Communication Styles and Reader Preferences of LLM and Human Experts in Explaining Health Information

Jiawei Zhou[†1], Kritika Venkatachalam[†1], Minje Choi[2], Koustuv Saha[3], and Munmun De Choudhury[1]

[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA
[2]Amazon, Seattle, WA, USA
[3]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, USA
[†]Authors contributed equally to this research

[*]Corresponding Author:
Jiawei Zhou
Georgia Institute of Technology
756 W Peachtree St. NW
Atlanta, GA, USA, 30308
j.zhou@gatech.edu

ABSTRACT

Background: With the wide adoption of large language models (LLMs) in information creation assistance, it is essential to examine their alignment with human communication styles and values. We situate this study within the context of fact-checking health information, given the critical challenge of correcting misconceptions and building trustworthiness. Recent studies have explored the potential of LLM-based fact-checking, but communication style differences between LLMs and human fact-checkers and associated reader perceptions remain under-explored. In this light, our study evaluates the communication styles of LLMs, focusing on how their explanations differ from those of humans in three core components of health communication: information linguistic features, sender persuasive strategies, and receiver value alignments.
Methods: We compiled a dataset of 1498 health misinformation explanations from authoritative fact-checking organizations. Using this dataset, we employed chain-of-thought prompting with zero-shot and few-shot variations to generate LLM fact-checking responses to inaccurate health information. We drew from health communication theories and categorized communication styles along linguistic, persuasion, and value-based dimensions and measured how closely the LLM-generated responses aligned with professional explanations. Then, we examined human preferences with 99 participants who were unaware of LLM involvement and rated randomized fact-checking articles with switching orders.
Results: Our findings reveal that LLM-generated articles showed significantly lower scores in persuasive strategies, certainty expressions, and alignment with social values and moral foundations. However, human evaluation demonstrated a strong preference for LLM content, with over 60% responses favoring LLM articles for clarity, completeness, and persuasiveness.
Conclusion: Our results suggest that LLMs' structured approach to presenting information may be more effective at engaging readers despite scoring lower on traditional measures of quality in fact-checking and health communication.

Keywords: health misinformation, health communication, large language models, fact-checking, linguistic analysis

Background

With the widespread advancement and proliferation of Large Language Models (LLMs) and generative AI, these technologies are increasingly being used for co-writing and writing assistance across a range of domains—including health communication tasks. Recent studies have demonstrated their potential to generate customized responses and provide support for complex information needs in health-related contexts[1,2]. People are turning to LLMs for quick advice on health issues, seeking guidance not only for themselves but also for their family members[3–5]. However, effective health communication goes beyond simply conveying information; it is also important to frame ideas[6] and align with fundamental human values, since it directly impacts how information is received, trusted, and acted up towards health outcomes[7]. In practice, communicating healthcare content to match the audience's emotional and social contexts is key to meaningful health communication[7,8]. For example, empathic

communication that acknowledges and respects individual experiences can enhance message receptivity and trust[9]. Similarly, audience-centered approaches that adapt language and presentation styles to specific knowledge levels and cultural contexts are shown to improve the clarity and effectiveness of communication[8].

One particularly critical health communication task is explaining misinformation. Expert fact-checking, as one of the crucial traditional methods of misinformation rebuttal, has long-established standards for verifying information[10–12]. Recent studies in automated fact-checking systems, specifically leveraging LLMs, have shown promise in streamlining this process[6,13–15]. Although LLMs have shown technical promise, evidence is still needed to examine communication styles and reader acceptability of LLM-generated content in assessing its potential to assist communicating health misconceptions.

To address the above gap, this work examines whether LLM-generated explanations diverge from human explanations in both linguistic styles and value alignment and whether these differences influence reader preferences. Situating this work in fact-checking false claims during the COVID-19 pandemic, a health crisis time during which opinions were polarized[16] and people were susceptible to vulnerability and rumors[17], we seek to answer the following research questions (RQs):

RQ1. If and how do communication styles differ between LLMs and human fact-checking?

RQ2. Is there a difference in human preference between LLMs and human-generated fact-checking, and why?

To address the RQs, we collected a dataset of 1,498 human-written fact-checking articles from prior work[18,19]. Using chain-of-thought prompting[20], we constructed prompts based on journalism information assessment guidelines[21] and generated LLM fact-checking articles using a set of state-of-the-art models: GPT-4[22], Llama-2-70B[23], and Mistral-7B[24] with zero-shot and few-shot variations. To evaluate how closely LLM-generated responses aligned with human communication, we operationalized communication styles across three key components in health communication: information, receiver, and sender[8]. Specifically, we analyzed 1) information linguistic features of certainty, cognitive processing, and readability, 2) receiver value alignments in social values and moral foundations, and 3) sender persuasive strategies. We found that LLM-generated fact-checking articles showed distinct differences in communication style compared to human writing. LLMs employed fewer persuasive strategies with lower language impact and evidence use, expressed less certainty, and scored lower on moral foundations for authority and fairness. They also demonstrated lower alignment with social values like tradition and conformity. However, this neutrality can help minimize cultural bias. While LLMs display more cognitive expressions, their higher readability scores suggest the use of more complex language.

Then, to explore human preferences for communication styles between human and LLM content, we recruited 99 participants to rate and provide feedback on the clarity, completeness, and persuasiveness of randomly selected fact-checking article pairs. Participants were not informed that some articles were AI-generated, and the order of article pairs was randomized to minimize bias. Our results showed a notable preference for LLM-generated content despite the limitations identified in linguistic analysis. Approximately 62% of participants preferred LLM articles across all three dimensions – clarity, completeness, and persuasiveness, for their focused presentation, objective and neutral tone, professional appearance, and language accessibility.

