Speaking at the Right Level: Literacy-Controlled Counterspeech Generation with RAG-RL

Xiaoying Song[1] Anirban Saha Anik[1] Dibakar Barua[1]
Pengcheng Luo[2] Junhua Ding[1] Lingzi Hong[1]
[1] University of North Texas
[2] Peking University
{xiaoyingsong, anirbansahaanik, dibakarbarua}@my.unt.edu
luopc@pku.edu.cn, {junhua.ding, lingzi.hong}@unt.edu

Abstract

Health misinformation spreading online poses a significant threat to public health. Researchers have explored methods for automatically generating counterspeech to health misinformation as a mitigation strategy. Existing approaches often produce uniform responses, ignoring that the health literacy level of the audience could affect the accessibility and effectiveness of counterspeech. We propose a **Controlled-Literacy** framework using retrieval-augmented generation (RAG) with reinforcement learning (RL) to generate tailored counterspeech adapted to different health literacy levels. In particular, we retrieve knowledge aligned with specific health literacy levels, enabling accessible and factual information to support generation. We design a reward function incorporating subjective user preferences and objective readability-based rewards to optimize counterspeech to the target health literacy level. Experiment results show that Controlled-Literacy outperforms baselines by generating more accessible and user-preferred counterspeech. This research contributes to more equitable and impactful public health communication by improving the accessibility and comprehension of counterspeech to health misinformation.

1 Introduction

Health misinformation online poses a serious risk to public health as it can mislead people, cause bad health choices, and reduce trust in doctors and medical advice (Do Nascimento et al., 2022). A user on Reddit once claimed that "I won't take a mammogram because it is the squishing that causes cancer." Such health misinformation not only misleads individuals into avoiding appropriate medical actions but also increases public confusion (Fridman et al., 2023). Therefore, it is essential to curb the spread of health misinformation.

Social media users actively participate in combating health misinformation by posting counterspeech (Kuo and Chen, 2025; Anik et al., 2025; Micallef et al., 2020). However, user-generated responses often lack credible evidence, decreasing the effectiveness of counterspeech (Yue et al., 2024a). While experts and fact-checkers can provide factual support in debunking health misinformation, it becomes less efficient when confronted with an increasing volume of health misinformation (Yue et al., 2024b). Generating supportive counterspeech at scale remains a challenge.

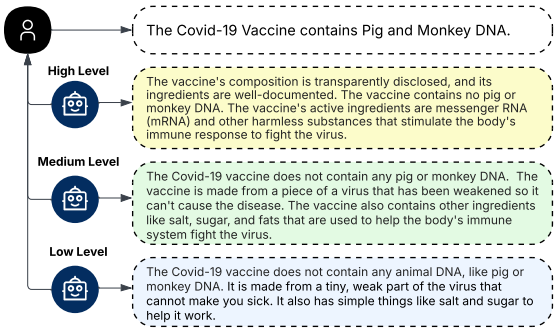

Figure 1: An example of health misinformation, paired with three counterspeech responses tailored to low, medium, and high health literacy levels.

Recent studies have explored the use of Large Language Models (LLMs) to generate counterspeech in response to health misinformation (Yue et al., 2024a,b; Wang et al., 2024b; He et al., 2023). However, these methods often overlook a critical issue: Is the generated counterspeech accessible and persuasive to users with different health literacy levels? Health literacy refers to the capability to understand, judge, and use health information to make good choices and stay healthy (World Health Organization, 2024). The health literacy level of the audience can significantly affect the effectiveness of counterspeech (Liu et al., 2020). Users with lower health literacy often struggle to comprehend counterspeech grounded in scientific research (Shahid et al., 2022; Liu et al., 2020) and understand information that contains complex or technical language (Centers for Disease Control

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2812–2830
November 4-9, 2025

and Prevention, 2024; Chen et al., 2018). In contrast, users with a higher literacy level may lose interest or perceive the content as oversimplified if it is tailored to a low comprehension level (August et al., 2024; Martínez Silvagnoli et al., 2022). For example, simplistic explanations such as *"It is made from a tiny, weak part of the virus that cannot make you sick ..."* in Figure 1 (low-level counterspeech) may be inappropriate for users with advanced health literacy. Such mismatches in communication can critically undermine the effectiveness of counterspeech.

In this work, we aim to generate counterspeech for users with diverse health literacy levels. To address this question, we propose a controllable RAG framework, the **Controlled-Literacy**. The approach uniquely retrieve knowledge adapting to diverse health literacy levels, considering that the complexity of retrieved knowledge directly influences the style and clarity of the generated responses (Ke et al., 2025). We then integrate Reinforcement Learning (RL) to further optimize the generation, ensuring that the outputs align with the target health literacy level.

Our Controlled-Literacy framework is capable of producing counterspeech that is more accessible, user-preferred, polite, and also factually accurate when addressing health misinformation. We summarize our contribution as follows: (1) We introduce the novel insight that effective counterspeech should be aligned with the health literacy level of its target audience. (2) We propose **Controlled-Literacy**, a framework that integrates RAG and RL for generating accessible counterspeech that accounts for both *objective readability and subjective user preferences.* (3) A new dataset is curated, **MisinfoLiteracy**, consisting of health misinformation posts paired with counterspeech responses tailored to users with different health literacy levels.

2 Related Work

2.1 Counterspeech to Misinformation Generation

Counterspeech has been proven effective in mitigating misinformation (Peng and Grimmelmann, 2024; Siebert and Siebert, 2023). Previous studies focus on generating counterspeech with desirable attributes. For example, Anik et al. (2025) and Yue et al. (2024a) combined external scientific sources to generate evidence-based counterspeech in response to misinformation. Yue et al. (2024b) generated factually accurate counterspeech by synthesizing contrastive arguments derived from fact-checking sources. Hong et al. (2024) adapted the generation of counterspeech according to preferred conversation outcomes, resulting in more positive conversation engagement. He et al. (2023) proposed a reinforcement learning-based framework to generate counterspeech with enhanced politeness, factuality, and refutational strength. Despite these efforts, existing studies have not yet investigated counterspeech generation tailored to users with different health literacy levels.

2.2 Accessible Language Generation

Accessible language generation is widely explored in multiple domains, such as health care (Yao et al., 2024; Rahman et al., 2024; Luo et al., 2022; Phatak et al., 2022), education (Wang et al., 2024a; Malik et al., 2024; Rooein et al., 2023), and finance (Wang et al.; Kosireddy et al., 2024; Perez-Rojas et al., 2023). In health care, previous studies primarily focus on simplifying or summarizing complex medical terminology. Yao et al. (2024) introduced a dataset with medical terms paired with lay definitions, aiming to enhance the accessibility of medical information for non-experts. In education, studies have investigated tailoring learning materials to learners at varying proficiency levels. For instance, Malik et al. (2024) have experimented with methods to control the language proficiency level of LLM-generated content to suit different types of learners. In finance, the emphasis has been on democratizing access to financial information for a broader audience. Kosireddy et al. (2024) used small language models to make financial question-answering more accessible for people with limited resources. Existing approaches primarily optimize for objective readability metrics while neglecting subjective factors such as user perceptions.

3 Methodology

3.1 Task Definition

We follow the framework proposed by Nutbeam (2000), which defines and categorizes health literacy into three hierarchical levels: *Functional Health Literacy* involves essential reading and writing skills necessary for understanding everyday health information. *Interactive Health Literacy* encompasses more advanced cognitive skills, enabling individuals to communicate effectively and apply new health information to changing circum-

