The resulting dataset consists of $N = 30$ unique videos. Analysis at the video level re-

veals a highly skewed stance distribution: 24 videos (80%) promote the use of *gomutra*, while 6 (20%) debunk it. To account for videos featuring interactive, multi-speaker formats, our gender analysis was conducted at the speaker level ($N_{speakers} = 33$). This also yields a skewed demographic breakdown of 20 male (60.6%) and 13 female (39.4%) speakers. All computational experiments and data processing were conducted using Google Colaboratory. Detailed hardware specifications and compute budget information are provided in the Appendix.

To provide concrete examples of the health narratives present in our dataset, we highlight a few translated excerpts from the transcripts. Many videos frame the substance as a comprehensive cure for lifestyle diseases, making broad claims such as: *"As a way of life, those who are overweight should drink cow urine... It is a permanent solution to constipation, a medicine for obesity, it purifies the blood, and detoxifies the liver, kidneys, and the entire body."* Other videos attempt to manufacture medical credibility by invoking institutional authority, stating: *"A patent has already been obtained for using cow urine to manufacture medicines for cancer and other diseases, and currently, the patent certificate is displayed on your screen."* Furthermore, transcripts frequently target specific ailments by blending traditional terminology with modern medical jargon, claiming that *"because it reduces phlegm and wind, people suffering from high cholesterol, fatty liver, or blockages in the body, low bone density, or more bloating, all of them can consume it."*

Experimental Setup and LLM Evaluation

Multiple Large Language Models were evaluated for their ability to interpret these competing narratives. First, to analyze macro-level discourse, GPT-4o was employed using a standardized prompt to identify two primary linguistic features: (i) traditional metaphors, defined as narratives deeply connected to cultural heritage (e.g., Ayurvedic medicines, religious references like divine or amrit) to establish credibility; and (ii) scientific terms, comprising pseudo-scientific language that emphasizes empirical evidence and clinical terminology (e.g., antioxidants, toxins). Term densities for these identified features were computed per 100 words across both promoting and debunking stances. To quantify the GPT-4o performance against the human-annotated ground truth, we computed precision (P), recall (R), and the F1-score ($F1$). These metrics are based on the identification of True Positives (TP), False Positives (FP), and False Negatives (FN). Specifically for this study, a true positive is defined as an extracted scientific or traditional term that exists within the set of human-annotated ground-truth terms by a native speaker. These evaluation metrics were calculated for each transcript utilizing the standard mathematical formulations described in the Appendix. The final evaluation scores reported in the main text represent the macro-average of these metrics across all transcripts in the manually annotated validation subset.

Second, to operationalize a theoretically grounded linguistic marker of persuasive intent that is both computationally tractable and linguistically interpretable, intensifiers were selected as a focal linguistic feature because of their established roles in persuasive and deceptive communication (Cheung et al. 2025; Holmes 1990). An intensifier is a word or a phrase specifically used to strengthen, exaggerate, or emphasize a claim. It is a lexical item that operates as a degree modifier on an adjective, an adverb, or occasionally a verb phrase. For example, *very* (as in "The treatment is very effective"), *completely* (as in "It completely cures the disease"), and *absolutely* (as in "This is absolutely proven that cow urine has anti-cancer properties"). Unlike neutral descriptors, intensifiers signal epistemic uncertainty, a well-documented characteristic of health misinformation, which tends toward absolute claims rather than the hedged language typical of evidence-based communication (Van Der Linden 2022). By quantifying intensifier use across promotional and debunking content, we evaluated the following models: GPT-4o-mini, Gemini 2.5 Pro, and DeepSeek-V3.1. To ensure accurate computation, repeated intensifiers within a single model output were filtered out to yield unique counts. For this intensifier analysis, we designed a structured prompting strategy to evaluate model sensitivities, comparing zero-shot (no examples provided) against structured few-shot environments. Furthermore, motivated by prior work demonstrating that surface-level stylistic variations in prompts can significantly affect LLM outputs (Guan et al. 2025; Salinas and Morstatter 2024; Sclar et al. 2024; Zhuo et al. 2024), we investigated the impact of prompt personas by comparing a formal instruction style against a "friendly" conversational tone (e.g., "Hi LLM!"), to analyze whether social cues encourage the models to produce more nuanced, human-like language extraction (full prompts are provided in the Appendix). In addition, to assess cross-model reliability of intensifier identification, we computed pairwise Cohen's Kappa (κ) (Cohen 1960; Landis and Koch 1977) on a presence/absence matrix of the unique intensifiers identified across all conditions. Cohen's Kappa was selected over simple percentage agreement because it corrects for chance agreement, making it appropriate for comparing binary identification decisions across models with different output volumes. For each model pair, a vector of 1s and 0s was constructed indicating whether each intensifier was flagged by that model in any of its four conditions; Kappa was then computed on these vectors. Condition-level Kappa was additionally computed by collapsing across models to compare formal versus friendly tone conditions and zero-shot versus few-shot settings independently.