Together, our study provides implications for the potential use of LLMs in fact-checking and health communication. While LLMs demonstrate strengths in structured presentation and linguistic styles, their limitations in reasoning support, evidence engagement, and moral alignment raise concerns about their reliability in high-stakes domains. The contrasts between language analysis and human preferences suggest that perceived professionalism and clarity may overshadow deeper considerations of nuanced reasoning and holistic coverage. Future work should explore strategies to enhance LLMs' alignment with human reasoning and values while maintaining their communicative advantages.

Methods

We retrieved fact-checking COVID-19 misinformation articles sourced from FNIR[18] and FakeNews[19] datasets that span from February 2020 to July 2020. Both datasets were peer-reviewed and contained source links to fact-checking articles by authoritative fact-checking platforms such as Poynter [1] and FactCheck [2]. After removing non-English content, we collected a total of 1,498 human-written fact-checking articles (referred to as Human) with an average length of 765 (range [47-6473]). We then generated LLM fact-checking articles using chain-of-thought prompting, comprising 1,498 articles in the zero-shot variation and 1492 articles in the few-shot variation (referred to as LLM) with an average length of 335 words (range [9-2135]). We compared the communication styles of the two datasets by extracting linguistic features, value alignments, and persuasive strategies. Then, with approval from our Institutional Review Board, we recruited 99 participants to assess the human preference in reasoning styles.

[1] https://www.poynter.org/
[2] https://www.factcheck.org/

LLM Fact-checking Articles

We used a set of state-of-the-art models to generate LLM fact-checking articles, including GPT-4[22], Llama-2-70B[23], and Mistral-7B[24]. We employed chain-of-thought prompting[20], a widely adopted method known to assist LLMs in performing complex reasoning, with zero-shot and few-shot variations. In the zero-shot prompt, we did not present any examples of human fact-checked articles, which allows us to focus solely on the LLMs' independent reasoning process. For the few-shot prompt, we supplemented with examples of human fact-checked articles for both true and false claims.

To generate fact-checking articles that follow the same general journalism guidelines as human articles, we constructed our prompt based on an information assessment guideline[21] built upon prior research in journalism communication[25] and information credibility indicators[26]. Three key criteria included in this guideline were source credibility, availability of evidence, and consideration of alternative explanations. We incorporated these criteria by instructing the models to write a detailed fact-checking article for the given claim while considering the credibility of sources for the claim, any evidence that supports the information, and any alternative explanation. Our prompt contained two subtasks (Table 1): classify the veracity of a given claim and provide reasoning for the classification.

Based on this prompting strategy, we collected responses from the LLMs across five runs, ensuring random selection, ordering, and alteration of the few-shot examples given to the models. Following common settings in few-shot learning[27,28], we used one true example and one false example for each run from a total of 10 few-shot examples. This setup ensured that each evaluation included balanced guidance for the model in determining claim veracity.

Communication Styles

To compare LLM-generated and human-written fact-checking articles, we analyzed communication styles across three key components – information, sender, and receiver[8] – by examining linguistic features, value alignments, and persuasive strategies.

- *Information Linguistic Features:* For information, we examined linguistic features of certainty, cognitive attributes, and readability[29]. Certainty is the confidence expressed in statements, which is critical in discussing healthcare risk in physician-patient communication[30]. We used the model developed by Pei and Jurgens[31] to capture both the level and aspects of certainty in text. Cognitive attributes represent human cognitive processing, such as causal and insightful thinking and is essential to understanding. We used a validated psycholinguistic lexicon, Linguistic Inquiry and Word Count (LIWC)[32] to measure cognitive attributes in writing, a tool that has been extensively utilized in empirical studies on misinformation and detection research[33–35]. Lastly, to measure the accessibility of language, we used Automated Readability Index (ARI)[36] and Flesch-Kincaid grade level[37] that indicate the grade level required to understand the text, with a lower score meaning the text is easier to read.
- *Sender Persuasive Strategies:* On the sender side, we evaluated persuasive strategies employed in the articles in achieving communication goals. We analyzed the persuasive tactics employed to influence opinion or belief[38,39] based on a previous work that examined the strategies of credibility, evidence, and impact[39]. This hierarchical weakly-supervised latent variable model leverages partially labeled data to predict persuasive strategies in text at both the document and sentence levels.
- *Receiver Value Alignment:* On the receiver side, we assessed value alignment, which directly impacts the effectiveness of health communication[7], by analyzing references to social values and moral foundations. To measure human social values, we utilized Schwartz values[40,41] to capture social values of security, power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, conformity, and tradition. Then, to analyze moral considerations in text, we used Mformer[42], a fine-tuned large language model that assesses moral foundations, such as authority and fairness, as defined by Moral Foundations Theory[43].

Human Evaluation

To examine human preference for communication styles in fact-checking, we recruited 100 participants from the crowd-sourcing research platform Prolific [3] and 99 participants passed attention checks and completed the survey. Table 2 describes participant demographics. We defined the inclusion criteria as adults who can read English and live in America. Participants were informed that the study aimed to examine different strategies for writing fact-checking responses to potential misinformation, without knowing that some articles were generated by LLMs or which specific articles were AI-generated. The actual research purpose was disclosed to participants at the end of the study. We conducted pilot testing with 40 participants to define our survey before launching the study, and included several attention checks to ensure valid responses. The survey was hosted on a website created for this study and consisted of three parts:

- *Demographics*: We asked participants about age, gender, race/ethnicity, and level of education.

[3] https://www.prolific.com/

- *Article Rating*: Each participant rated randomly selected five pairs of fact-checking articles, with the order of LLM and human content shuffled. The article pairs were randomly selected from a subset due to the length difference between Human and LLM datasets. This subset was selected from the initial dataset where the human fact-checking articles were within 250-350 words to ensure no identifiable difference between LLM-generated articles and human articles. For each article pair, without informing the involvement of AI generations and with the order of human-AI article pairs randomized, participants answered three rating questions on a seven-point scale. The three rating questions were: Which article uses more clear and understandable language? Which article provides all the necessary information to understand the fact being checked? Which article do you think is more persuasive?
- *Preference Reason*: Additionally, we included one open-ended question for each pair of articles about why they preferred a certain article (or the lack of preference). Participants were encouraged to provide at least two factors.

