Ask or Answer: A Decision Framework for Multi-Turn Health Misinformation Intervention

Xiaoying Song[1] **Anirban Saha Anik**[1] **Jinyu Liu**[1]
Qitao Tan[2] **Geng Yuan**[2] **Lingzi Hong**[1]
[1] University of North Texas
[2] University of Georgia
{xiaoyingsong, anirbansahaanik, jinyuliu}@my.unt.edu, lingzi.hong@unt.edu
{qitaotan, geng.yuan}@uga.edu

Abstract

Correcting health misinformation in dialogue requires more than producing a factual rebuttal: users differ in what they know, what they believe, and what they need to hear, so an effective intervention often depends on first asking the right clarifying question. Yet existing methods either respond immediately or probe indiscriminately, treating clarification as either unnecessary or always beneficial. We propose **Reward-Optimized Probe-and-Respond (RO-PnR)**, a framework that learns *when asking is worth its cost*. At each turn, RO-PnR chooses between probing for more information and committing to a final correction, guided by a turn-level reward that weighs the expected gain from probing against its interaction cost. To capture how user heterogeneity affects probing value, we model each simulated user with a latent state along health literacy and belief commitment. Experiment shows that RO-PnR achieves the highest cost-adjusted utility across three health-misinformation datasets and three base models, using ∼30% fewer turns than always-probe baselines.

1 Introduction

Health misinformation distorts public health, undermining vaccine uptake and adherence to public-health guidance (Van Der Linden, 2022), and dialogue-based agents are an increasingly promising channel for correcting it (Karinshak et al., 2023). Effective correction depends on more than producing accurate facts; it depends on what the specific user knows, what they believe, and why they believe it (Ecker et al., 2022b).

Prior work on health misinformation correction has focused on the response itself: generating evidence-grounded rebuttals (Yue et al., 2024; Anik et al., 2025) or tailoring them to user characteristics such as health literacy (Song et al., 2025; Peng and Grimmelmann, 2024). These methods improve the quality of individual responses, but they commit to a reply based on the original post alone, without engaging the user to surface the underlying concern. A separate line of work on multi-turn dialogue uses clarification questions to handle ambiguity (Chen et al., 2024; Wu et al., 2025), with recent decision-theoretic methods weighing the value of asking against its cost (Dong et al., 2026; Li et al., 2024). However, these approaches typically ask follow-up questions to clarify users' ambiguous queries for task completion, without considering users' underlying state. In health misinformation intervention, users differ in their knowledge and beliefs, so the effects of asking versus answering may vary across users. This makes correction a decision problem, not just a generation problem.

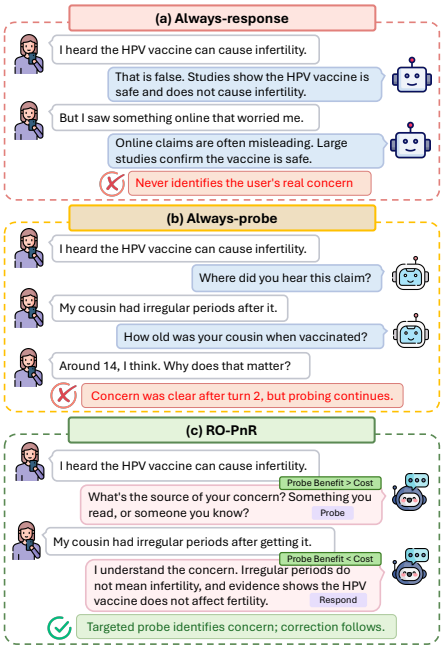

Figure 1: Different strategies on the same misinformation post. RO-PnR weighs benefit vs. cost before acting.

At every interaction turn, the agent must decide whether to respond now with the information it has or ask the user a clarifying question first (Chen et al., 2024; Zhang and Choi, 2025). Responding

too eagerly commits to a generic correction before the user's real concern surfaces (Figure 1a). Asking too eagerly, probing at every opportunity, wastes the user's attention and adds little when the context is already sufficient (Figure 1b). The right behavior depends on whether the expected benefit of one more question outweighs the cost of asking it (Dong et al., 2026).

The trade-off is sharpened by user heterogeneity. Users vary in health literacy: what they can understand and reason about (Nutbeam, 2000; Berkman et al., 2011), and in belief commitment (Ecker et al., 2022b; Wittenberg and Berinsky, 2020): how firmly they hold the misconception. A low-literacy user who is open to revision may need a single targeted probe; a high-literacy user who already articulated their concern may need none (Song et al., 2025); a strong-believer user may need careful probing precisely to gather the context that will make a correction non-confrontational (Swire-Thompson et al., 2022; Lewandowsky et al., 2012). The value of asking, therefore, is itself user-dependent. Agents that don't model this heterogeneity will systematically over- or under-probe.