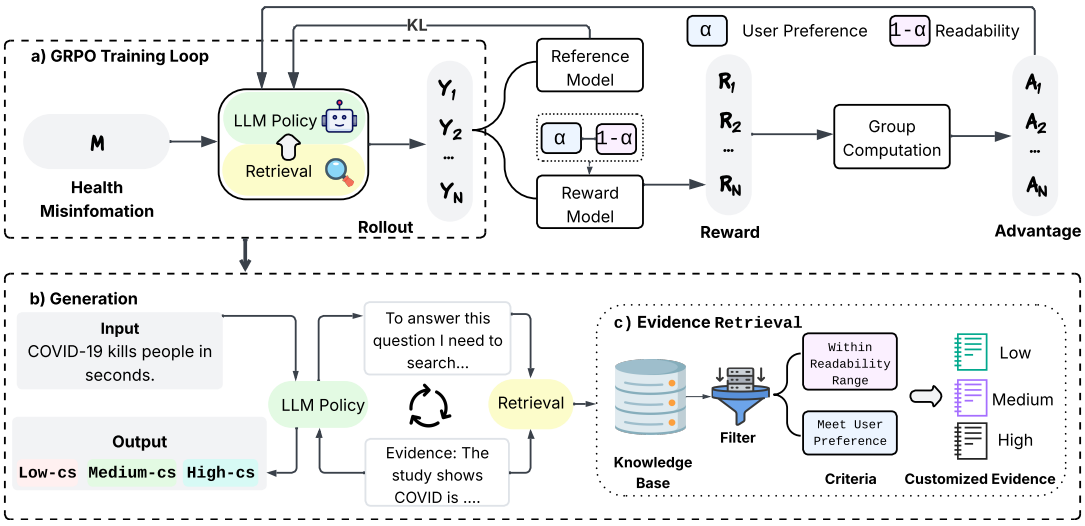

Figure 2: Overview of our *Controlled-Literacy* counterspeech generation framework tailored to users with different health literacy levels. (a) The GRPO training loop integrates evidence retrieval into the LLM policy and optimizes it using a hybrid reward function combining user preference (weight α) and readability (weight $1-\alpha$). Rewards are aggregated through group computation to compute the advantage signal. (b) During inference, the model takes a health misinformation input and retrieves customized evidence to generate counterspeech adapted to low, medium, or high health literacy users. (c) The evidence retrieval module selects content from the knowledge base by filtering it according to the target readability range and user preference thresholds, ensuring personalized support for the generation process.

stances. *Critical Health Literacy* includes the ability to critically analyze health information and understand determinants of health. Based on the framework, we define three target health literacy levels: *low*, *medium*, and *high*, corresponding to Functional Health Literacy, Interactive Health Literacy, and Critical Health Literacy, respectively.

Given the complex definition of health literacy, measuring health literacy in its entirety remains a methodological challenge, especially at scale and in automated settings. In our framework, we first adopt the Flesch-Kincaid Reading Ease (FKRE) score as a practical, scalable proxy for estimating the functional dimension of health literacy. FKRE quantifies text difficulty based on syntactic features, which directly influence comprehension. This makes it an approximate tool for distinguishing materials accessible to users with varying levels of health literacy. For example, texts with higher FKRE scores (e.g., 80–100) are generally easier to read and more appropriate for individuals with lower health literacy. Referring to the previous study (Rooein et al., 2023), we collapse the FKRE score into three categories, easy (80-100), medium (60-79), and hard (0-59).

However, FKRE alone does not capture users' ability to critically interpret or act upon health information, dimensions that fall under interactive and critical health literacy. FKRE also fails to reflect whether users perceive the content as helpful, trustworthy, or respectful. To address this limitation, our approach augments FKRE with simulated user preference ratings, derived from LLMs conditioned to emulate users at different health literacy levels. These simulated raters assess counterspeech based on perceived accessibility, clarity, and helpfulness, providing a subjective complement to the FKRE score. Together, these signals enable our model to generate content that aligns with the syntactic simplicity suitable for a target group and resonates with the group's communication expectations and information-processing styles.

The task is formulated in the following: given a piece of health misinformation m and relevant retrieved knowledge k, the function f generates a counterspeech response c that is tailored to match the health literacy level l of a target user group.

$$f : (m, k, l) \mapsto c$$

3.2 Controlled-Literacy Pipeline

The Controlled-Literacy Pipeline includes three sections (See Figure 2): Knowledge Base Construction, Health Literacy-Adaptive Evidence Retrieval, and Controlled Literacy RL Generation.

Knowledge Base Construction Users with different levels of health literacy get information from

different sources and prefer content with varying complexity and detail (Chen et al., 2018). We collect knowledge from a diverse set of reliable sources to ensure the inclusiveness and representativeness of knowledge across user groups. These include federal health agencies such as the Centers for Disease Control and Prevention (CDC) and academic and research databases like the Johns Hopkins Children's Center. In addition to diverse sources, we employ FKRE to evaluate the readability of each document, categorized as easy (80–100), medium (60–79), or hard (0–59), to ensure that materials are at different readability levels.

Health Literacy-Adaptive Evidence Retrieval We utilize a hybrid retrieval method to get evidence from the knowledge base, which has been proven to perform better than a single method (Sawarkar et al., 2024). Our retrieval module integrates two retrieval methods: keyword-based (R_k) and semantic retrieval (R_s). The hybrid retriever (R_h) integrates the strengths of two methods: $R_h = R_k \cup R_s$. When retrieving the top-N documents ($R_h = \{d_1, d_2, \ldots, d_N\}$), these documents are concatenated into a single context: $C = \text{concat}(d_1, d_2, \ldots, d_N)$.The concatenated context C is then paired with the input query q to construct the prompt for the LLM to generate responses r.

After retrieving knowledge, we filter the evidence using both the FKRE score and LLM-simulated user preference ratings. We design a 1–5 Likert-style scale to evaluate their preference for the generated counterspeech, where a rating of 3 or higher indicate that users with a certain health literacy level find the content to be at least acceptable. We retain evidence that falls within the target FKRE range and receives a preference rating of 3 or higher, leaving out content users find unhelpful or confusing (rated 1 or 2). This ensures the selected content is both syntactically accessible and aligned with users' communication expectations.

Controlled Literacy RL Generation A counterspeech generator is trained to adapt to the target audience's health literacy level. It balances a readability-based reward with a user preference reward to optimize outputs, enabling the generation to be more accessible to a user group. The FKRE score assesses the difficulty of texts. Rather than enforcing strict constraints, we consider a response desirable if its FKRE score falls within the corresponding target range. This design allows for greater flexibility during optimization and encourages more natural counterspeech generation (Ibrahim et al., 2024). We adopt a double-sigmoid function to optimize outputs into the target range of readability:

$$r_{\text{read}} = \sigma\left(\frac{\text{F} - L}{s}\right) - \sigma\left(\frac{\text{F} - R}{s}\right) \quad (1)$$

F is the FKRE score. L and R are the left and right boundaries of the target FKRE range. s is a scaling factor controlling the transition smoothness.

User preference measures whether the generated counterspeech is accessible and easy for users with the target health literacy level to comprehend and apply. Since no publicly available datasets exist for this task and employing human annotators is costly, we leverage LLMs with customized instructions to simulate users from different health literacy levels.

We ultimately incorporate both evaluations to optimize generation, ensuring that the counterspeech objectively fits the target readability level and subjectively aligns with user preferences. We define the final reward as a weighted combination of two components: the readability reward and the user preference reward. Specifically,

$$r(x, y) = \alpha \cdot r_{\text{read}}(x, y) + (1 - \alpha) \cdot r_{\text{pref}}(x, y) \quad (2)$$

where $\alpha \in [0, 1]$ is a hyperparameter that controls the balance between promoting text that fits the target readability range and ensuring subjective accessibility for users across different health literacy levels. $r(x, y)$ is the composite reward. r_{read} is the readability reward based on the FKRE score, The preference reward r_{pref} is derived from LLM-based Likert-style scoring (1–5) simulated for a given health literacy level.

After confirming the reward function, we use Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to optimize the counterspeech. GRPO generates n responses for each prompt and computes their relative advantages based on an aggregated reward signal. The policy is then updated to increase the likelihood of higher-ranked responses while constraining divergence from the reference policy π_ϕ using KL regularization. The optimization objective is:

$$\mathbb{E}_{(x,y_i)\sim\pi_\theta}\left[r(x, y_i) - \beta\,\text{KL}\left(\pi_\theta(y_i \mid x) \,\|\, \pi_\phi(y_i \mid x)\right)\right] \quad (3)$$

where $r(x, y)$ is the composite reward function, and β controls the regularization strength.

4 Experiment Results

4.1 Datasets

MisinfoLiteracy We utilize PRAW API [1] to collect health misinformation from Reddit, focusing on topics related to the coronavirus disease (COVID-19). We collect Reddit posts containing health-related keywords (e.g., "vaccines," "COVID-19," "alternative medicine"). We obtain 3,872 posts from high-engagement subreddits (e.g., r/health or r/science, etc. See details in Appendix A). Then, we employ human annotators to identify and filter health misinformation. We document the details in Appendix B. The final dataset contained 440 posts labeled as health misinformation.

MisinfoCorrect This public dataset contains 789 misinformation tweets processed by (He et al., 2023) with pairs of misinformation and corresponding counterspeech responses. We use this dataset to fine-tune our LLMs to prevent them from rejecting prompts containing misinformation. Additionally, as the dataset originates from a different platform and exhibits a distinct linguistic style, it serves as a testbed for our cross-generalization experiments.

Check-COVID The dataset is a public benchmark dataset designed to facilitate the fact-checking of COVID-19-related news claims (Wang et al., 2023). The dataset comprises 1,504 claims, which are either directly taken from news articles or manually written by annotators to represent common misinformation. The dataset is also used in our cross-generalization experiments.