To examine the effects of demographic characteristics and relative communication style differences on preferences, we performed multivariate ordinal logistic regression models[44,45] with outcome variables of preferences in language clarity, information completeness, and persuasiveness. To help explain our regression findings, we also analyzed participants' reasons for their preferences toward LLM-generated articles. Following the general inductive approach[46], two authors first independently engaged in a close reading of all participant responses to gain a preliminary understanding while taking low-level notes to capture factors mentioned in answers using their own words (e.g., *"facts in everyday terms"*) or quoting participant responses (e.g., *"didn't overwhelm you with information not pertinent to the story"*). Then, they discussed their notes and grouped low-level notes into higher-level themes, such as *"neutrality in presenting information"*. Throughout the analysis, the research team engaged in ongoing discussions to refine and clarify emerging themes.

Results

Correctness

We conducted two sets of prompt experiments: a zero-shot prompt to evaluate the models' inherent ability and style, and a few-shot prompt where we provided human examples of fact-checking articles to help LLMs capture phrasing strategies used by humans. The few-shot prompting performed slightly better with 93.68% in recall, gaining a 2.95% improvement over zero-shot.

Communication Styles

ANOVA tests showed significant differences for all the communication style dimensions of linguistic features, persuasive strategies, and value alignments (Table 3). To ensure a robust comparison, we also applied Tukey's HSD test to further analyze these differences and identify specific pairs with statistically significant variations. Figure 1 presents the score distribution plots for both human and LLM assessments. For brevity, LLM scores were averaged across the three LLMs used for this study: GPT-4, Llama-2-70B, and Mistral-7B. Individual models' scores were presented in Table 5 in Supplementary Materials.

Information linguistic features

Certainty: LLM-generated content scored significantly lower in terms of certainty than human writing, although the mean difference between the groups was smaller as compared to the other dimensions tested.
Cognitive attributes: LLM-generated content involved significantly more cognitive processing expressions. With higher explanations of cognitive justifications, LLMs could achieve better reasoning in writing.
Readability: We found LLM articles had significantly higher readability scores, meaning that they are more complex and harder to read than human writings. The lower competence of LLM in readability is especially problematic for health claims and misinformation corrections because it may hinder understanding across a diverse audience and weaken trust by failing to present clear and accessible information.

Sender persuasive strategies

LLM-generated articles presented lower language impact and less evidence compared to human-authored articles (Figure 1a and 1b). The Tukey HSD test confirmed these observations with a negative mean difference for all LLM-human pairs.

Receiver value alignment

Social values: LLM articles presented lower levels of power (Figure 1d) and achievements (Figure 1g) than human writings. LLMs' less frequent presentations of control or dominance suggest that human authors tended to be better at establishing authority and guiding readers toward correct information in effective health communication. LLM's lower emphasis on competence indicates that human authors tended to be skilled at presenting information in a way that aligns with societal expectations of accuracy and reliability. We also found that LLM-generated articles exhibited lower tradition (Figure 1f) and conformity (Figure 1e) scores than human writing. A lower tradition score suggests that LLM content was less culturally grounded, which can affect how people view customs and religions. However, this neutrality can be a strength, helping to avoid cultural bias and promote inclusive content. A lower conformity score suggests that LLMs are less likely to spread biases or

stick to cultural stereotypes like human writers might. Lower conformity could be a positive sign, meaning the content is more neutral and credible.
Moral foundations: LLM content scored significantly lower in moral dimensions than human content, suggesting that LLMs are less likely to emphasize fairness and authority in writing.

Human Evaluation

Reader preference

To assess preference between LLMs and human-generated articles, 99 participants each rated five random content pairs and together provided a total of 495 responses. In general, more individuals preferred LLM-generated articles regarding clarity of language, inclusion of essential information, and persuasiveness. The Wilcoxon Signed-Rank Test rejected the null hypothesis of no preference and showed a significant preference for LLM-generated articles. Figure 2 shows the human preferences, and regression analyses of individual and content-related factors influencing these preferences are provided in Table 4.

- *Language clarity:* 308 out of 495 responses (62.23%) preferred LLM-generated articles for their clarity and ease of reading, with 28.08% strongly preferred LLMs. This suggests that LLM content might be better at helping people understand complex health information. Regression results showed that male individuals and individuals identifying as Black/African American and White are more likely to prefer LLM-generated articles. At the same time, people were more likely to prefer articles that expressed more security or tradition-related values or utilized less language impact. While participants' education levels and content readability were not significantly related to language clarity preferences, we found that participants with higher education levels were more likely to find LLMs clearer, and that LLM articles with lower readability scores (i.e., easier to read) were rated as clearer.
- *Necessary information:* 307 out of 495 responses (62%) preferred LLM-generated articles in providing comprehensive information on health claims. Conversely, a substantial proportion of participants (29.9%) expressed a preference for human-generated content. Regression results indicated that male participants tended to prefer LLM fact-checking and its reasoning style as more thorough in including necessary information.
- *Persuasiveness:* 306 out of 495 responses (61.8%) found LLM-generated articles to be more persuasive, with 125 responses (25.25%) strongly preferring LLMs. 29.09% participants were inclined towards human-generated articles, with 41 responses (8.28%) strongly preferring human articles. Contrary to prior literature showing authorities as trusted sources with high persuasiveness due to source credibility[47,48], our regression results indicated that when LLM articles used fewer authority-related expressions than their paired human articles, participants were more likely to find LLMs more persuasive.