We propose Reward-Optimized Probe-and-Respond (RO-PnR), a framework that learns when asking is worth its cost. At each turn, RO-PnR chooses between probing and responding based on a turn-level decision reward that weighs the expected gain from probing against the cost of one more question. Because the value of probing depends on hidden user characteristics, we model each simulated user with a latent state along health literacy and belief commitment, and train the policy on turn-level decisions. As illustrated in Figure 1c, RO-PnR learns to ask one well-targeted question when context is missing, and to commit immediately when it is not. We validate this approach across three health-misinformation datasets and three base models. RO-PnR achieves the highest cost-adjusted utility while using 30% fewer turns than always-probe baselines, with the largest gains on low-literacy and strongly-committed users, exactly the slices where the probing decision matters most.

Our contributions are: (a) We frame health misinformation intervention as a probe-or-respond decision problem grounded in domain-specific user dynamics, moving beyond the single-turn, one-size-fits-all paradigm of prior counterspeech work. (b) We introduce **RO-PnR**, a decision-theoretic policy tailored to health misinformation intervention that learns when asking is worth its cost, accounting for user heterogeneity along health literacy and belief commitment.

2 Related Work

2.1 Health Misinformation Intervention

Research on health misinformation intervention has progressed from generic factual rebuttals to retrieval-augmented generation that grounds corrections in scientific evidence (Yue et al., 2024; He et al., 2023) and audience-aware approaches that tailor responses to the user's health literacy or belief commitment (Song et al., 2025; Peng and Grimmelmann, 2024; Anik et al., 2025). These methods improve the quality of individual responses but remain single-turn: the agent commits to a reply based on the misinformation content, without engaging the user to surface the underlying concern or adapting the response as new information emerges. This is particularly limiting in heterogeneous user settings, where the right intervention depends on context (Joseph et al., 2025).

2.2 Decision-Theoretic Probing in Dialogue

Recent work treats clarification as an active decision rather than a default behavior. Some methods train models to choose between asking and answering when a request is ambiguous (Chen et al., 2024; Zhang and Choi, 2025), others use fixed-schedule probing that asks a predetermined set of questions before responding (Fu and Du, 2025), and decision-theoretic approaches weigh the benefit of asking against its communication cost (Dong et al., 2026; Li et al., 2024). Closer to our setting, Wu et al. (2025) introduces multi-turn-aware rewards that estimate the long-term value of a response via simulated future conversations, and Wan et al. (2025) adds a curiosity reward to reduce uncertainty about users' latent state. We build on this direction by combining forward-looking reward estimation with an explicit interaction cost, and by grounding the latent user state in domain-specific dimensions (health literacy and belief commitment) rather than a generic user variable.

2.3 User Heterogeneity Modeling

Recent studies adapt dialogue agents to user-specific traits, either by conditioning responses on explicit profiles (Salemi et al., 2024) or by inferring latent representations from interaction (Wang et al., 2025). Research has shown that responses tailored

to user literacy level are more effective (Song et al., 2025; Peng and Grimmelmann, 2024). Our work differs: (1) We treat the user state as a hidden variable that drives the agent's *decision to probe* and shapes its response; (2) We model two coupled user-state dimensions, health literacy and belief commitment, which jointly shape misinformation susceptibility (Ecker et al., 2022b; Nan et al., 2022) and are key to health misinformation intervention.

3 Methodology

3.1 Task Definition

We formulate the problem as a multi-turn health misinformation intervention task under hidden user heterogeneity. Given a health-misinformation post x, the agent interacts with a user over a short dialogue and follows a RO-PnR paradigm. At each turn, the agent may either *probe* to elicit missing users' concern or *respond* with a final correction.

3.2 RO-PnR Policy

The policy casts intervention as a sequential decision problem: the agent must decide not only *when* to probe (Zhu et al., 2025), but also *when* to stop gathering information and commit to a final response (Dong et al., 2026).

Let the dialogue history at turn t be

$$h_t = \{x, (u_1, a_1), \ldots, (u_{t-1}, a_{t-1}), u_t\},$$

where u_t denotes the current user utterance. Conditioned on h_t, the agent selects an action

$$a_t \sim \pi(\cdot \mid h_t),$$

where π denotes the RO-PnR policy and

$$a_t \in \{\textsc{Probe}, \textsc{Respond}\}.$$

At each turn, the agent must decide whether the current information in h_t is sufficient to generate a final adaptive response, or whether asking one additional clarification question is likely to yield enough benefit to justify its interaction cost. Accordingly, the policy chooses PROBE only when the expected value of acquiring additional user information outweighs the cost of continuing the dialogue; otherwise, it chooses RESPOND.