Complete code is available at Github repository: https://github.com/MsAnamtaKhan/gomutra_health_misinformation

Core Findings

Our analysis addresses each of the three research questions posed in the introduction and yields three principal findings:

RQ1: Asymmetric Rhetorical Framing: Our analysis revealed a significant asymmetry in linguistic strategy between promotional and debunking *gomutra* content. We found that to establish credibility, the speakers leveraged scientific terms (e.g., *betadine*, *detoxifies*, *immunity*, *arthritis*) to project medical legitimacy, while simultaneously em-

ploying traditional metaphors (e.g., *Ayurveda*, *amrit*, *sanjeevani*) to anchor their claims in a deep cultural context. As shown in Figure 2, promotional content exhibited substantially higher densities of both traditional metaphors ($\sim$ 4.4 per 100 words) and scientific terms ($\sim$ 3.2 per 100 words), with traditional metaphors dominating. Debunking content, by contrast, showed markedly lower densities overall ($\sim$1.0 traditional metaphors and $\sim$1.5 scientific terms per 100 words), but crucially employed a more *balanced* rhetorical approach, integrating both scientific terminology and traditional references rather than relying on either exclusively. This balance was statistically confirmed: promotional (positive) content showed significantly higher density of traditional terms (t = 2.61, $p < 0.05$; Mann-Whitney U = 130.5; $p < 0.05$), while debunking (negative) content's more even distribution was similarly significant (t = 2.63, $p < 0.05$; Mann-Whitney U = 133.0; $p < 0.05$) (MacFarland and Yates 2016). To evaluate the performance of the LLM in extracting scientific and traditional terms, we randomly selected one-third of the dataset, consisting of both promoting and debunking content, for manual annotation by a single author (a native speaker of Hindi). We calculated the precision, recall, and F1-score, as shown in Table1, by comparing the human ground-truth annotations against the GPT-4o outputs for each transcript, reporting the average across the subset. For scientific terms, the model achieved an average precision of 61%, a recall of 53%, and an F1-score of 52%. For traditional terms, it achieved a precision of 64%, a recall of 56%, and an F1-score of 59%. These results demonstrate that our automated extraction pipeline provides a sufficiently robust and reliable baseline for evaluating the broader dataset in this exploratory study.

Linguistic Strategy	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Scientific terms	61	53	52
Traditional terms	64	56	59

Table 1: Evaluation of GPT-4o term extraction performance against human annotation.

This asymmetry creates a direct challenge for LLM-based detection. The finding that debunking content employs traditional terminology reflects a communicatively rational strategy: to persuade an audience already invested in traditional medicine, effective debunkers must engage the cultural framework on its own terms rather than simply opposing it with scientific language, which risks being dismissed as culturally alien or elitist. This mirrors well-documented patterns in science communication, where meeting audiences within their existing belief frameworks is more persuasive than direct contradiction (Van der Linden et al. 2025). A model that treats high traditional metaphor density as a marker of misinformation will therefore perform well on promotional content, but will systematically misclassify this sophisticated debunking strategy, flagging culturally competent counter-speech as the very misinformation it is trying to correct.

[0]Note that GPT-4 and GPT-4o were employed for transcript translation and intensifier/prompt-sensitive analysis, respectively.

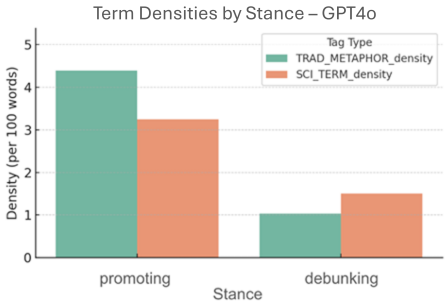

Figure 2: Comparison of traditional metaphor and scientific term densities (per 100 words) given by GPT-4o across promoting and debunking stances.

RQ2: "Friendly" Prompt Doesn't Always Mean LLMs are More Expressive Contrary to our hypothesis, varying prompt persona from formal to friendly (see Appendix), zero-shot to few-shot (i.e., providing a few examples) produced no consistent difference in LLM output quality across models. Table 3 illustrates this inconsistency clearly: the effect of the friendly persona differed in both direction and magnitude across models. Gemini 2.5 Pro showed only a marginal decrease from formal to friendly in zero-shot (21 vs 45 unique intensifiers respectively), while GPT 4o-mini showed a dramatic drop under the friendly condition (48 formal vs 34 few-shot formal, and 14 friendly vs 24 few-shot friendly). DeepSeek-V3.1, on the other hand showed the opposite pattern in few-shot settings, producing slightly more intensifiers under the friendly condition (13 formal vs. 16 friendly). Similarly, few-shot prompting did not consistently outperform zero-shot across models: for GPT-40-mini, few-shot reduced intensifier counts under the formal condition, while for DeepSeek-V3.1, it increased them under the friendly condition.

Inter-model agreement was consistently poor across all pairs (Table 2). Gemini and GPT-4o-mini showed below-chance agreement ($\kappa = -0.256$), indicating systematic divergence rather than random disagreement. Gemini and DeepSeek-V3.1 showed near-zero agreement ($\kappa = -0.006$), suggesting effectively independent identification. GPT-4o-mini and DeepSeek-V3.1 showed the highest agreement, though still only fair ($\kappa = 0.214$). Condition-level analysis revealed comparable divergence: formal versus friendly prompting produced $\kappa = -0.431$, and zero-shot versus few-shot produced $\kappa = -0.452$, indicating that both tone and shot setting caused models to identify systematically different intensifiers rather than merely different quantities.