4.2 Baseline

Instructional Prompt We examine whether the generation could achieve the target without further training. We experiment with several prompt settings and document the best-performing prompts in Appendix D.

RAG RAG integrates external knowledge into generation to avoid hallucinations. We build a RAG system with evidence selection to generate counterspeech for different health literacy levels. The experiment tests the capability of RAG with no generation optimization. The prompt is derived from the Instructional Prompt setting.

4.3 Experiment Setup

We experiment with various LLMs, including models with similar parameter sizes but different architectures (e.g., LLaMA3.1-8B-Instruct[2] vs. Qwen2.5-7B-Instruct[3]), and models from the same family with varying parameter sizes (e.g., LLaMA3.1-8B-Instruct vs. LLaMA3.2-1B-Instruct[4]). This setup allows us to investigate model performance across diverse architectures and parameter scales.

Knowledge Base We collect information related to COVID-19 from diverse sources to construct the knowledge base, including COVID-19 fact sheets from the CDC, materials from Johns Hopkins Children's Center to help kids understand the pandemic, and a collection of research articles about COVID-19 (Wang et al., 2020). The full list of sources is documented in Appendix C.

Evidence Retrieval We incorporate both keyword-based and semantic retrieval methods. In both cases, the misinformation statement alone is used as the retrieval query, excluding the full prompt to prevent the introduction of irrelevant noise during the retrieval process. Retrieved results from both methods are merged using an `AND` operation, and the combined candidates are subsequently ranked to select the *Top-k* relevant knowledge chunks for generation. We use LLaMA3.1-8B-Instruct as an example and experiment with multiple *Top-k* selections (See Appendix F).

Evidence Selection We filter for evidence that falls within the target FKRE range: easy (80–100), medium (60–79), or hard (0–59), and has a user preference rating equal to or above 3 (the midpoint of a 5-point scale). This ensures the retrieved knowledge is both appropriately readable and aligned with user preferences, enabling more customized and supportive generation.

RL Optimization We fine-tune LLMs using GRPO to generate counterspeech tailored to users with low, medium, and high health literacy. The policy backbone is a LoRA-adapted supervised model, initially fine-tuned on the MisinfoCorrect dataset, which pairs health misinformation with corresponding counterspeech to prevent response refusal when prompted with misinformation. For RL, we use the MisinfoLiteracy dataset, splitting it into 80% for GRPO training and 20% for inference evaluation. The reward function combines two

[1] https://praw.readthedocs.io/

[2] Available at: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] Available at: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] Available at: https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

Method	Literacy Level	Politeness	Target Distance (↓)	User Preference	Factual Accuracy
LLaMA-8B					
Instructional Prompt	low	0.39 (0.23)	2.16 (3.69)	0.75 (0.03)	0.86
	medium	0.49 (0.22)	5.99 (8.24)	0.74 (0.06)	0.89
	high	0.35 (0.19)	0.07 (0.74)	0.75 (0.04)	0.87
	Avg.	**0.41** (0.21)	**2.74 (4.22)**	**0.75** (0.04)	**0.87**
RAG	low	0.72 (0.25)	1.30 (3.23)	0.71 (0.13)	0.89
	medium	0.66 (0.22)	4.45 (6.35)	0.71 (0.13)	0.88
	high	0.41 (0.21)	0.08 (0.75)	0.72 (0.09)	0.90
	Avg.	**0.60** (0.23)	**1.94 (3.44)**	**0.71** (0.12)	**0.89**
Controlled-Literacy	low	0.84 (0.15)	1.21 (2.21)	0.74 (0.07)	0.89
	medium	0.69 (0.20)	1.50 (3.81)	0.73 (0.10)	0.90
	high	0.98 (0.02)	0.00 (0.00)	0.75 (0.00)	0.93
	Avg.	**0.84 (0.15)**	**0.90 (2.01)**	**0.74 (0.04)**	**0.91**
Qwen-7B					
Instructional Prompt	low	0.36 (0.22)	0.37 (3.54)	0.74 (0.06)	0.71
	medium	0.48 (0.23)	4.25 (6.50)	0.73 (0.13)	0.88
	high	0.41 (0.18)	0.37 (3.54)	0.73 (0.12)	0.89
	Avg.	**0.42 (0.15)**	**1.66 (4.53)**	**0.73 (0.09)**	**0.83**
RAG	low	0.69 (0.25)	0.60 (1.97)	0.69 (0.16)	0.79
	medium	0.52 (0.18)	3.07 (5.55)	0.68 (0.18)	0.92
	high	0.45 (0.19)	0.01 (0.13)	0.73 (0.07)	0.91
	Avg.	**0.55 (0.13)**	**1.23 (2.55)**	**0.70 (0.13)**	**0.87**
Controlled-Literacy	low	0.77 (0.18)	2.09 (3.38)	0.73 (0.11)	0.84
	medium	0.55 (0.19)	4.08 (5.67)	0.74 (0.08)	0.94
	high	0.86 (0.22)	0.00 (0.02)	0.75 (0.04)	0.92
	Avg.	**0.73 (0.16)**	**2.06 (3.02)**	**0.74 (0.08)**	**0.90**
LLaMA-1B					
Instructional Prompt	low	0.65 (0.28)	8.17 (9.36)	0.66 (0.17)	0.50
	medium	0.40 (0.25)	17.30 (23.39)	0.58 (0.22)	0.58
	high	0.70 (0.24)	0.19 (2.87)	0.68 (0.14)	0.54
	Avg.	**0.58 (0.16)**	**8.55 (11.87)**	**0.64 (0.17)**	**0.54**
RAG	low	0.50 (0.31)	16.90 (14.34)	0.45 (0.25)	0.63
	medium	0.41 (0.21)	9.64 (12.20)	0.36 (0.25)	0.71
	high	0.42 (0.22)	0.51 (4.92)	0.42 (0.22)	0.63
	Avg.	**0.44 (0.05)**	**9.02 (10.49)**	**0.34 (0.24)**	**0.66**
Controlled-Literacy	low	0.73 (0.17)	2.08 (3.62)	0.68 (0.15)	0.61
	medium	0.63 (0.21)	1.86 (4.10)	0.71 (0.12)	0.75
	high	0.85 (0.27)	0.00 (0.00)	0.67 (0.13)	0.76
	Avg.	**0.74 (0.11)**	**1.31 (2.57)**	**0.69 (0.13)**	**0.71**

Table 1: Counterspeech generation results on **MisinfoLiteracy** categorized by health literacy levels: low, medium, and high, along with the overall average. The mean(variance) across samples is reported for Politeness, Target Distance, and User Preference. Factual Accuracy reports the percentage of responses that are factually correct. User preference presents the evaluation by intended user category. Higher mean values for politeness, user preference, and factual accuracy indicate better performance, while lower mean values for target distance indicate better alignment. The best overall performance in each category is highlighted in gray.

components: (1) a double-sigmoid FKRE-based score that promotes readability within a specific target range (e.g., 80–100 for low literacy), and (2) a GPT-4o-mini-based user preference score, rated on a 1–5 Likert scale, measuring how well the generated counterspeech aligns with user expectations. These two signals are weighted equally (0.5 readability, 0.5 user preference) and aggregated to compute the total reward. For each prompt, LLMs generate four responses and compute their relative advantages based on an aggregated reward signal. The GRPO training loop then updates the policy to increase the likelihood of higher-reward responses while constraining divergence from a fixed reference policy (initialized from the supervised LoRA-adapted model). This setup enables the model to learn responses that are both accessible and effective across varying levels of health literacy. The GRPO training is configured with the following hyperparameters: a per-device batch size of 4, gradient accumulation steps of 1, a learning rate of 5×10^{-6}, and training epochs of 3. For each prompt, the model generates 4 completions with a maximum token length of 200. The GRPO-specific regularization coefficient β is set to 0.2.

Generation For all generations, we set `max_new_tokens` to 200, `do_sample=False`, `temperature=0.5`, and `top_p=0.9`.

4.4 Evaluation

We evaluate the generated responses across four key dimensions: target distance, user preference, politeness, and factual accuracy. In the following, we provide detailed definitions and measurements for each of these evaluation dimensions.

Target Distance This metric indicates how close the readability of a generated response is to the target range. We use the FKRE score to measure readability. As outlined in Section 3.2, FKRE scores are categorized into three distinct ranges correspond-

Literacy Level	Category	Tolerant Match	Cohen's Kappa
Low	Computing	0.94	0.65
	Human	0.96	0.75
Medium	Computing	0.96	0.65
	Human	0.92	0.69
High	Computing	0.96	0.65
	Human	0.88	0.63

Table 2: Agreement between human-based and computing-based evaluations and inter-annotator agreement among human evaluators on User Preference.

ing to different health literacy levels. The target distance quantifies the deviation between the actual FKRE score of a response and its designated target range. A lower target distance indicates better alignment with the intended readability level.