3.3 Latent User Simulation

The RO-PnR policy operates over the observed dialogue history h_t. However, the evolution of this history depends on how different users respond to the intervention. To capture such hidden heterogeneity, we model each simulated user with a latent state $z \in \mathcal{Z}$ that governs how they react to the agent's actions.

Prior work identifies two dimensions as central to health misinformation intervention: *health literacy* (Berkman et al., 2011; Sørensen et al., 2012; Nan et al., 2022) and *belief commitment* (Ecker et al., 2022a; Walter and Murphy, 2018; Lewandowsky et al., 2012). Health literacy refers to a user's ability to understand and process health-related information, which directly affects how they interpret evidence and make health decisions (Ogbadu-Oladapo et al., 2026). Belief commitment reflects the strength with which a user holds a misinformation-related belief, as well as their resistance to corrective information (Siebert and Siebert, 2023; Wittenberg and Berinsky, 2020); These two dimensions are tightly coupled: A user's response to correction depends on both their ability to understand the evidence and their willingness to revise prior beliefs. Effective intervention must therefore account for both comprehension capacity and openness to belief revision.

Motivated by these findings, we model each user along these two dimensions and define the latent state space as

$$\mathcal{Z} = \mathcal{L} \times \mathcal{B},$$

where $\mathcal{L}$ denotes health literacy and $\mathcal{B}$ denotes belief commitment. Following Nutbeam (2000), we discretize health literacy as

$$\mathcal{L} = \{\textit{functional},\ \textit{interactive},\ \textit{critical}\},$$

corresponding to increasing capacity to understand, apply, and critically evaluate health information. Since there is no established categorization of belief commitment, we draw on prior work (Siebert and Siebert, 2023; Wittenberg and Berinsky, 2020) to define

$$\mathcal{B} = \{\textit{strong},\ \textit{hesitant},\ \textit{open}\},$$

reflecting decreasing resistance to belief revision. Together, $(L, B) \in \mathcal{Z}$ provide a compact characterization of user heterogeneity that supports adaptive intervention. We provide detailed explanations of user modeling in Appendix A and Figure 3.

3.4 Reward Design

We train the RO-PnR policy with a turn-level decision reward (Zhou and Zanette, 2024; Gao et al.,

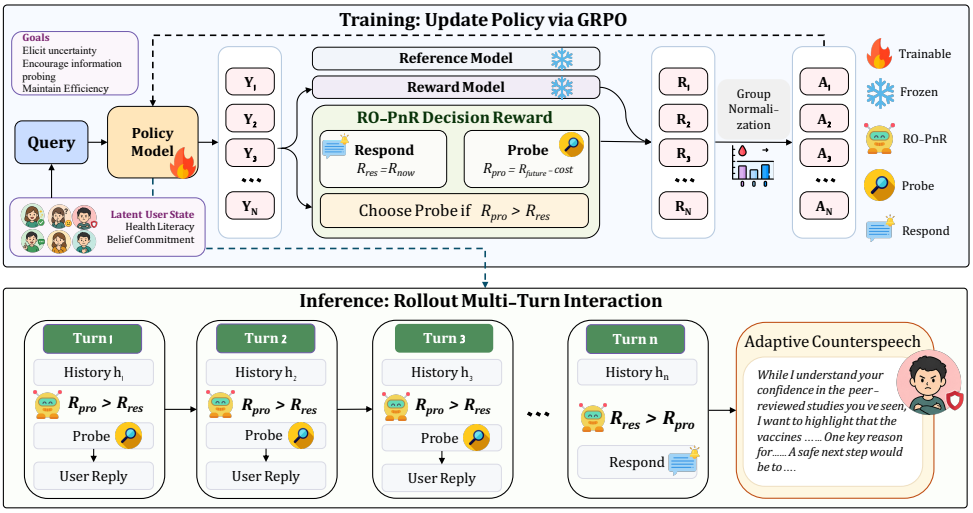

Figure 2: **Overview of the RO-PnR framework.** *Top (training):* the policy generates candidate actions conditioned on the query and hidden user state; a frozen reference and reward model score the dialogue, and the reward compares RESPOND (current quality) against PROBE (expected future quality minus cost). Group-normalized advantages then update the policy via GRPO. *Bottom (inference):* the policy interacts with the user over multiple turns, deciding at each step whether to probe or produce a final adaptive response.

2024) built in two layers, each addressing a distinct aspect of the probe-or-respond trade-off.