This inconsistency is itself a substantive finding: LLMs do not respond to social register cues the way human interlocutors do, and cultural competency in automated analysis cannot be achieved through surface-level persona adjustment alone. Deeper architectural and training-level solutions

Comparison	Cohen's κ
Gemini vs GPT-4o-mini	-0.256
Gemini vs DeepSeek-V3.1	-0.006
GPT-4o-mini vs DeepSeek-V3.1	0.214
Formal vs Friendly	-0.431
Zero-shot vs Few-shot	-0.452

Table 2: Pairwise Cohen's κ across models and prompt conditions.

are required in future studies.

RQ3: Speaker Specific Rhetorical Patterns and Gender Bias Risk We identified a distinct gender-based pattern in their rhetorical strategy across the corpus, though we note that the dataset is skewed toward male speakers, reflecting the broader gender imbalance in *gomutra* promotional content on YouTube. As detailed in Table 4, male speakers favored a rhetoric of certainty and exaggeration. They deployed absolute-certainty amplifiers ("never", "permanent") at a distinctly higher rate than their female counterparts (averaging 5.64 occurrences per 1,000 words versus just 0.67). Furthermore, men utilized universality markers ("all," "completely," "everywhere") more frequently (11.51 vs. 9.04 per 1,000 words) and relied heavily on authoritative, first-person framing ("I"), which appeared an average of 10.27 times per 1,000 words for males compared to only 2.61 for females. Notably, the use of hyperbolic terms ("immortal," "miracle") was exclusively observed in male speech within this sample, averaging 1.99 instances per 1,000 words. Female speakers, by contrast, employed relational and inclusive language ("we," "our body") paired with gentle encouragement, building trust through community rather than top-down authority. This shift toward communal framing is quantitatively reflected in their higher frequency of inclusive pronouns ("we/our"), which appeared 8.94 times per 1,000 words in female speech compared to 7.89 in male speech, standing in stark contrast to the male preference for the singular "I."

		Shot Setting	
Model	**Prompt Style**	**Zero-shot**	**Few-shot**
Gemini	Formal	21	35
	Friendly	45	44
GPT4o-mini	Formal	48	34
	Friendly	14	24
DeepSeek-V3.1	Formal	13	16
	Friendly	12	9

Table 3: Comparison of the Number of Unique Intensifiers Generated across Different Models (Gemini, GPT4o-mini, and DeepSeek-V3.1) and Prompt Styles (Formal vs. Friendly), Contrasting Zero-shot and Few-shot Settings.

These patterns are consistent with broader findings in gender and communication research (Lakoff 1973; De Francisco 1992; Holmes 1990). Contemporary studies confirm that these dynamics persist in modern digital spaces; for instance, recent work highlights that female social media users continue to employ more indirect, affiliative language such as hedging (Aydın 2025), whereas male digital influencers tend to rely heavily on direct, assertive framing (Liu, Zhao, and Feng 2022). These gendered communication styles have direct implications for automated moderation. LLMs trained without awareness of gendered rhetorical norms risk systematically misreading female-coded persuasion, which avoids absolutes and relies on relational framing, as less confident or less promotional than it actually is. This introduces a structural gender bias into moderation outcomes that is not a minor calibration issue but a fairness concern with real deployment consequences.

Implications

Our findings do not resolve how LLMs should be used to analyse culturally embedded health misinformation; they reveal the conditions under which such analysis would be unreliable, and what would need to change. There are three main implications of our work from this perspective. First, we reconfirm a finding that has been much talked about in scholarship and the mainstream media alike, that LLMs fail to parse rhetorical signals specific to cultures outside of those on whose data they are built upon. This, in effect means they fail to capture nuance in culturally embedded misinformation. This is articulated in the finding that an understanding of tradition-specific metaphor and its employment is essential in comparing the discursive approaches of promoters and debunkers of gomutra-related theories. Second, we find that even slight alterations in the nature of conversations with LLMs has sizable impacts on the prompt responses. Finally, we find that there are unique elements of gendered behavior that can additionally undermine LLMs' ability to offer fair analyses and outputs of complex narrative content.

The proliferation of such content presents three intertwined challenges: *Epistemically,* culturally embedded misinformation resists verification because it operates through belief systems rather than factual falsehood. This is observed through both the invocation of identity through the call to tradition, but also in the subtle ways in which patriarchy manifests itself subtly in distinctions between male and female influencers' output of the same narrative. *Infrastructurally,* platform algorithms amplify this content even before analyses can occur (Chen and Shu 2024; Tao et al. 2024). That both debunkers and promoters use a mix of science and tradition underlines the perceived value for rhetorical inversion among influencers, who as sophisticated parties in the information economy, understand affective values attached to the cases they wish to make. The hybrid epistemic registers appeal to different elements in the social network to push virality. *Rhetorically,* promotional content exploits cultural familiarity and traditional authority in ways that differ fundamentally from straightforwardly false claims, and that LLMs trained predominantly on Western corpora are structurally ill-equipped to recognise. Here, we see that there is use of obvious linguistic cues, that machines can be trained to recognize, but also less obvious metaphorical references that require an understanding of cultural elements specific to India to accurately debunk misinformation.