Literacy Level	Pearson	Spearman	Kendall's Tau
Low level	0.68	0.65	0.59
Medium level	0.67	0.60	0.56
High level	0.61	0.73	0.70

Table 3: Correlation between human ratings and LLM-generated ratings across literacy levels

User Preference This dimension captures how a response is subjectively perceived by users with different health literacy levels. It offers additional insight into how well a counterspeech message connects with its intended audience, for example, how people with low literacy levels view the effectiveness of messages designed for them, as well as those designed for medium or high literacy audiences. Due to the high cost of recruiting real users with diverse literacy levels, we employ LLMs as evaluators. These LLMs are guided with tailored instructions to simulate users at specific health literacy levels. We use a 1–5 Likert-style scale to assess user preference, with detailed evaluation prompts documented in the Appendix D.2. We further conduct human validation for the user preference evaluation (see details in Appendix E). Before the evaluation, we administer a brief health literacy survey adapted from Rasmussen et al. (2023), consisting of several short questions to assess participants' health literacy levels. Based on the survey results, we recruit participants representing low, medium, and high health literacy groups to ensure diversity and alignment with target user profiles. The results in Table 2 show that the agreement rate among human annotators exceeds 0.85, with Cohen's Kappa values above 0.60, indicating substantial agreement. The agreement rate between human judgments and LLM-generated evaluations is above 0.90, and the Cohen's Kappa exceeds 0.60, indicating a reliable evaluation by the LLM evaluator. In addition, we have conducted a correlation analysis between human ratings and LLM-generated ratings of user preference (see Table 3). The results suggest that the LLM's behavior adapts well to different user profiles. We document the analysis in Appendix E Correlation Analysis.

Politeness Politeness means the degree of respectfulness and courtesy (Song et al., 2025b). Politeness in responses is essential in communication, which helps to foster user engagement (Shan et al., 2022). A polite counterspeech helps avoid potential backlash and is more likely to be accepted by users, as it encourages a respectful tone (Song et al., 2025c; Yue et al., 2024a; He et al., 2023). We utilize the Multilingual Politeness Classification Model, which is a computational tool designed to assess the politeness of responses (Srinivasan and Choi, 2022).

Factual Accuracy It measures the reliability and trustworthiness of the generated response by ensuring that the information provided is correct (Zhou et al., 2024). Presenting scientifically accurate facts in counterspeech can effectively correct misinformation and maintain user trust (Yue et al., 2024a). We employ LLM-based evaluation to assess whether the facts presented in counterspeech are correct. While LLMs exist hallucinations, studies find that when appropriately prompted, they could be strong fact-checkers (Wang et al., 2025; Guan et al., 2024; Chen et al., 2023). Considering both the cost and need for updated information, we prompt gpt-4o-mini-2024-07-18[5] with guidance, such as encouraging web search and explanation generation to assist with fact checking (Prompts detailed in D.3). To validate the reliability of the evaluation, we conduct a human assessment, which achieves a tolerant match score above 0.80 and a Cohen's Kappa above 0.70, indicating strong reliability (See further details in Appendix E).

4.5 Results

We present the evaluation of results in Table 1. There are several findings we conclude as follows.

RAG helps improve the readability and factual accuracy, but shows limited gains in user preference. RAG incorporates customized evidence into generation, which enhances the read-

[5]https://platform.openai.com/docs/models/gpt-4o-mini

ability of responses (e.g., LLaMA-8B: RAG vs. Instructional Prompt: 1.94 vs. 2.74; Qwen-7B: 1.23 vs. 1.66), and factual accuracy (e.g., LLaMA-8B: RAG vs. Instructional Prompt: 0.89 vs. 0.97; Qwen-7B: 0.87 vs. 0.83; LLaMA-1B: 0.66 vs. 0.54). However, its impact on user preference is limited (LLaMA-8B: 0.71 vs. 0.75; Qwen-7B: 0.70 vs. 0.73; LLaMA-1B: 0.34 vs. 0.64). It indicates that the customized evidence may help improve the readability at the syllabus and word level and provide more accurate facts, but the current integration may have limited capability to improve the generation quality, highlighting the need to further optimize the generation module.

Controlled-Literacy achieves higher overall performance. Our Controlled-Literacy framework consistently achieves strong performance across all three evaluation dimensions. For LLaMA-8B, they demonstrate higher politeness (0.84), better alignment with target readability levels (Target Distance: 0.90), strong user preference (0.74), and higher factual accuracy (0.91). Although Instructional Prompt slightly outperforms in user preference (0.75), it shows inferior performance in politeness (0.41), target alignment (2.74), and factual accuracy (0.87). In the case of Qwen-7B and LLaMA-1B, Controlled-Literacy achieves the best overall performance across all metrics: politeness (0.73 and 0.74), target distance (2.06 and 1.31), user preference (0.74 and 0.69), and factual accuracy (0.90 and 0.71). Moreover, we observe that smaller models (e.g., LLaMA-1B) benefit more significantly from Controlled-Literacy, exhibiting greater improvements in overall performance.

5 Cross Evaluation

We employ two evaluation methods to assess our proposed framework. Firstly, we test our methods on *Check-COVID* and *MisinfoCorrect* to examine their generalization ability. These two datasets reflect different sources of misinformation: Twitter and online news media, respectively. Second, we perform cross-evaluation to analyze how users with a particular health literacy level respond to counterspeech generated for other literacy levels. This allows us to demonstrate that our method is more effective when tailored to the intended target health literacy level.

Controlled-Literacy generalizes robustly across datasets. We report the results of the first evaluation in Table 8 and Table 9 in Appendix G. Across the CheckCovid and MisInfoCorrect datasets, the Controlled-Literacy method consistently achieves the best average performance in all four metrics: (1) Politeness: Highest across nearly all model configurations (e.g., 0.85 in LLaMA-8B for CheckCovid, 0.86 in LLaMA-8B for MisInfoCorrect). (2) Target Distance: Consistently the lowest among all baselines, indicating superior alignment with the intended readability level. (3) User Preference: Typically equal to or higher than other methods (e.g., 0.75 in most configurations). (4) Factual accuracy: Always higher than RAG and Instructional Prompt, which indicates that counterspeech generated by Controlled-Literacy is more reliable.

Counterspeech/User	Low User	Medium User	High User
Instructional Prompt			
low	**0.75 (0.03)**	**0.75 (0.04)**	0.68 (0.12)
medium	0.73 (0.08)	0.74 (0.06)	0.74 (0.07)
high	0.55 (0.16)	0.62 (0.15)	**0.75 (0.04)**
RAG			
low	**0.73 (0.09)**	**0.73 (0.10)**	0.50 (0.17)
medium	0.72 (0.11)	**0.73 (0.11)**	0.66 (0.17)
high	0.45 (0.18)	0.51 (0.18)	**0.72 (0.09)**
Controlled-Literacy			
low	**0.74 (0.07)**	**0.74 (0.05)**	0.61 (0.16)
medium	0.69 (0.16)	0.73 (0.10)	0.62 (0.23)
high	0.62 (0.15)	0.63 (0.13)	**0.75 (0.00)**

Table 4: Cross evaluation of user preference.

The improvements in politeness, target distance, and factual accuracy are particularly notable in smaller models (e.g., LLaMA-1B), indicating that literacy control is especially beneficial for low-capacity settings when dealing with informal health misinformation.

User preference peaks when counterspeech matches literacy level. The second evaluation results in Table 4 show that the top preference scores for each user group generally occur when the counterspeech matches the user's literacy level. For instance, in the Instructional Prompt setting, low and high users show the highest preference for low-level (0.75) and high-level (0.75) responses, respectively. Similarly, under the RAG setting, the top scores appear for low (0.73), medium (0.73), and high (0.75) users when aligned with their corresponding levels. The Controlled-Literacy method shows a similar trend, with low users preferring low-level counterspeech (0.74) and high users favoring high-level responses (0.75).

However, there are notable exceptions where users prefer counterspeech designed for different literacy levels. For example, in the Instructional Prompt setting, medium users rated low-level coun-

terspeech slightly higher (0.75) than medium-level (0.74). In the RAG setting, medium users rated both low and medium-level counterspeech equally (0.73). Likewise, in the Controlled-Literacy setting, medium users showed a higher preference for low-level responses (0.74) compared to medium-level (0.73). These findings suggest that while aligning counterspeech with user literacy level generally yields the best user preference outcomes, users may sometimes prefer responses with a lower readability level, possibly due to better accessibility. This inspires us to further optimize our generation strategy by aligning responses more closely with the lower end of the target health literacy range in the future.