Reasons for witnessed LLM preference

To understand factors contributing to the observed preference for LLM fact-checked articles, we analyzed participants' reasons for preferences without knowing the involvement of AI generations and summarized five factors:

- *Focused presentation*: Participants preferred LLM-generated fact-checking articles for their clear organization, direct presentation, and focused content – often starting with the main point and avoiding unnecessary details. The paired human-written ones, on the other hand, were sometimes described as scattered, convoluted, and prone to including excessive or off-topic information, and sometimes giving undue emphasis to false claims.
- *Objective standpoint*: LLM-generated fact-checking articles were often characterized as more objective and fact-based with minimal opinions, while human-written ones were sometimes associated as opinionated or personal.
- *Professional vibe*: Many participants perceived LLM-generated fact-checking articles as more professional-sounding with cross-checking among multiple organizations, well-explained background information, and references from reputable sources. Comparatively, some human-written articles used *"non-reputable citations"* or *"social media evidence"*.
- *Language accessibility*: Participants found LLM-generated fact-checking articles easier to follow with clear and digestible language that involved *"everyday terms to present facts"* in an accessible manner.
- *Non-emotional tone*: Participants also pointed out that LLM-generated fact-checking articles employed a calm and neutral tone, while some human-written articles were perceived as *"angry"* or *"accusatory"* and sometimes resembling clickbait grabbing attention.

Discussion

Our study presents empirical evidence on the linguistic features and human preferences between LLM-generated and human-written fact-checking articles addressing health misinformation. Language analysis indicated some potential limitations in LLM-generated content compared to human-written ones. Specifically, LLM-generated fact-checking articles engaged less with persuasive strategies of language impact, credibility, and evidence; displayed less certainty in writing; and aligned less

with moral foundations of authority and fairness as well as important social values like tradition, self-direction, and conformity. These findings suggest that LLMs may prioritize writing-level persuasion over reasoning support and moral alignment. In addition, LLM articles tended to be more challenging to read for people without high education levels. Although they presented more cognitive processing explanations, this characteristic could potentially make the text feel too analytical rather than intuitive and accessible.

Despite these potential limitations of LLMs presented by language analysis, our human evaluation results suggested a stronger preference towards LLM articles in writing clarity, inclusion of necessary information, and persuasiveness. Although this preference for clarity contrasts with the language analysis, our qualitative findings suggest that readers appreciate LLMs' use of digestible terms and their structured and focused presentation style, even if readability metrics indicate otherwise. Similarly, we were surprised to find that human preferences in informational completeness and persuasiveness also contradicted the language analysis. Specifically, LLMs engaged with less evidence and authoritative and fair expressions while aligned less with social values – dimensions that typically contributed to the comprehensiveness of information and effectiveness of persuasion[49]. However, qualitative results suggest that LLMs' tendency to cite reputable sources, even when citations are incorrect or non-existent, contributes to a perception of professionalism and evidence support. Together, our findings suggest that people may prioritize surface-level information presentation – such as structured and focused answers, an objective and neutral tone, and a professional appearance – over characteristics traditionally associated with high-quality fact-checking and health communication like nuanced reasoning, moral alignment, and holistic coverage.

LLMs for Fact-checking Explanation: Taken together, our findings suggest the potential of leveraging LLMs in fact-checking workflows while balancing their strengths and limitations. Human preference towards LLM articles despite opposite findings from the language analyses implies that the way LLMs present information may create an impression of completeness and persuasiveness, even if the content itself may fall short on depth or rigor. Thus, while LLMs may not replace human fact-checkers due to their inferior performance on credibility and evidence and weaker alignment with moral and social values, we see potential in them complementing human efforts by presenting and organizing information in a recipient-friendly manner. Future work can study how to fine-tune models to enhance factors valued by high-quality fact-checking such as evidence support and value alignment.

LLMs in Health Communication: In a broader sense, our work sheds light on the potential role of LLMs in assisting health communication. The observed human preference for LLM-generated content in language clarity, information completeness, and persuasiveness suggests the potential for using LLMs in organizing and revising patient-facing materials. However, their weaker performance on reasoning support and moral alignment could pose risks in health contexts where cultural sensitivity, trust-building, and reliability are critical[8]. These limitations underscore the need for careful evaluation and responsible use of LLMs in high-stakes contexts like healthcare and combating misinformation. Future research can explore developing strategies to fine-tune and monitor LLMs while implementing human oversight to ensure reliability, accountability, and ethical alignment in their applications.

Conclusion

Our study highlights a disconnection between expert-derived communication benchmarks and lay reader preferences in the context of LLM-generated health information corrections. Despite scoring lower on traditional dimensions of high-quality health communication, such as persuasive strategies, certainty expression, and alignment with moral and social values, LLM-generated fact-checking responses were consistently preferred by participants. This preference appears to be driven by LLMs' structured presentation, clarity, and neutral tone, which may foster an impression of completeness and professionalism, even when the underlying content lacks nuance or rigor. Together, these findings indicate both the potential and the limitations of deploying LLMs in fact-checking and health communication workflows. While LLMs can enhance the accessibility of complex information, their current shortcomings in value-sensitive communication raise important concerns for use in high-stakes settings. Future research should focus on fine-tuning LLMs to improve alignment with ethical and communicative standards in health, and on developing frameworks for integrating LLMs responsibly alongside human expertise.

List of abbreviations

LLM: large language model; AI: artificial intelligence; LIWC: linguistic inquiry and word count; ARI: automated readability index; ANOVA: analysis of variance; HSD: honestly significant difference; CI: confidence interval

Declarations

Ethics approval and consent to participate
This study was considered exempt with minimal risk by the Institutional Review Board of Georgia Institute of Technology. No personal information was collected, and participants were informed that the presented statements contained misinformation, followed by fact-checking explanations.

Consent for publication
Not applicable.