Linguistic Markers	Mean per 1000 Words	
	Male	Female
Certainty amplifiers (”never”, ”permanent”)	5.64	0.67
Universality markers (”all”, ”completely”)	11.51	9.04
First-person ”I”	10.27	2.61
Inclusive ”we/our”	7.89	8.94
Hyperbolic terms	1.99	0

Table 4: Comparison of linguistic markers between male and female speakers, measured as a mean per 1,000 words.

The Benchmark Problem Existing misinformation benchmarks are built predominantly on Western, English-language datasets where rhetorical norms, cultural references, and authority structures differ substantially from the Indian context studied here. Our results show that this gap is not merely a matter of translation: LLMs systematically misread the rhetorical inversions specific to this domain, where debunking content *engages* with traditional narratives rather than simply opposing them. A model trained to treat scientific language as a marker of credibility will misread sophisticated debunking content that strategically deploys cultural references for persuasive effect, mistaking culturally competent counter-speech for the very misinformation it is trying to counter.

The case for Indic LLMs Our work here provides a persuasive case for regionally trained language models for fine-grained analysis of layered discourse. Western general-purpose LLMs struggle with sarcasm, cultural idiom, and code-switching patterns characteristic of multilingual Indian health discourse - features that are not edge cases but structural properties of how this content circulates (Gumma et al. 2024; Shukla et al. 2026). What we see here goes beyond the understanding of words and terms to a deeper meaning of what is being said. Since innuendo and figures of speech are commonly found in an misinformation, it makes for a useful case study, but arguably a range of domains in casual or even literary conversation have affective elements that will be lost without culturally thoughtful human engagement.

Do LLMs have a gender problem? The gendered rhetorical patterns we identify have direct implications for LLM-assisted discourse analysis. A model that reads absolute-certainty language as the primary marker of promotional content will systematically underread female-coded persuasion, which operates through relational framing and inclusive language rather than hyperbole. These are not minor calibration issues but structural analytical blind spots that risk misrepresenting whose voices get flagged and whose do not. While our work offers more direct insight into gendered style creeping into social media commentary, a larger question remains whether there is a case for more work into how LLMs understand a world in which the overwhelming majority of conversations it is built on are in the voices of men.

These findings are preliminary and derived from a small, culturally specific corpus (a small-scale analysis). We do not claim direct generalizability to other traditional medicine misinformation domains, though we believe our work, examining asymmetric rhetorical strategies, testing prompt sensitivity, and auditing for gendered patterns, is replicable across culturally diverse discourse contexts. What this study demonstrates is that LLM-assisted discourse analysis in non-Western contexts requires cultural grounding that goes beyond translation and prompt design, it requires rethinking what counts as a reliable analytical instrument in the first place.

Limitations

This study has several limitations that future work should address. First, the corpus of 30 transcripts, while carefully selected, limits statistical power and generalizability; findings should be treated as hypothesis-generating rather than definitive. Second, our reliance on a multilingual translation model introduces potential noise, particularly for code-switched or dialect-specific content. Translation quality was not explicitly evaluated, and models trained exclusively on Indian languages may perform differently, potentially affecting downstream applications. Third, we probe only three LLMs and do not compare against traditional supervised classifiers, which would provide a more complete picture of the detection landscape. Fourth, stance (promoting/debunking) and speaker's gender were annotated by a single author. Finally, the retirement of GPT-4o mini from OpenAI's consumer interface during the study period prevented a fully controlled comparison across all conditions. These limitations define a clear agenda for future work: a larger, human-annotated corpus; direct comparison with Indic LLMs and traditional classifiers; and extension to related traditional medicine misinformation domains in other cultural contexts. Additionally, all prompts were administered in English to Western-developed LLMs. Future work should explore prompting in Indian languages and providing original (untranslated) transcripts directly to the model, which may yield culturally closer responses.

Conclusion

This paper examined how culturally embedded health misinformation - specifically, promotional and debunking content around *gomutra* on Indian YouTube - differs rhetorically, using LLMs as analytical instruments for a post-facto examination of discourse. Our analysis reveals an asymmetry: promotional content relies more heavily on traditional metaphor and cultural authority, while effective debunking integrates both scientific and traditional framing to build credibility. Yet, we do find that promotional content

also uses science and scientific terms to emphasize empirical credibility. LLMs, we find, are sensitive to these differences in ways that are not yet well-controlled, responding inconsistently to prompt tone, misreading gendered rhetorical strategies, and struggling with the cultural inversions that characterize sophisticated debunking. These findings have implications beyond *gomutra*. As LLMs are increasingly proposed as scalable moderation tools across languages and culture, this study illustrates that cultural competency cannot be retrofitted through prompt engineering alone; it requires training data, benchmarks, and evaluation frameworks that reflect the rhetorical diversity of the global information environment.

Culturally embedded health misinformation on YouTube does not look like ordinary misinformation; it blends sacred traditional language with pseudo-scientific claims in ways that even sophisticated debunking content mirrors, and LLMs used to analyse this discourse are systematically misled by these rhetorical patterns in ways that differ by culture, gender, and prompt design.