6 Qualitative Analysis

To understand why the optimized counterspeech achieves higher performance across politeness, target distance, and factual accuracy, yet still falls short of a perfect user preference score, we conduct a qualitative analysis with human experts. We engage two experts in misinformation studies to review and analyze the reasons, identify their strengths, and highlight remaining limitations. We randomly select 50 samples from the best-performing model, Controlled-Literacy, using LLaMA3.1-8B-Instruct, for high health literacy level of users.

They summarize several key elements that make the counterspeech achieve high performance: (1) Evidence-based reasoning: most of the counterspeech cite scientific sources (CDC, WHO, peer-reviewed studies) and explain the details, which matches the expectations of rational justification for high health literacy level of users. (2) Precise terminology. The counterspeech uses domain-specific vocabulary, such as "pathogenesis," "clinical trials" and so on. (3) Structured argumentation. The counterspeech often follows a consistent structure: acknowledgement $\rightarrow$ clarification $\rightarrow$ evidence $\rightarrow$ implication $\rightarrow$ recommendation, satisfying expectations for logical coherence and depth.

However, our best-performing responses also exhibit several shortcomings that need further improvement. (1) Repetitive openings. Many counterspeech begin with similar phrases (e.g., "Thank you for sharing..."), which lack lexical variety. It tends to be less human-like, potentially decreasing the effectiveness of counterspeech. (2) Heavy information density. While users with high health literacy generally prefer more technical content, some responses are overly dense and packed with information. It may diminish clarity and fatigue readers. (3) Detached tone. Although the counterspeech maintains a coherent and logically structured academic tone, it often lacks emotional resonance or personal storytelling. Incorporating more empathetic or narrative elements could enhance reader engagement and strengthen the persuasive impact.

7 Conclusion

We propose a Controlled-Literacy method to create counterspeech that matches users' health literacy levels when addressing health misinformation. This framework enables the integration of RAG and RL to generate accessible and audience-appropriate counterspeech. To control the evidence used during generation, we construct a knowledge base that contains diverse evidence and filter the evidence for each target group after retrieval. Our reward design combines user preference and readability, ensuring the generated content aligns with the health literacy needs of different user groups. Experimental results show that Controlled Literacy produces counterspeech that is more accessible, polite, user-preferred by intended groups, and factually accurate. Furthermore, cross-generalization experiments demonstrate the robustness of our method across various types of health misinformation.

Limitations

Limited coverage in retrieved knowledge. Although we collect knowledge from diverse sources, our current approach does not incorporate real-time information, which is critical for addressing dynamic and evolving health misinformation. In the future, we will incorporate web search into our generation to enhance the timeliness and relevance of retrieved knowledge.

Need for finer-grained user group categorization. In our study, we only consider three health literacy level. However, real-world users exhibit a wide range of characteristics and information needs. We aim to develop a more nuanced and detailed user segmentation strategy to better align with diverse health literacy profiles.

Insufficient evaluation framework. Our evaluation mostly relies on computing-based measurement. While informative, these do not fully capture

how real users perceive or benefit from the generated counterspeech. Moving forward, we plan to incorporate more human evaluations from diverse user groups to obtain a comprehensive assessment of counterspeech effectiveness.

Ethics Statement

We ensure that our study adheres to ethical guidelines by carefully evaluating associated risks and benefits. We collect data from Reddit under Reddit's Terms of Service using PRW API. Reddit is a public forum. When users sign up to Reddit, they consent to make their data available to the third party, including the academy. Therefore, we can use Reddit data without further seeking user consent following the ethical rules (Procter et al., 2019). We have masked users' identifiable information before analysis and modeling. We will make sure the dataset is exclusively used for non-commercial research purposes[6]. We acknowledge the potential risks of users being re-identified with anonymized data or misuse of the data by individuals, but the benefits will outweigh such risks.

References

Anirban Saha Anik, Xiaoying Song, Elliott Wang, Bryan Wang, Bengisu Yarimbas, and Lingzi Hong. 2025. Multi-agent retrieval-augmented framework for evidence-based counterspeech against health misinformation. *arXiv preprint arXiv:2507.07307*.

Tal August, Kyle Lo, Noah A Smith, and Katharina Reinecke. 2024. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the" general" audience. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Centers for Disease Control and Prevention. 2024. Understanding health literacy. https://www.cdc.gov/health-literacy/php/about/understanding.html. Accessed: 2025-04-23.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069*.

Xuewei Chen, Jennifer L Hay, Erika A Waters, Marc T Kiviniemi, Caitlin Biddle, Elizabeth Schofield, Yuelin Li, Kimberly Kaphingst, and Heather Orom. 2018. Health literacy and use and trust in health information. *Journal of health communication*, 23(8):724–734.

Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544.

Ilona Fridman, Skyler Johnson, and Jennifer Elston Lafata. 2023. Health information and misinformation: a framework to guide research and practice. *JMIR medical education*, 9:e38687.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4523–4536.

Lingzi Hong, Xiaoying Song, Anirban Saha Anik, and Vanessa Frias-Martinez. 2025. Dynamic fusion of large language models for crisis communication. In *Proceedings of the International ISCRAM Conference*.

Sinan Ibrahim, Mostafa Mostafa, Ali Jnadi, Hadi Salloum, and Pavel Osinenko. 2024. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. *IEEE Access*.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, and 1 others. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Tagore Rao Kosireddy, Jeffrey D Wall, and Evan Lucas. 2024. Exploring the readiness of prominent small language models for the democratization of financial literacy. *arXiv preprint arXiv:2410.07118*.

[6] https://www.reddit.com/wiki/api-terms/

Hsin-Yu Kuo and Su-Yen Chen. 2025. Predicting user engagement in health misinformation correction on social media platforms in taiwan: Content analysis and text mining study. *Journal of Medical Internet Research*, 27:e65631.

Chenxi Liu, Dan Wang, Chaojie Liu, Junnan Jiang, Xuemei Wang, Haihong Chen, Xin Ju, and Xinping Zhang. 2020. What is the meaning of health literacy? a systematic review and qualitative synthesis. *Family medicine and community health*, 8(2):e000351.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to tolkien: Controlling the language proficiency level of llms for content generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15670–15693.

Leia Martínez Silvagnoli, Caroline Shepherd, James Pritchett, and Jason Gardner. 2022. Optimizing readability and format of plain language summaries for medical research articles: cross-sectional survey study. *Journal of medical Internet research*, 24(1):e22122.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE international Conference on big data (big data)*, pages 748–757. IEEE.

Don Nutbeam. 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health promotion international*, 15(3):259–267.

Kenny Peng and James Grimmelmann. 2024. Rescuing counterspeech: A bridging-based approach to combating misinformation. *Available at SSRN 4993772*.

Nelson Perez-Rojas, Saul Calderon-Ramirez, Martin Solis-Salazar, Mario Romero-Sandoval, Monica Arias-Monge, and Horacio Saggion. 2023. A novel dataset for financial education text simplification in spanish. *arXiv preprint arXiv:2312.09897*.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10(11):e38095.

Rob Procter, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matt Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. *arXiv preprint arXiv:1908.11732*.

Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.

Stinne Eika Rasmussen, Anna Aaby, Anne Søjbjerg, Anna Mygind, Helle Terkildsen Maindal, Olli Paakkari, and Kaj Sparle Christensen. 2023. The brief health literacy scale for adults: adaptation and validation of the health literacy for school-aged children questionnaire. *International Journal of Environmental Research and Public Health*, 20(22):7071.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do llms adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *arXiv preprint arXiv:2404.07220*.

Rabia Shahid, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. 2022. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC health services research*, 22(1):1148.

Yi Shan, Meng Ji, Wenxiu Xie, Xiaobo Qian, Rongying Li, Xiaomin Zhang, and Tianyong Hao. 2022. Language use in conversational agent–based health communication: Systematic review. *Journal of Medical Internet Research*, 24(7):e37403.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233.

Jana Siebert and Johannes Ulrich Siebert. 2023. Effective mitigation of the belief perseverance bias after the retraction of misinformation: Awareness training and counter-speech. *Plos one*, 18(3):e0282202.

Xiaoying Song, Anirban Saha Anik, Eduardo Blanco, Vanessa Frias-Martinez, and Lingzi Hong. 2025a. A dynamic fusion model for consistent crisis response. *arXiv preprint arXiv:2509.01053*.

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025b. Assessing the human likeness of ai-generated counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3547–3559.

Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2025c. Echoes of discord: Forecasting hater reactions to counterspeech. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4892–4905.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen Mckeown. 2023. Check-covid: Fact-checking covid-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127.