Availability of data and materials
The datasets are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests

Funding
JZ, MC, and MDC were partly supported by U.S. National Science Foundation grant #2137724. This research project has benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

Authors' contributions
JZ and KV contributed to study conceptualization and design, data collection and analysis, result interpretation, and manuscript writing. MC contributed to study conceptualization and design, and data collection and analysis. KS contributed to study conceptualization, result interpretation, and manuscript writing. MDC contributed to and supervised study conceptualization and design, result interpretation, and manuscript writing. All authors read and approved the final manuscript.

Acknowledgements
We thank the study participants for contributing to this work.

Tables and Figures

(a) Claim Verification (Boolean)

```
Let's think step by step.
Establish the credibility of sources and return TRUE or FALSE for
the given claim.
[MASK]
<claim-to-verify>
```

(b) Fact-Checking Article Generation

```
Write a detailed fact-checking article of around 800 words for the given
claim.  Give a detailed explanation.
Let's think step by step.  Establish the credibility of sources.
Is there any evidence that supports the information or claim?  What might
be an alternative explanation or understanding?
[MASK]
<claim-to-verify>
<classification-result>
```

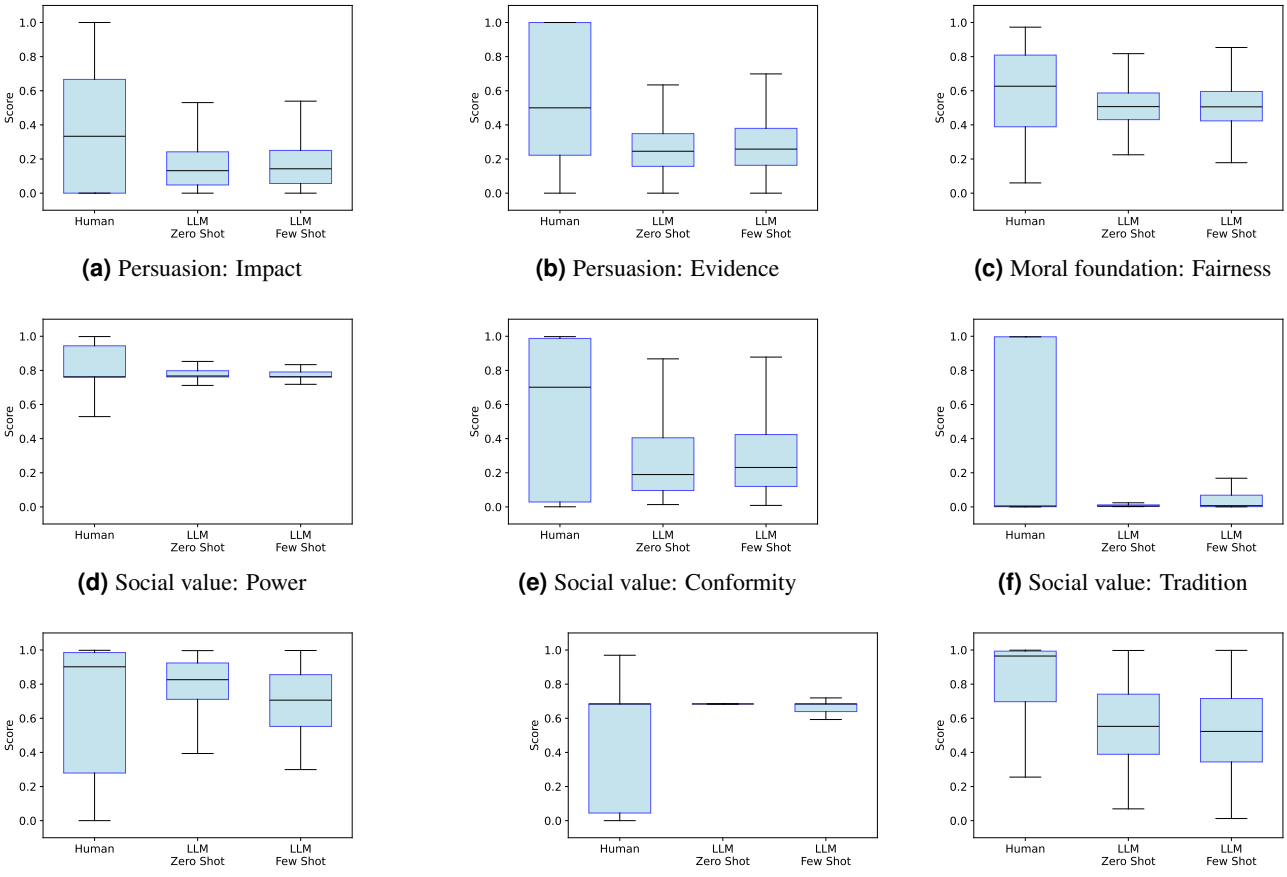
Table 1. Prompt templates for fact-checking tasks, with the [MASK] token serving as a placeholder for few-shot examples.

Demographic	Number of participants
Age Distribution	
18-24	10
25-34	31
35-44	28
45-54	14
55-64	9
65 or above	7
Gender Distribution	
Male	36
Female	62
Prefer not to say	1
Ethnicity Distribution	
American Indian or Alaska Native	0
Asian	7
Black or African American	23
Hispanic or Latino or Spanish Origin	9
Native Hawaiian or Pacific Islander	2
White	53
Other	5
Education Distribution	
High School or less	13
Some college or technical certification	28
Bachelor's degree	39
Master's degree or higher	19
Prefer not to say	0

Table 2. Overview of participant demographics (N = 99).