In the future, we plan to expand this work to larger corpora and also compare the performances of Western models with Indic language models to prevent the loss of cultural nuance inherent in translation, thereby allowing us to strictly preserve Hindi-English code-switching and capture richer sociolinguistic signals (Velutharambath, Sassenberg, and Klinger 2026; Pérez-Montero et al. 2025). Additionally, we also plan to broaden our scope beyond textual analysis to incorporate multimodal analyses, including the full visual context of the videos (Shang et al. 2025); leveraging Optical Character Recognition (OCR) to extract on-screen text and identifying visual cues, such as clinical props, medical reports, and affiliate links, to better map the use of authority and financial incentives driving these claims. Furthermore, we also intend to benchmark current models against a more comprehensive suite of baselines, including classical machine learning classifiers, specialized Indic LLMs (Danish, Liu, and Alshmrany 2025), and foundational linguistic metrics such as lexicon, part-of-speech (POS), and dependency-based structural counts.

Ethical considerations

Our research touches upon deeply rooted cultural beliefs surrounding *gomutra*. To navigate this respectfully, our analysis focuses strictly on the rhetorical strategies that influencers use to spread unverified health claims, rather than presenting a critque of the cultural or traditional practices that underlie the willingness to think of *gomutra* as medicinal. In this, our goal is not to critique traditional medicine more broadly, but rather to critically examine public-facing actors and the algorithmic systems that amplify these narratives. All data was collected exclusively from publicly available YouTube videos. We intentionally avoided collecting Personally Identifiable Information (PII) from viewers or commenters. To comply with platform guidelines, our aggregated dataset will be released strictly for non-commercial, academic research under an ethical use agreement.

Finally, we acknowledge the inherent fairness risks in our gender-based analysis. Annotating speaker gender as a binary variable (male/female) based on visual presentation is deconstructive; it fails to capture the full spectrum of gender identity and carries the risk of misgendering. This methodological compromise was, however, necessitated by the lack of self-reported demographic data. We emphasize that this categorization was used solely as a diagnostic lens to analyze linguistic patterns and expose gender-based biases in LLM outcomes, rather than to essentialize communication styles or reinforce restrictive gender binaries.

Acknowledgements

We would like to thank all the reviewers for their valuable comments and feedback. We are also grateful to Ananya Sharedalal for proposing the idea of investigating cow urine and for compiling the initial list.

References

Adilazuarda, M. F.; Mukherjee, S.; Lavania, P.; Singh, S. S.; Aji, A. F.; O'Neill, J.; Modi, A.; and Choudhury, M. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15763–15784.

Alarabid, A. 2025. Leveraging Large Language Models for Misinformation Detection: A Focus on Public Health Misinformation on Social Media. In *Proceedings of the International Conference on Computing and Information Technology*. Springer.

Asian News International. 2022. Former PM Morarji Desai also used to drink cow urine for medicinal benefits: Ashwini Choubey. News article. Updated February 4, 2022.

Aydın, F. 2025. Examining Gender Differences in Social Media Language. *Bulletin of Language and Literature Studies*, 2(1).

Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J. A.; Wornow, M.; Swaminathan, A.; Lehmann, L. S.; et al. 2025. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 333(4): 319–328.

Chaturvedi, A. 2024. Indian State Says Yoga Guru Misled Public with COVID, Other Cures. News article.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI magazine*, 45(3): 354–368.

Cheung, N.; Grieve, A.; Wilby, K.; Tran, T.; and Lim, A. 2025. Use of Linguistic Communication Strategies (Hedges and Intensifiers) in Simulated Pharmacy Education Shared Decision-Making. *American Journal of Pharmaceutical Education*, 89(10): 101492.

Chhikara, G.; Kumar, A.; and Chakraborty, A. 2025. Through the Prism of Culture: Evaluating LLMs' Understanding of Indian Subcultures and Traditions. *arXiv preprint arXiv:2501.16748*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Danish, M.; Liu, H.; and Alshmrany, S. 2025. A Comparative Study of mBERT and IndicBERT for Natural Language Processing in Indic Languages. In *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*. IEEE.

Daria, S.; and Islam, M. R. 2021. The use of cow dung and urine to cure COVID-19 in India: a public health concern. *International Journal of Health Planning and Management*.

De Francisco, V. L. 1992. Deborah Tannen, You Just Don't Understand: Women and Men in Conversation. New York: William Morrow & Co., 1990. Pp. 330. *Language in Society*, 21(2): 319–324.

Denniss, E.; and Lindberg, R. 2025. Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International*, 40(2): daaf023.

Essar, M. Y.; Kazmi, S. K.; Hasan, M. M.; Costa, A. C. d. S.; and Ahmad, S. 2021. The rampant use of cow dung to treat COVID-19. *Health Science Reports*, 4(4): e363.

Garg, P.; and Fetzer, T. 2025. How Much Does Context Affect the Accuracy of AI Health Advice? *arXiv preprint arXiv:2504.18310*. Version 2, revised February 24, 2026.

Guan, B.; Roosta, T.; Passban, P.; and Rezagholizadeh, M. 2025. The Order Effect: Investigating Prompt Sensitivity to Input Order in LLMs. *arXiv*, abs/2502.04134.

Gulf News. 2019. Indian Health Minister Ashwini Kumar Choubey Is Working on Cow Urine to Prepare Medicines, India.

Gumma, V.; Raghunath, A.; Jain, M.; and Sitaram, S. 2024. HEALTH-PARIKSHA: Assessing RAG Models for Health Chatbots in Real-World Multilingual Settings. *arXiv preprint arXiv:2410.13671*.

Holmes, J. 1990. Hedges and boosters in women's and men's speech. *Language & Communication*, 10(3): 185–205.