Haining Wang, Jason Clark, Hannah McKelvey, Leila Sterman, Zheng Gao, Zuoyu Tian, Sandra Kübler, and Xiaozhong Liu. 2024a. Science out of its ivory tower: Improving accessibility with reinforcement learning. *arXiv e-prints*, pages arXiv–2410.

Haiyang Wang, Yuchen Pan, Xin Song, Xuechen Zhao, Minghao Hu, and Bin Zhou. 2024b. F2rl: Factuality and faithfulness reinforcement learning framework for claim-guided evidence-supported counterspeech generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4457–4470.

Ling Wang, Jinglin Li, Boyang Zhuang, Shasha Huang, Meilin Fang, Cunze Wang, Wen Li, Mohan Zhang, and Shurong Gong. 2025. Accuracy of large language models when answering clinical research questions: Systematic review and network meta-analysis. *Journal of Medical Internet Research*, 27:e64486.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, and 1 others. 2020. Cord-19: The covid-19 open research dataset. In *Annual Meeting of the Association for Computational Linguistics*.

Neng Wang, Hongyang Yang, and Christina Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

World Health Organization. 2024. Health literacy. Accessed: 2025-04-23.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024a. Evidence-driven retrieval augmented response generation for online misinformation. In *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, pages 5628–5643. Association for Computational Linguistics (ACL).

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024b. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

A Subreddits List

Table 5 presents the list of subreddits analyzed in this study. These subreddits were selected to cover a broad spectrum of discussions on science, public health, medical topics, vaccines, and conspiracy theories.

Subreddit Name
r/science, r/Wuhan_Flu, r/CoronavirusCirclejerk, r/DebateVaccines, r/vaccinelonghaulers, r/ChurchofCOVID, r/vaccineautismevidence, r/ScienceUncensored, r/CoronavirusUS, r/conspiracy, r/medical_advice, r/joerogan, r/Conservative, r/vaccineaddiction, r/thingsprovaxxerssay

Table 5: Subreddit list

B Guidelines and Annotation process for detecting health misinformation

B.1 Annotation guidelines and process

After collecting the Reddit dataset, the next step is to annotate posts into two categories: "Health Misinformation" and "Not Health Misinformation.". To ensure annotation consistency, all annotators received comprehensive guidelines, including examples of health misinformation, classification criteria, and citations of reputable sources (e.g., CDC, WHO, and NIH) for verification purposes.

Annotation guidelines for health misinformation labeling

The goal of this annotation task is to classify posts based on whether they contain health-related

misinformation. Posts will be assigned one of two labels:

Label 1: "Health Misinformation", If the post contains health-related misinformation. Label 0: "Not Health Misinformation", If the post does not contain any health-related misinformation or is unrelated to health information.

Definition of Health Misinformation: Any false, misleading, or unverified claim related to health, medicine, diseases, treatments, vaccines, nutrition, or wellness. Misinformation includes claims that contradict established medical research, public health guidelines, or authoritative sources such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC).

B.2 Human validation

A human validation step was conducted to validate the reliability of these automatically labeled posts. Three annotators independently fact-checked a random sample of 100 model-labeled posts, assessing their accuracy. The agreement rate between the annotators is 88.1%, 89.1%, and 85.1%, and Cohen's Kappa scores are $\kappa \geq 0.73$, $\kappa \geq 0.75$, and $\kappa \geq 0.67$, respectively. This shows substantial agreement between the annotators. For the disagreement, we conducted further discussion and fact-checking, which concluded us with the final label. With the final label, our model agreement rate is 88.1% and Cohen's Kappa, $\kappa \geq 0.73$ demonstrates a substantial agreement between the model's classification and human judgment.

Document Name	Source
COVID-19 Activity Book	Johns Hopkins Children's Center Child Life Program
Covid-19-Plain-Language-Flyer-with-Facemask	Independent Living Center of the Hudson Valley
COVID-19-Vaccines-A-Plain-Language-Guide	HealthMatters Program
FINAL Diagnostic Testing	U.S. Centers for Disease Control and Prevention
Healthy School Year	U.S. Centers for Disease Control and Prevention
Pediatric Testing Materials	Johns Hopkins Children's Center Child Life Program
Plain Language COVID Fact Sheets	Maryland Developmental Disabilities Council
Plain Language Guide for COVID-19_Group-Home	St. Clair County Community Mental Health (SCCCMH)
COVID-19 Teen Info Sheet	U.S. Centers for Disease Control and Prevention
Symptoms Testing	U.S. Centers for Disease Control and Prevention
CDC Global Response to COVID-19_CDC Archive	U.S. Centers for Disease Control and Prevention
CORD-19	COVID-19 Open Research Dataset (CORD-19), Allen Institute for AI
COVID-19-Global-Response-fact-sheet	U.S. Centers for Disease Control and Prevention
FAQ COVID-19 Data and Surveillance_CDC Archive	U.S. Centers for Disease Control and Prevention

Table 6: Knowledge bases categorized by source

C Knowledge Bases

We detail the knowledge bases designed for different health literacy levels in Table 6. We collect knowledge from a diverse set of reliable sources to ensure inclusiveness and representativeness across user groups, referring to previous studies (Song et al., 2025a; Hong et al., 2025; Anik et al., 2025). These sources include federal health agencies such as the Centers for Disease Control and Prevention (CDC), academic and research institutions like the Johns Hopkins Children's Center and the Allen Institute for AI, as well as community organizations including the Maryland Developmental Disabilities Council, St. Clair County Community Mental Health (SCCCMH), and the Independent Living Center of the Hudson Valley. These organizations provide materials written in plain language, often tailored to individuals with diverse health literacy backgrounds.

D Prompt Design

D.1 Instructional Prompt

The Instructional Prompt experiments aim to explore the full potential of LLMs in generating high-quality responses without relying on complementary techniques such as fine-tuning or external knowledge bases. To this end, we explicitly define the task and emphasize it at the beginning of each prompt to ensure clarity. We further design customized prompts tailored to different user groups, considering factors such as language style, evidence presentation, structural organization, and tone. This approach provides detailed and structured guidance to steer the LLMs toward generating responses that are both audience-appropriate and task-specific.

Low Health Literacy

```
<|Target Fkre|>80-100
<|Audience|> Low Health Literacy
<|Task|> Generate Counterspeech
```

You are an expert in health communication and plain language. Your audience has low health literacy — they have only basic reading and writing skills.

Your task is to write a counterspeech response to the following health misinformation, tailored specifically for this audience.

Your response must meet the following criteria:

1. Simple and Clear Language: Use everyday words and short sentences. Avoid medical jargon and complex terms.
2. Evidence-Based: Provide a fact-based correction in a way that's easy to understand.

3. Clarity and Accessibility: Use simple examples or analogies to help explain your point.
4. Polite and Respectful: Be kind and supportive. Do not shame or blame the person who may believe the misinformation.

Your response must be the counterspeech only — do not include any extra explanation or commentary.

Health misinformation to address: `"{comment}"`

Medium Health Literacy

```
<|Target Fkre|>60-79
<|Audience|> Medium Health Literacy
<|Task|> Generate Counterspeech
```

You are an expert in health communication with a focus on individuals with medium health literacy. This audience possesses the cognitive and social skills needed to actively participate in healthcare, communicate effectively with providers, and apply new information to changing circumstances Your task is to generate a counterspeech response to a piece of health misinformation, tailored to this audience.

Your response should meet the following criteria:

1. Clear and Understandable Language: Use plain words and short, simple sentences. Avoid complex grammar. You may include basic medical terms, but explain them clearly and briefly.
2. Evidence-Based Correction: Give a fact-based explanation using trusted health information. Keep the explanation short, logical, and easy to follow.
3. Organized and Structured: Present your response in a simple and clear format. Use short paragraphs or bullet points if needed.
4. Polite and Respectful: Be kind and supportive. Do not shame or blame the person who may believe the misinformation.

Your response must be the counterspeech only — do not include any extra explanation or commentary.

Health misinformation to address: `"{comment}"`

High Health Literacy

```
<|Target Fkre|>0-59
<|Audience|> High Health Literacy
<|Task|> Generate Counterspeech
```

You are an expert in health communication and digital literacy, specializing in engaging audiences with high health literacy who encompasses the ability to critically analyze information, understand social determinants of health, and engage in collective actions to address health disparities. Your task is to generate a counterspeech response to a piece of health misinformation.

Your response should meet the following criteria:

1. Advanced Language: Use precise, nuanced language that reflects the audience's ability to analyze, synthesize, and apply complex health information.
2. Evidence-Based Correction: Correct the misinformation with accurate, research-backed health information.
3. Clarity and Depth: Employ clear, well-structured arguments and sophisticated examples or analogies that resonate with an informed audience.
4. Polite and Respectful: Be kind and supportive. Do not shame or blame the person who may believe the misinformation.