Dimension	F-statistic	p-adj
Persuasive Strategies		
Impact	303.87	***
Credibility	133.57	***
Evidence	518.88	***
Certainty		
Certainty	63.97	***
Moral Foundations		
Authority	51.02	***
Fairness	94.32	***
Social Values		
Security	72.34	***
Power	154.38	***
Benevolence	53.85	***
Conformity	216.51	***
Tradition	464.75	***
Achievement	65.93	***
Hedonism	167.75	***
Stimulation	361.72	***
Self-direction	277.42	***
Universalism	146.51	***
Readability		
ARI	806.21	***
Flesch-Kincaid	996.88	***
Cognitive Processes		
Cogproc	618.82	***

Table 3. ANOVA results for framing of reasoning dimensions (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)



(a) Persuasion: Impact **(b)** Persuasion: Evidence **(c)** Moral foundation: Fairness

(d) Social value: Power **(e)** Social value: Conformity **(f)** Social value: Tradition

(g) Social value: Achievement **(h)** Social value: Stimulation **(i)** Social value: Self-direction

Figure 1. Framing dimensions with significant differences between LLMs and human writing

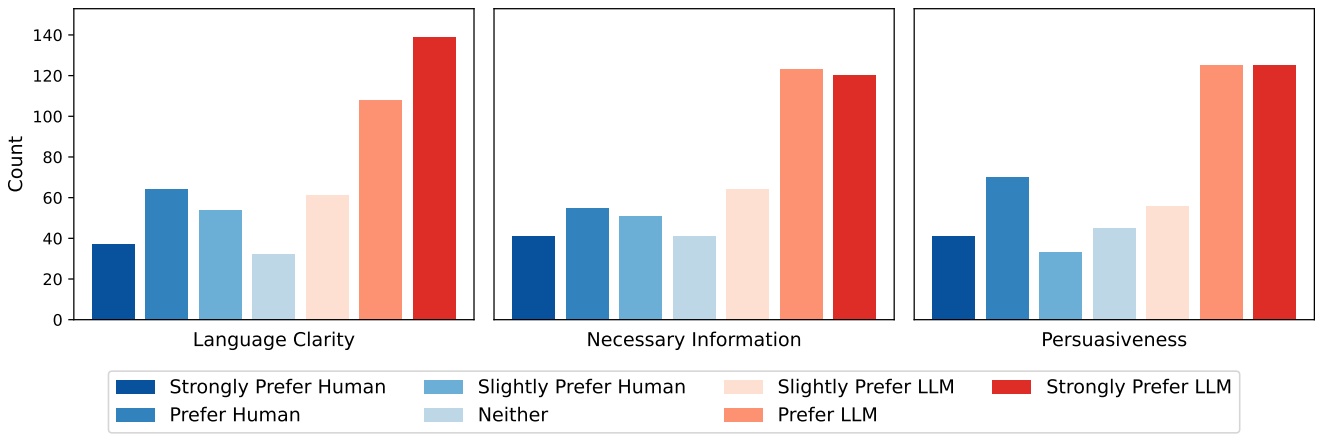


Figure 2. Human preferences for language clarity, necessary information and persuasiveness

Variable	Language clarity			Necessary information			Persuasiveness		
	Odds Ratio	p	**95% CI**	**Odds Ratio**	p	**95% CI**	**Odds Ratio**	p	**95% CI**
Demographic characteristics									
Age	1.0345		(-0.0923 - 0.1601)	1.0952		(-0.0337 - 0.2156)	1.0587		(-0.0697 - 0.1837)
Education	1.1123		(-0.0971 - 0.3099)	1.0875		(-0.1176 - 0.2854)	1.1438		(-0.0702 - 0.3389)
Gender (Masculine)	1.4994	*	(0.0594 - 0.7507)	1.5093	*	(0.0686 - 0.7548)	1.1923		(-0.1691 - 0.5208)
Gender (Prefer not to say)	1.6311		(-1.0901 - 2.0686)	3.1977		(-0.7203 - 3.0452)	2.5849		(-0.7033 - 2.6027)
Ethnicity (Black/African American)	2.0809	*	(0.0679 - 1.3977)	1.7438		(-0.0954 - 1.2075)	1.7117		(-0.1182 - 1.1931)
Ethnicity (Hispanic/Latino/Spanish)	1.0816		(-0.733 - 0.8898)	0.7371		(-1.1178 - 0.5076)	1.0856		(-0.7197 - 0.884)
Ethnicity (Native Hawaiian/Pacific Islander)	0.7029		(-1.5006 - 0.7956)	0.4227		(-2.0514 - 0.3292)	0.4667		(-1.9647 - 0.4406)
Ethnicity (Other)	1.1164		(-0.7429 - 0.9631)	1.1393		(-0.7186 - 0.9794)	1.3533		(-0.5545 - 1.1595)
Ethnicity (White)	2.1562	*	(0.1407 - 1.396)	1.0656		(-0.5446 - 0.6717)	1.2865		(-0.3613 - 0.8651)
Relative linguistic differences									
Cognitive process	1.0207		(-0.1461 - 0.1872)	0.8809		(-0.2919 - 0.0382)	0.9115		(-0.2564 - 0.0712)
Social Value: security	1.2434	*	(0.0352 - 0.4005)	1.0761		(-0.1084 - 0.255)	1.1411		(-0.0554 - 0.3194)
Social Value: power	1.0395		(-0.1563 - 0.2337)	1.0669		(-0.1286 - 0.2581)	1.0245		(-0.1728 - 0.2211)
Social Value: achievement	0.9113		(-0.2611 - 0.0753)	1.1002		(-0.0724 - 0.2633)	0.9899		(-0.1792 - 0.1589)
Social Value: hedonism	0.989		(-0.1557 - 0.1335)	1.0647		(-0.0909 - 0.2164)	1.0838		(-0.0782 - 0.2392)
Social Value: stimulation	1.0511		(-0.1195 - 0.2193)	1.0265		(-0.1444 - 0.1967)	1.0698		(-0.1011 - 0.2361)
Social Value: universalism	0.9519		(-0.2527 - 0.1542)	0.8968		(-0.3131 - 0.0953)	0.9814		(-0.217 - 0.1794)
Social Value: benevolence	1.0099		(-0.1679 - 0.1875)	1.1932		(-0.0301 - 0.3833)	1.1319		(-0.0735 - 0.3213)
Social Value: conformity	0.9216		(-0.2479 - 0.0847)	0.8983		(-0.2649 - 0.0503)	0.9884		(-0.1776 - 0.1541)
Social Value: tradition	1.2185	*	(0.0242 - 0.3711)	1.1383		(-0.0336 - 0.2927)	1.1698		(-0.0077 - 0.3214)
Moral Foundation: authority	0.8882		(-0.2883 - 0.0511)	0.8512		(-0.3408 - 0.0187)	0.8208	*	(-0.3712 - -0.0238)
Moral Foundation: fairness	1.0195		(-0.1775 - 0.2161)	0.9566		(-0.2388 - 0.1501)	1.0002		(-0.1875 - 0.1879)
Certainty	0.9226		(-0.2396 - 0.0784)	0.9176		(-0.2458 - 0.0739)	0.9619		(-0.1935 - 0.1158)
Persuasive Strategy: Credibility	0.9146		(-0.2904 - 0.1118)	0.8947		(-0.3107 - 0.0881)	1.0189		(-0.1916 - 0.229)
Persuasive Strategy: Impact	0.8069	*	(-0.3994 - -0.0297)	0.8563		(-0.341 - 0.0307)	0.9715		(-0.2162 - 0.1584)
Persuasive Strategy: Evidence	0.9662		(-0.1976 - 0.1289)	0.99		(-0.1687 - 0.1485)	1.0423		(-0.1208 - 0.2037)
Readability: ARI	0.9605		(-0.2192 - 0.1386)	0.9983		(-0.1792 - 0.1757)	0.9583		(-0.2215 - 0.1363)