Hurford, B.; Rana, A.; and Sachan, R. S. K. 2022. Narrative-based misinformation in India about protection against Covid-19: Not just another "moo-point". *Indian Journal of Medical Ethics*, 7(1): 22–26.

India Today. 2019. Cow Urine to Be Used for Preparing Medicines, Treating Cancer: Health Minister Ashwini Kumar Choubey. News article. Published September 8, 2019.

Jin, Y.; Chandra, M.; Verma, G.; Hu, Y.; De Choudhury, M.; and Kumar, S. 2024. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, 2627–2638. Association for Computing Machinery.

Kaiser Family Foundation. 2024. Poll: Most Who Use Artificial Intelligence Doubt AI Chatbots Provide Accurate Health Information. KFF News Release.

Kelland, K.; and Satter, R. 2020. Trump's COVID-19 Disinfectant Ideas Horrify Health Experts. News article. Updated April 26, 2020.

Kemp, S. 2025a. Digital 2025: Global Overview Report. DataReportal.

Kemp, S. 2025b. Digital 2025: India. DataReportal.

Khalil, M.; Mohamed, F.; and Shoufan, A. 2025. Evaluating the Quality of Medical Content on YouTube Using Large Language Models. *Scientific Reports*, 15(1): 9906.

Khandelwal, K.; Tonneau, M.; Bean, A. M.; Kirk, H. R.; and Hale, S. A. 2024. Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 231–239. Association for Computing Machinery.

Kong, W.; Song, S.; Zhao, Y.; Zhu, Q.; and Sha, L. 2021. TikTok as a Health Information Source: Assessment of the Quality of Information in Diabetes-Related Videos. *Journal of Medical Internet Research*, 23(9): e30409.

Lakoff, R. 1973. Language and woman's place. *Language in society*, 2(1): 45–79.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Li, H. O.-Y.; Pastukhova, E.; Brandts-Longtin, O.; Tan, M. G.; and Kirchhof, M. G. 2022. YouTube as a Source of Misinformation on COVID-19 Vaccination: A Systematic Analysis. *BMJ Global Health*, 7(3): e008334.

Liu, M.; Zhao, R.; and Feng, J. 2022. Gender performances on social media: A comparative study of three top key opinion leaders in China. *Frontiers in psychology*, 13: 1046887.

Liu, Z. 2025. Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. *Journal of Transcultural Communication*, 3(2): 224–244.

Ma, Y.; Zhang, X.; Ren, J.; Wang, R.; Wang, M.; and Chen, Y. 2025. Linguistic features of AI mis/disinformation and the detection limits of LLMs. *Nature Communications*.

MacFarland, T. W.; and Yates, J. M. 2016. *Mann–Whitney U Test*, 103–132. Cham: Springer International Publishing. ISBN 978-3-319-30634-6.

Madathil, K. C.; Rivera-Rodriguez, A. J.; Greenstein, J. S.; and Gramopadhye, A. K. 2015. Healthcare information on YouTube: A systematic review. *Health Informatics Journal*, 21(3): 173–194.

Notermans, C. 2019. Prayers of Cow Dung: Women Sculpturing Fertile Environments in Rural Rajasthan (India). *Religions*, 10(2): 71.

OpenAI. 2023. GPT-4 Technical Report.

Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F. G.; Song, I.; Oh, A.; and Augenstein, I. 2025. Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, 51(3): 907–1004.

Peng, W.; Lim, S.; and Meng, J. 2023. Persuasive strategies in online health misinformation: a systematic review. *Information, Communication & Society*, 26(11): 2131–2148.

Pérez-Montero, A.; Gargova, S.; Lloret, E.; and Moreda, P. 2025. Detecting Deception in Disinformation Across Languages: The Role of Linguistic Markers. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 943–952.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*.

Rodrigues, F.; Newell, R.; Rathnaiah Babu, G.; Chatterjee, T.; Sandhu, N. K.; and Gupta, L. 2024. The social media Infodemic of health-related misinformation and technical solutions. *Health Policy and Technology*, 13(2): 100846.

Salinas, A.; and Morstatter, F. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4629–4651. Bangkok, Thailand: Association for Computational Linguistics.

Sclar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Shang, L.; Zhang, Y.; Deng, Y.; and Wang, D. 2025. MultiTec: a data-driven multimodal short video detection framework for healthcare misinformation on TikTok. *IEEE Transactions on Big Data*, 11(5): 2471–2488.

Shukla, V.; Sharma, H.; Reganti, A. N.; Wasmatkar, S.; Kumar, B.; and Singh, V. 2026. Lost in Translation? A Comparative Study on the Cross-Lingual Transfer of Composite Harms. *arXiv preprint arXiv:2602.07963*.

State Council Information Office of the People's Republic of China. 2020. Fighting COVID-19: China in Action. White paper.

Tao, Y.; Viberg, O.; Baker, R. S.; and Kizilcec, R. F. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9): pgae346.

Thakur, N.; Cui, S.; Knieling, V.; Khanna, K.; and Shao, M. 2024. Investigation of the Misinformation about COVID-19 on YouTube Using Topic Modeling, Sentiment Analysis, and Language Analysis. *Computation*, 12(2): 28.

Van Der Linden, S. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine*, 28(3): 460–467.