Your response must be the counterspeech only — do not include any extra explanation or commentary.

Health misinformation to address: `"{comment}"`

Health Literacy Level	*Top-k*	Politeness	Target Distance	User Preference	Factual Accuracy
Low	top_10	0.72 (0.25)	1.30 (3.23)	0.71 (0.13)	**0.89**
	top_5	**0.73 (0.25)**	**1.14 (2.97)**	0.71 (0.12)	0.87
	top_3	0.71 (0.25)	1.23 (3.14)	**0.73 (0.09)**	0.84
Medium	top_10	0.66 (0.22)	4.93 (6.81)	0.71 (0.13)	0.81
	top_5	0.70 (0.22)	4.98 (6.76)	0.72 (0.10)	0.87
	top_3	**0.71 (0.23)**	**4.45 (6.35)**	**0.73 (0.11)**	**0.88**
High	top_10	0.41 (0.21)	0.08 (0.75)	**0.72 (0.09)**	**0.90**
	top_5	**0.43 (0.21)**	**0.04 (0.72)**	0.71 (0.10)	0.89
	top_3	0.42 (0.21)	0.42 (2.56)	0.71 (0.10)	0.81

Table 7: Comparison of *Top-k* selections.

D.2 Evaluation Prompt - User Preference

Low Health Literacy

Assume you are a user with low health literacy—someone who may struggle to understand basic health information such as medication labels, appointment slips, or preventive care guidelines. You are presented with a piece of health misinformation and a counterspeech response written for your understanding.

```
Misinformation_Comment:
"{misinfo_comment}"
Counterspeech_Response:
"{counterspeech}"
```

Evaluate the counterspeech based on its clarity and effectiveness in correcting the misinformation.

Use the following 1–5 Likert-style scale:

- 1: Poor – Very difficult to understand and does little to correct the misinformation.
- 2: Fair – Somewhat understandable but leaves confusion or only partially corrects the misinformation.
- 3: Good – Generally clear and corrects the misinformation to a fair extent.
- 4: Very Good – Clear, easy to understand, and effectively corrects the misinformation.
- 5: Excellent – Extremely clear, very easy to understand, and completely corrects the misinformation.

Provide only the score (an integer from 1 to 5) as your final output.

Medium Health Literacy

Assume you are a user with medium health literacy—someone who can understand and act on straightforward health information but may struggle with complex or abstract concepts. You are presented with a piece of health misinformation and a counterspeech response written for your understanding.

```
Misinformation_Comment:
"{misinfo_comment}"
Counterspeech_Response:
"{counterspeech}"
```

Evaluate the counterspeech based on its clarity and effectiveness in correcting the misinformation.

Use the following 1–5 Likert-style scale:

- 1: Poor – Overly complex or ambiguous, difficult to understand and fails to correct the misinformation.
- 2: Fair – Somewhat clear but includes complexity that hinders full understanding.
- 3: Good – Generally clear and reasonably corrects the misinformation, though some parts may be slightly complex.
- 4: Very Good – Clear, straightforward, and effectively corrects the misinformation with minimal complexity.
- 5: Excellent – Extremely clear, easy to understand, and fully corrects the misinformation in an accessible way.

Provide only the score (an integer from 1 to 5) as your final output.

High Health Literacy

Assume you are a user with high health literacy—someone capable of analyzing, synthesizing, and applying complex health information across diverse contexts. You are presented with a piece of health misinformation and a counter-speech response written for your understanding.

`Misinformation_Comment: "{misinfo_comment}"`
`Counterspeech_Response: "{counterspeech}"`

Evaluate the counterspeech based on its clarity and effectiveness in correcting the misinformation.

Use the following 1–5 Likert-style scale:

- 1: Poor – Oversimplified or incomplete, lacking sufficient depth to correct the misinformation.
- 2: Fair – Addresses the misinformation but misses nuances or provides a partial correction.
- 3: Good – Generally clear and corrects the misinformation adequately, though some complexity may be missing.
- 4: Very Good – Clear, comprehensive, and effectively corrects the misinformation with a well-balanced explanation.
- 5: Excellent – Extremely clear, insightful, and provides a nuanced, well-supported correction that fully addresses the complexities.

Provide only the score (an integer from 1 to 5) as your final output.

D.3 Evaluation Prompt - Factual Accuracy

You are an expert fact-checker. Your task is to determine whether the following counter-speech is factually correct. You may search the web to verify the claims made in the response.

Counter-Speech Response:
`"{model_response}"`

Evaluation Instructions:

- If the counter-speech is factually correct and does not contain misinformation, output: `Label: 1`
- If the counter-speech contains false or misleading claims, output: `Label: 0`

Provide the label and explanations.

Output Format:
`Label: (0 or 1)`
`Explanations:`

E Evaluation Validation

Considering that **user preference** evaluations rely on LLMs, which may introduce potential bias and may not fully capture real users' perspectives, we additionally conduct a human evaluation. We recruit six annotators representing low, medium, and high health literacy levels. To select annotators with diverse health literacy, we administer a brief screening survey following the method proposed by Rasmussen et al. (2023). Using this method, respondents rate each item on a 4-point scale using the following categories: Not at all true (1 point), Not completely true (2 points), Somewhat true (3 points), and Absolutely true (4 points). A cumulative score is then calculated. Based on the scoring criteria from the HLSAC, the initial version of the B-HLA categorizes health literacy as follows: low health literacy (10–25 points), moderate health literacy (26–35 points), and high health literacy (36–40 points). Two annotators are selected from each health literacy category, resulting in a balanced sample across literacy levels. All annotators are proficient in English and have no prior involvement in the project. Annotators are compensated at a rate of $10 per hour, in accordance with ethical guidelines for human-subject research. Each annotator completes the evaluation in approximately two hours. The final annotator group consisted of individuals aged 15 to 35 years, with educational backgrounds ranging from high school to graduate-level studies. The group includes three female and three male participants, all based in the United States. All annotators reported regular access to online health information, though their confidence and comprehension varied, as reflected in the screening tool scores.

We randomly sample 50 health misinformation paired with low/medium/high counterspeech generated by Instructional Prompt using LLaMA3.2-

Method	Literacy Level	Politeness	Target Distance (↓)	User Preference	Factual Accuracy
LLaMA-8B					
Instructional Prompt	low	0.52 (0.19)	2.69 (4.25)	0.74 (0.06)	0.87
	medium	0.60 (0.17)	4.94 (7.49)	0.74 (0.08)	0.86
	high	0.51 (0.12)	0.03 (0.56)	0.74 (0.07)	0.76
	Avg.	**0.54 (0.16)**	**2.55(4.10)**	**0.74 (0.07)**	**0.83**
RAG	low	0.72 (0.19)	2.11 (4.06)	0.74 (0.06)	0.82
	medium	0.74 (0.16)	3.21 (5.82)	0.74 (0.07)	0.85
	high	0.57 (0.13)	0.05 (0.60)	0.70 (0.11)	0.88
	Avg.	**0.68 (0.16)**	**1.79 (3.49)**	**0.73 (0.08)**	**0.85**
Controlled-Literacy	low	0.82 (0.12)	0.74 (1.90)	0.75 (0.00)	0.92
	medium	0.75 (0.15)	2.40 (4.26)	0.75 (0.02)	0.91
	high	0.99 (0.00)	0.00 (0.00)	0.75 (0.02)	0.96
	Avg.	**0.85 (0.09)**	**1.05 (2.05)**	**0.75 (0.01)**	**0.93**
LLaMA-1B					
Instructional Prompt	low	0.68 (0.21)	4.36 (7.40)	0.65 (0.19)	0.57
	medium	0.50 (0.19)	17.29 (11.74)	0.55 (0.23)	0.59
	high	0.84 (0.17)	0.06 (0.60)	0.65 (0.16)	0.46
	Avg.	**0.67 (0.19)**	**7.24 (6.58)**	**0.62 (0.19)**	**0.54**
RAG	low	0.52 (0.19)	16.90 (14.68)	0.41 (0.24)	0.69
	medium	0.55 (0.15)	9.19 (10.50)	0.43 (0.23)	0.59
	high	0.57 (0.17)	0.11 (1.27)	0.43 (0.21)	0.51
	Avg.	**0.55 (0.17)**	**8.73 (8.82)**	**0.42 (0.23)**	**0.60**
Controlled-Literacy	low	0.80 (0.14)	2.29 (3.69)	0.71 (0.13)	0.69
	medium	0.79 (0.17)	2.22 (3.81)	0.72 (0.10)	0.74
	high	0.83 (0.27)	0.05 (0.73)	0.68 (0.13)	0.75
	Avg.	**0.81 (0.19)**	**1.52 (2.74)**	**0.70 (0.12)**	**0.73**
Qwen-7B					
Instructional Prompt	low	0.46 (0.18)	4.93 (7.12)	0.73 (0.10)	0.84
	medium	0.59 (0.18)	4.79 (7.69)	0.74 (0.11)	0.95
	high	0.53 (0.13)	0.58 (6.05)	0.74 (0.10)	0.97
	Avg.	**0.53 (0.16)**	**3.43 (6.95)**	**0.74 (0.10)**	**0.92**
RAG	low	0.58 (0.17)	5.20 (6.98)	0.74 (0.07)	0.88
	medium	0.64 (0.16)	7.09 (8.22)	0.75 (0.04)	0.95
	high	0.63 (0.14)	0.03 (0.38)	0.73 (0.08)	0.92
	Avg.	**0.62 (0.16)**	**4.11 (5.19)**	**0.74 (0.06)**	**0.92**
Controlled-Literacy	low	0.79 (0.16)	1.94 (3.71)	0.75 (0.04)	0.90
	medium	0.62 (0.13)	1.36 (2.67)	0.75 (0.04)	0.97
	high	0.88 (0.18)	0.00 (0.00)	0.74 (0.05)	0.98
	Avg.	**0.76 (0.16)**	**1.10 (2.13)**	**0.75 (0.04)**	**0.95**