Table 4. Regression results for reader preferences in language clarity, information provision, and persuasiveness. (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

References

1. Kuroiwa, T. *et al.* The potential of chatgpt as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J. medical Internet research* **25**, e47621 (2023).
2. Benichou, L. The role of using chatgpt ai in writing medical scientific articles. *J. Stomatol. Oral Maxillofac. Surg.* **124**, 101456, DOI: https://doi.org/10.1016/j.jormas.2023.101456 (2023).
3. Wang, Z., Yang, Z., Azimi, I. & Rahmani, A. M. Differential private federated transfer learning for mental health monitoring in everyday settings: A case study on stress detection. In *Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2024).
4. Galitsky, B. A. Llm- based personalized recommendations in health. *Preprints* DOI: 10.20944/preprints202402.1709.v1 (2024).
5. Guo, Z. *et al.* Large language models for mental health applications: Systematic review. *JMIR Ment Heal.* **11**, e57400, DOI: 10.2196/57400 (2024).
6. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. & Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Palmer, M., Hwa, R. & Riedel, S. (eds.) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937, DOI: 10.18653/v1/D17-1317 (Association for Computational Linguistics, Copenhagen, Denmark, 2017).
7. Kreuter, M. W. & McClure, S. M. The role of culture in health communication. *Annu. Rev. Public Heal.* **25**, 439–455 (2004).
8. Schiavo, R. *Health communication: From theory to practice* (John Wiley & Sons, 2013).
9. Mheidly, N. & Fares, J. Leveraging media and health communication strategies to overcome the covid-19 infodemic. *J. public health policy* **41**, 410–420 (2020).
10. Bautista, J. R., Zhang, Y. & Gwizdka, J. Healthcare professionals' acts of correcting health misinformation on social media. *Int. J. Med. Informatics* **148**, 104375 (2021).
11. Swire-Thompson, B., Lazer, D. *et al.* Public health and online misinformation: challenges and recommendations. *Annu. Rev Public Heal.* **41**, 433–451 (2020).
12. Guo, Z., Schlichtkrull, M. & Vlachos, A. A survey on automated fact-checking. *Transactions Assoc. for Comput. Linguist.* **10**, 178–206, DOI: 10.1162/tacl_a_00454 (2022). https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf.
13. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, DOI: 10.18653/v1/N19-1423 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
14. Hoes, E., Altay, S. & Bermeo, J. Leveraging chatgpt for efficient fact-checking, DOI: 10.31234/osf.io/qnjkf (2023).
15. Zhang, X. & Gao, W. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method (2023). 2310.00305.
16. Hart, P. S., Chinn, S. & Soroka, S. Politicization and polarization in covid-19 news coverage. *Sci. communication* **42**, 679–697 (2020).
17. Tasnim, S., Hossain, M. M. & Mazumder, H. Impact of rumors and misinformation on covid-19 in social media. *J. preventive medicine public health* **53**, 171–174 (2020).
18. Saenz, J. A., Kalathur Gopal, S. R. & Shukla, D. Covid-19 fake news infodemic research dataset (covid19-fnir dataset), DOI: 10.21227/b5bt-5244 (2021).
19. Shahi, G. K. & Nandini, D. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19* (ICWSM, 2020).
20. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. neural information processing systems* **35**, 24824–24837 (2022).
21. Zhou, J., Zhang, Y., Luo, Q., Parker, A. G. & De Choudhury, M. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–20, DOI: 10.1145/3544548.3581318 (Association for Computing Machinery, New York, NY, USA, 2023).

22. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

23. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

24. Jiang, A. Q. *et al.* Mistral 7b (2023). 2310.06825.

25. Kovach, B. & Rosenstiel, T. *Blur: How to know what's true in the age of information overload* (Bloomsbury Publishing USA, 2011).

26. Zhang, A. X. *et al.* A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, 603–612 (2018).

27. Parnami, A. & Lee, M. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291* (2022).

28. Mittal, S., Jung, H., ElSherief, M., Mitra, T. & De Choudhury, M. Online myths on opioid use disorder: A comparison of reddit and large language model. *Proc. ICWSM* (2025).