Van der Linden, S.; Albarracín, D.; Fazio, L.; Freelon, D.; Roozenbeek, J.; Swire-Thompson, B.; and Van Bavel, J. 2025. Using psychological science to understand and fight health misinformation: An APA consensus statement. *American Psychologist*.

Velutharambath, A.; Sassenberg, K.; and Klinger, R. 2026. What if Deception cannot be Detected? A Cross-linguistic Study on the Limits of Deception Detection from Text. *Computational Linguistics*, 1–71.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The Spread of True and False News Online. *Science*, 359(6380): 1146–1151.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1950–1976. Miami, Florida, USA: Association for Computational Linguistics.

Appendix

Prompts for RQ1

Scientific Terms Identification Prompt

User Instruction:

You are an expert linguistic annotator specializing in scientific discourse. Carefully read the transcript below (ID: transcript_id).

Extract ONLY terms that are explicitly scientific or pseudo-scientific|biomedical, chemical, laboratory, clinical, anatomical, or technical vocabulary typically associated with modern medicine or science.

Include:

- Specific chemicals, compounds, or medicines (e.g., enzyme, uric acid).
- Diseases, medical conditions, or biological processes (e.g., cancer, immune response).
- Diagnostic tools or procedures (e.g., MRI, ECG, blood test).
- Scientific-sounding buzzwords used to lend credibility (e.g., antioxidants, toxins).

Exclude:

- Ayurvedic or traditional medical terms.
- Mythological or religious references.
- General terms like 'health' or 'natural.

Return ONLY a comma-separated list of unique terms (no duplicates, no explanations). If none, return ''.

Transcript: transcript

Traditional/Cultural Terms Identification Prompt

User Instruction:

You are an expert linguistic annotator specializing in traditional, Ayurvedic, and religious discourse. Carefully read the transcript below (ID: transcript_id).

Extract ONLY traditional, Ayurvedic, mythological, religious, or culturally

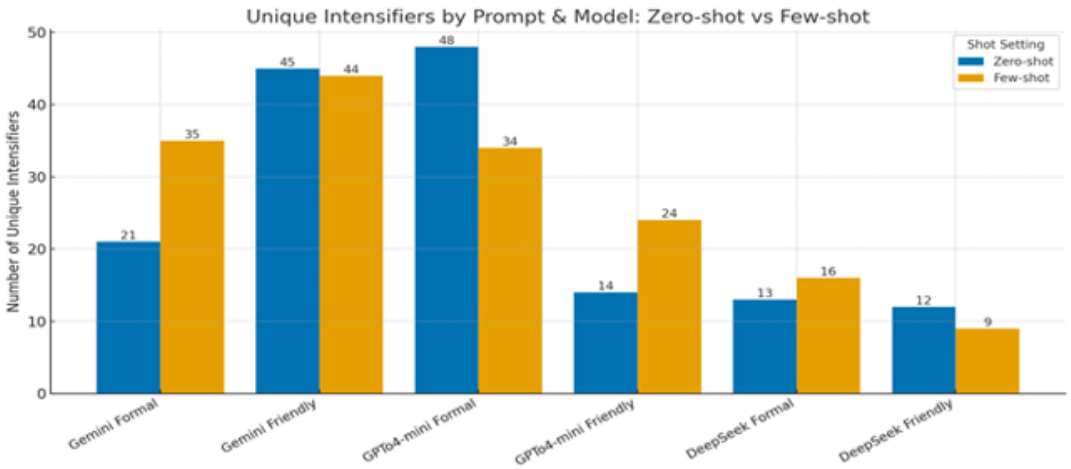

Figure 3: Comparison of the number of unique intensifiers generated across different models (Gemini, GPT4o-mini, and DeepSeek-V3.1) and prompt styles (formal vs. friendly), contrasting zero-shot and few-shot settings.

symbolic terms.

Include:

- Ayurvedic medicines, herbs, or therapies (e.g., Triphala, Ashwagandha).
- Mythological or religious references (e.g., divine, Sanjeevani, sacred cow).
- Traditional metaphors or culturally symbolic phrases (e.g., divine gift, Amrit).

Exclude:

- Modern biomedical, chemical, or technical vocabulary.
- General non-specific terms like "healthy" or "pure"

Return ONLY a comma-separated list of unique terms (no duplicates, no explanations). If none, return ''.

Transcript: transcript

Prompts for RQ2

Formal Prompt

User Instruction:

You are provided with a document containing transcripts from YouTube videos that discuss misinformation about *gomutra* (cow urine). Carefully review the provided texts and complete the following tasks:

Tasks:

1. Identify all intensifiers in each transcript.
   Definition of an intensifier: An intensifier is a word or a phrase specifically used to strengthen, exaggerate, or emphasize a claim. It is a lexical item that operates as a degree modifier on an adjective, an adverb, or occasionally a verb phrase. It augments the head it modifies, without altering the core semantic denotation of that head. For example, \very" (as in \The food was very good"), \strongly" (as in \I strongly believe in the power of education"), \too" (as in \The person was driving too fast\).
2. For each identified intensifier:
   - Quote the exact intensifier.
   - Identify the word it modifies.
   - Indicate its precise position within the text (e.g., beginning, middle, end of the sentence/paragraph).
   - Explain clearly how the intensifier amplifies or contributes to misinformation about *gomutra*.
3. Summarize any patterns or common techniques observed across the transcripts regarding the use of intensifiers in promoting misinformation