Table 8: Cross generalization performance on **CheckCovid Dataset**. The best overall performance in each category is highlighted in gray.

1B-Instruct. The evaluation guidelines provided to the LLM are also given to the human annotators to ensure consistency. To assess inter-annotator agreement, we compute both a tolerant match rate, acknowledging the difficulty of achieving exact agreement on a 1–5 scale, and the weighted Cohen's Kappa, which accounts for the degree of disagreement by penalizing larger rating discrepancies more heavily.

Additionally, we recruit two annotators to assess the **factual accuracy** of the generated counterspeech. Annotators are instructed as follows: You are an expert fact-checker. Your task is to evaluate whether the following counter-speech is factually correct. You may search the web to verify the claims made in the counter-speech.

Evaluation Instructions:

* If the counterspeech is factually correct and does not contain misinformation, label it as 1.

* If the counterspeech contains false or misleading claims, label it as 0.

You may consult fact-checking sources such as Snopes, HealthFeedback, and FactCheck.org to support your judgment. Please provide a brief explanation for each label you assign.

Correlation Analysis We conducted a correlation analysis between human ratings and LLM-generated ratings of user preference using Pearson (Benesty et al., 2009), Spearman, and Kendall's Tau (Kendall, 1938), referring to Shen et al. (2023).

We present the results in Table 3. The results show that Pearson correlation is highest for low and medium literacy levels (0.67–0.68), suggesting the LLM matches human scores most closely in absolute value for users with lower literacy. Spearman and Kendall's Tau are highest at the high literacy level, indicating that the LLM is especially good at ranking outputs in the same order as human annotators at this level, even if the exact scores differ. These findings suggest that the LLM's behavior adapts well to different user profiles: (1) For low and medium literacy users, the focus is

Method	Literacy Level	Politeness	Target Distance (↓)	User Preference	Factual Accuracy
LLaMA-8B					
Instructional Prompt	low	0.41 (0.24)	2.17 (4.03)	0.75 (0.02)	0.98
	medium	0.42 (0.20)	2.94 (5.59)	0.75 (0.01)	0.98
	high	0.29 (0.17)	0.11 (0.93)	0.75 (0.04)	0.99
	Avg.	**0.37 (0.20)**	**1.74 (3.52)**	**0.75 (0.02)**	**0.98**
RAG	low	0.86 (0.15)	0.98 (2.55)	0.75 (0.00)	0.99
	medium	0.72 (0.24)	1.49 (2.84)	0.75 (0.03)	1.00
	high	0.34 (0.18)	0.13 (1.07)	0.74 (0.04)	0.97
	Avg.	**0.64 (0.19)**	**0.87 (2.15)**	**0.75 (0.02)**	**0.99**
Controlled-Literacy	low	0.92 (0.09)	0.78 (1.95)	0.75 (0.00)	1.00
	medium	0.68 (0.23)	0.89 (2.45)	0.75 (0.00)	1.00
	high	0.97 (0.02)	0.00 (0.00)	0.75 (0.00)	1.00
	Avg.	**0.86 (0.11)**	**0.56 (1.47)**	**0.75 (0.00)**	**1.00**
LLaMA-1B					
Instructional Prompt	low	0.73 (0.28)	12.41 (12.03)	0.64 (0.18)	0.62
	medium	0.47 (0.32)	15.82 (16.05)	0.59 (0.22)	0.72
	high	0.82 (0.18)	0.14 (2.54)	0.68 (0.15)	0.78
	Avg.	**0.67 (0.26)**	**9.46 (10.21)**	**0.64 (0.18)**	**0.71**
RAG	low	0.43 (0.30)	15.53 (12.60)	0.53 (0.23)	0.68
	medium	0.36 (0.25)	5.59 (7.89)	0.48 (0.23)	0.79
	high	0.33 (0.23)	0.51 (3.78)	0.49 (0.21)	0.76
	Avg.	**0.37 (0.26)**	**7.21 (8.09)**	**0.41 (0.22)**	**0.74**
Controlled-Literacy	low	0.84 (0.16)	4.54 (5.18)	0.68 (0.17)	0.74
	medium	0.61 (0.23)	1.05 (2.30)	0.72 (0.08)	0.70
	high	0.92 (0.20)	0.00 (0.00)	0.68 (0.13)	0.81
	Avg.	**0.79 (0.20)**	**1.86 (2.49)**	**0.69 (0.13)**	**0.75**
Qwen-7B					
Instructional Prompt	low	0.31 (0.22)	3.04 (5.74)	0.74 (0.06)	0.91
	medium	0.45 (0.23)	4.74 (10.06)	0.74 (0.10)	0.97
	high	0.38 (0.21)	0.02 (0.41)	0.74 (0.05)	1.00
	Avg.	**0.38 (0.22)**	**2.60 (5.40)**	**0.74 (0.07)**	**0.96**
RAG	low	0.59 (0.22)	3.96 (6.38)	0.75 (0.01)	0.97
	medium	0.52 (0.22)	4.84 (7.15)	0.75 (0.03)	0.96
	high	0.49 (0.26)	0.02 (0.17)	0.74 (0.04)	1.00
	Avg.	**0.53 (0.23)**	**2.94 (4.57)**	**0.75 (0.03)**	**0.98**
Controlled-Literacy	low	0.78 (0.19)	2.32 (2.99)	0.75 (0.00)	1.00
	medium	0.43 (0.16)	1.23 (2.53)	0.75 (0.03)	0.99
	high	0.77 (0.30)	0.00 (0.00)	0.75 (0.00)	1.00
	Avg.	**0.66 (0.22)**	**1.18 (1.84)**	**0.75 (0.01)**	**1.00**

Table 9: Cross generalization performance on **MisinfoCorrect Dataset**. The best overall performance in each category is highlighted in gray.

on numeric simplicity and readability, and LLM-generated scores align well with human ratings in both value and order. (2) For high literacy users, while the Pearson correlation is slightly lower, the LLM captures the ranking preferences of more sophisticated users very effectively.

F *Top-*k Comparison

We use the LLaMA3.1-8B-Instruct model within a RAG framework to examine the impact of Top-k evidence selection. Results in Table 7 reveal the following: for users with low health literacy, *Top-5* achieves the highest politeness score (0.73) and lowest target distance (1.14), *Top-3* yields the highest user preference score (0.73), and *Top-10* leads in factual accuracy (0.89). Given the relatively small differences in politeness, user preference, and target distance but a more substantial advantage in factual accuracy, we select *Top-10* for users with low health literacy. For users with medium health literacy, *Top-3* achieves the best overall performance across all dimensions: politeness (0.71), target distance (4.45), user preference (0.73), and factual accuracy (0.88), and is therefore selected. In the high health literacy setting, *Top-10* is chosen as it demonstrates the best performance in user preference (0.72) and factual accuracy (0.90).

G Cross Generalization Results

We apply our methods, Controlled-Literacy, to two distinct misinformation datasets: Check-COVID and MisinfoCorrect mentioned in Section 5. The results are in Table 8 and Table 9.

H Computing Resources

The computational resources applied in this research include a high-performance server equipped with an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 GPUs.

I Use of AI Assistants

We acknowledge the use of AI tools to assist with code writing and expression refinement. The authors developed all core ideas, methods, analyses, and conclusions. The final content reflects the authors' independent scholarly contributions.