29. Saha, K., Jain, Y., Liu, C., Kaliappan, S. & Karkar, R. Ai vs. humans for online support: Comparing the language of responses from llms and online communities of alzheimer's disease. *ACM Transactions on Comput. for Healthc.* (2025).

30. Berry, D. *Risk, communication and health psychology* (McGraw-Hill Education (UK), 2004).

31. Pei, J. & Jurgens, D. Measuring sentence-level and aspect-level (un)certainty in science communications. In Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-t. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9959–10011, DOI: 10.18653/v1/2021.emnlp-main.784 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021).

32. Pennebaker, J. W. Linguistic inquiry and word count: Liwc 2001 (2001).

33. Jiang, S. & Wilson, C. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proc. ACM on Human-Computer Interact.* **2**, 1–23 (2018).

34. Su, Q., Wan, M., Liu, X., Huang, C.-R. *et al.* Motivations, methods and metrics of misinformation detection: an nlp perspective. *Nat. Lang. Process. Res.* **1**, 1–13 (2020).

35. Zhang, X. & Ghorbani, A. A. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. & Manag.* **57**, 102025 (2020).

36. Smith, E. A. & Senter, R. *Automated readability index*, vol. 66 (Aerospace Medical Research Laboratories, Aerospace Medical Division, Air . . . , 1967).

37. Flesch, R. Flesch-kincaid readability test. *Retrieved Oct.* **26**, 2007 (2007).

38. Chen, S., Xiao, L. & Mao, J. Persuasion strategies of misinformation-containing posts in the social media. *Inf. Process. & Manag.* **58**, 102665 (2021).

39. Chen, J. & Yang, D. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests (2021). 2101.06351.

40. Schwartz, S. Basic human values: Theory, measurement, and applications. *Revue Francaise de Sociol.* **47**, 929–968+977+981 (2006).

41. Van Der Meer, M., Vossen, P., Jonker, C. M. & Murukannaiah, P. K. Do differences in values influence disagreements in online discussions? *arXiv preprint arXiv:2310.15757* (2023).

42. Nguyen, T. D. *et al.* Measuring moral dimensions in social media with mformer (2024). 2311.10219.

43. Graham, J. *et al.* Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, vol. 47, 55–130 (Elsevier, 2013).

44. Im, J. *et al.* Still out there: Modeling and identifying russian troll accounts on twitter. In *Proceedings of the 12th ACM conference on web Science*, 1–10 (2020).

45. Kuehn, K. M. & Salter, L. A. Assessing digital threats to democracy, and workable solutions: a review of the recent literature. *Int. J. Commun.* **14**, 22 (2020).

46. Thomas, D. R. A general inductive approach for analyzing qualitative evaluation data. *Am. journal evaluation* **27**, 237–246 (2006).

47. Pornpitakpan, C. The persuasiveness of source credibility: A critical review of five decades' evidence. *J. applied social psychology* **34**, 243–281 (2004).

48. Dutta-Bergman, M. *et al.* Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation. *J. medical Internet research* **5**, e893 (2003).

49. Thon, F. M. & Jucks, R. Believing in expertise: How authors' credentials and language use influence the credibility of online health information. *Heal. communication* **32**, 828–836 (2017).

Supplement

Dimensions	LLM	meandiff	*p*-adj
Persuasive Strategies			
Impact	GPT4	-0.2292	***
	Llama-2-70B	-0.1957	***
	Mistral-7B	-0.1906	***
Credibility	GPT4	-0.0586	***
	Llama-2-70B	0.0505	***
	Mistral-7B	0.1106	***
Evidence	GPT4	-0.3104	***
	Llama-2-70B	-0.2678	***
	Mistral-7B	-0.2418	***
Certainty			
Certainty	GPT4	-0.1723	***
	Llama-2-70B	-0.0221	**
	Mistral-7B	-0.0354	**
Moral Foundation			
Authority	GPT4	0.064	***
	Llama-2-70B	0.005	
	Mistral-7B	0.0025	
Fairness	GPT4	-0.0661	***
	Llama-2-70B	-0.0755	***
	Mistral-7B	-0.0997	***
Social Values			
Security	GPT4	0.1061	***
	Llama-2-70B	0.0311	***
	Mistral-7B	0.0571	***
Power	GPT4	0.138	***
	Llama-2-70B	0.1041	***
	Mistral-7B	0.125	***
Benevolence	GPT4	0.0708	***
	Llama-2-70B	0.0244	***
	Mistral-7B	0.0433	***
Conformity	GPT4	-0.1789	***
	Llama-2-70B	-0.2459	***
	Mistral-7B	-0.273	***
Tradition	GPT4	-0.2538	***
	Llama-2-70B	-0.2728	***
	Mistral-7B	-0.2715	***
Achievement	GPT4	0.09	***
	Llama-2-70B	-0.0385	***
	Mistral-7B	0.0151	
Hedonism	GPT4	-0.0665	***
	Llama-2-70B	-0.0639	***
	Mistral-7B	-0.066	***
Stimulation	GPT4	0.1641	***
	Llama-2-70B	0.1354	***
	Mistral-7B	0.1624	***
Self-direction	GPT4	-0.1754	***
	Llama-2-70B	-0.2972	***
	Mistral-7B	-0.2297	***
Universalism	GPT4	0.0288	***
	Llama-2-70B	0.0204	***
	Mistral-7B	0.0259	***
Readability			
ARI	GPT4	2.9773	***
	Llama-2-70B	1.9263	***
	Mistral-7B	2.2695	***
Flesch-Kincaid	GPT4	-2.8112	***
	Llama-2-70B	1.934	***
	Mistral-7B	1.8838	***
Cognitive Processes			
Cogproc	GPT4	3.2234	***
	Llama-2-70B	3.8885	***
	Mistral-7B	2.9714	***

Table 5. Tukey's HSD test results between Human and LLMs for Persuasive Strategies, Certainty, Moral Foundations and Social Values (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)