Friendly Prompt

User Instruction:

Hi Gemini 2.5 Pro! You're provided with a document containing transcripts from YouTube videos that discuss misinformation about *gomutra* (cow urine). Please carefully review these texts and complete the following tasks to help us with qualitative analysis:

Tasks:

1. Identify all intensifiers in each transcript.
   Definition of an intensifier: An intensifier is a word or a phrase specifically used to strengthen, exaggerate, or emphasize a claim. It is a lexical item that operates as a degree modifier on an adjective, an adverb, or occasionally a verb phrase. It augments the head it modifies, without altering the core semantic denotation of that head. For example, \very" (as in \The food was very good"), \strongly" (as in \I strongly believe in the power of education"), \too" (as in \The person was driving too fast\).
2. For each identified intensifier:
   - Please quote the exact intensifier.
   - Please identify the word it modifies.
   - Please indicate its precise position within the text (e.g., beginning, middle, end of the sentence/paragraph).
   - Please explain clearly how the intensifier amplifies or contributes to misinformation about *gomutra*.
3. Summarize any patterns or common techniques you've noticed across the transcripts regarding the use of intensifiers to promote misinformation.

Thanks for your help!

Few Shot Prompt

The model was (a) first conditioned with a 5-transcript sample to steer its output before (b) processing all the transcripts.

(a) User Instruction:

Below are a few transcripts from YouTube videos that discuss misinformation about *gomutra* (cow urine):

Transcript 1: [TEXT]

Transcript 2: [TEXT]

Transcript 3: [TEXT]

Transcript 4: [TEXT]

Transcript 5: [TEXT]

Carefully review the provided texts and complete the following tasks:

Tasks:

1. Identify all intensifiers in each transcript.
   Definition of an intensifier: An intensifier is a word or a phrase specifically used to strengthen, exaggerate, or emphasize a claim. It is a lexical item that operates as a degree modifier on an adjective, an adverb, or occasionally a verb phrase. It augments the head it modifies, without altering the core semantic denotation of that head. For example, \very" (as in \The food was very good"), \strongly" (as in \I strongly believe in the power of education"), \too" (as in \The person was driving too fast\).
2. For each identified intensifier:
   - Quote the exact intensifier.
   - Identify the word it modifies.
   - Indicate its precise position within the text (e.g., beginning, middle, end of the sentence/paragraph).
   - Explain clearly how the intensifier amplifies or contributes to misinformation about *gomutra*.
3. Summarize any patterns or common techniques observed across the transcripts regarding the use of intensifiers in promoting misinformation

(b) User Instruction:

You are now provided with a document containing all the transcripts from YouTube videos that discuss misinformation about *gomutra* (cow urine). Carefully review the provided texts and complete the following tasks:

Tasks:

1. Identify all intensifiers in each transcript.
   Definition of an intensifier: An intensifier is a word or a phrase

```
   specifically used to strengthen,
   exaggerate, or emphasize a claim. It
   is a lexical item that operates as a
   degree modifier on an adjective, an
   adverb, or occasionally a verb
   phrase. It augments the head it
   modifies, without altering the core
   semantic denotation of that head. For
   example, \very" (as in \The food was
   very good"), \strongly" (as in \I
   strongly believe in the power of
   education"), \too" (as in \The person
   was driving too fast\).
2. For each identified intensifier:
   • Quote the exact intensifier.
   • Identify the word it modifies.
   • Indicate its precise position
     within the text (e.g., beginning,
     middle, end of the
     sentence/paragraph).
   • Explain clearly how the intensifier
     amplifies or contributes to
     misinformation about gomutra.
3. Summarize any patterns or common
   techniques observed across the
   transcripts regarding the use of
   intensifiers in promoting
   misinformation
```

Computational Environment and Compute Budget

- **Compute Budget:** All experiments, data preprocessing, and API integrations were executed using a Google Colab Pro subscription (~$10 USD/month).
- **Hardware Allocation:**Tasks were run on dynamically allocated high-RAM environments equipped with standard Pro-tier GPUs (predominantly NVIDIA V100 or T4). The project operated entirely within the standard allotted compute units, requiring no additional premium cloud instances or external clusters.

Extended Transcription and Translation Protocols

- **Transcription Parameters:** We utilized Whisper's default decoding parameters (e.g., default temperature and beam search). To prevent language-switching artifacts, we explicitly forced the language parameter to Hindi (language="hi") and ran the model in FP32 precision (fp16=False).
- **Translation Parameters:** GPT-4 was run with a stochastic decoding temperature of 0.7. To mitigate this, the system prompt explicitly instructed the model to act as a culturally aware translator, strictly preserving the original meaning, tone, and rhetorical markers without paraphrasing. Furthermore, we preserve and release the original-language transcripts alongside the translations to enable future verification of these cross-lingual mappings.

Evaluation Metrics Formulation

Following evaluation metrics were calculated for each transcript via the following standard formulations:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{P \times R}{P + R}$$

- **True Positive (TP):** A scientific or traditional term correctly extracted by the LLM that perfectly aligns with the human annotation.
- **False Positive (FP):** A term incorrectly extracted by the LLM that was not present in the human-annotated ground truth.
- **False Negative (FN):** A valid term identified by the human annotator that the LLM failed to extract.