(1) Forward-looking value. The value of asking is not in the question itself, but in the better intervention it enables later in the dialogue. A reward that scores only the next utterance misses this: a probe may add little immediately but unlock a substantially long horizon benefit, like better final response (Wu et al., 2025). We therefore compare the response quality available now against the expected quality at conversation end after probing:

$$r_t(\textsc{Respond}) = R(h_t),$$
$$r_t(\textsc{Probe}) = \mathbb{E}\big[\, R(h_{\text{end}}) \mid h_t, \textsc{Probe} \,\big],$$

where $R(\cdot)$ scores a dialogue *Quality* using the average of audience alignment, personalized grounding, and tailored actionability considering dialogue history and response, and h_{end} is the state at which the agent eventually responds. (See Section 4.3 for *Quality* evaluation details.)

(2) Interaction cost. A forward-looking gain alone treats every probe as free, which would incentivize the agent to ask whenever any improvement is expected, even arbitrarily small ones. In practice, however, every clarification question imposes a real cost (Dong et al., 2026). An agent that ignores this cost will over-probe, degrading user experience even when the marginal information gain is negligible. We therefore subtract a per-question cost from the probe action that scales with the number of clarifications already asked, referring to (Dong et al., 2026):

$$r_t(\textsc{Probe}) = R_{\text{future}} - c \cdot n_t,$$

where $c > 0$ is the per-question cost and n_t is the number of clarification questions asked up to turn t. This increases the cost of probing as the dialogue progresses, requiring each additional question to yield greater expected benefit to balance cumulative cost. We further experiment with different cost selections in Section 6.

Putting it together. At each turn, the agent picks the action with the higher reward, choosing PROBE when

$$R_{\text{future}}(h_t) - R(h_t) \;>\; c \cdot n_t,$$

and RESPOND otherwise. Intuitively, the agent probes only when the expected long-horizon improvement from one more clarification outweighs the cumulative interaction cost. Early probes face a low bar, but later probes must justify a higher cost, encouraging the agent to gather information quickly and commit once further probing offers diminishing returns.

3.5 RO-PnR Optimization

We train the RO-PnR policy on a dataset of *turn-level decisions* rather than full dialogues, so that supervision aligns with the policy's per-step probe-or-respond choice. At each decision state extracted

from our multi-turn rollouts, we build paired supervision: a RESPOND record scored by the immediate response quality, and a PROBE record scored by the best continuation quality reachable after one further question. Pairing the two alternatives at the same state lets the policy learn the local trade-off directly. Construction details (rollout sources, grouping by post and profile, filtering) are deferred to Appendix B.

We first warm-start the policy with supervised fine-tuning on the best resolved trajectory in each group, then optimize it with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). GRPO suits our binary turn-level setting because it derives advantages by comparing candidate actions within the same decision group, eliminating the need for a separate value network. Each group is scored with the reward from Section 3.4, normalized into group-relative advantages, and used to update the policy. Because supervision operates directly at the turn level, the policy learns *when* to probe and *when* to respond at each step.

4 Experiment Setup

4.1 Dataset

For fine-tuning and evaluation, we construct multi-turn datasets with a simulated user environment for health misinformation intervention, see user simulation in Section A.3 and dataset construction details in Appendix B. Our primary dataset is derived from CounterHealth (See Appendix C), and is used for model fine-tuning. We additionally use two public health-misinformation datasets for evaluation, MisinfoCorrect(He et al., 2023) and PUBHEALTH (Kotonya and Toni, 2020).

4.2 User Simulation

Due to the cost and difficulty of recruiting real users for large-scale multi-turn evaluation, we use an LLM-based user simulator to generate user actions during interaction. Specifically, we use GPT-4o-mini[1], which has been increasingly adopted as a proxy for simulating user behavior in interactive dialogue settings (Kim et al., 2025; Laban et al., 2025; Liu et al., 2026; Lin et al., 2024). In our setting, the simulator is prompted to emulate users with varying levels of health literacy and belief commitment, allowing us to evaluate how agents respond to different users. The user-simulation prompt is shown in Figure 6, with additional details provided in Appendix A.3.

To evaluate whether the LLM-based user simulator faithfully follows the assigned user profile, we conduct human evaluation along two dimensions: Persona Accuracy and Persona Consistency, referring to (Wang et al., 2025). Persona Accuracy measures whether the simulated user accurately reflects the assigned levels of health literacy and belief commitment. Persona Consistency measures whether the simulated user maintains the assigned health literacy and belief commitment across the dialogue. We instruct human annotators to rate each dimension on a 5-point scale. The detailed annotation rubric is provided in Table 6. The results are discussed in Section 7.

4.3 Evaluation Metrics

We evaluate system performance along three complementary dimensions: *adaptation*, *factual reliability*, and *dialogue quality*, which together capture whether the system tailors the correction to the user, keeps it scientifically accurate, and reaches the goal efficiently. This goes beyond asking whether the final response is correct and also asks whether it is delivered in a way that is adaptive and effective.

Adaptation. Effective health misinformation correction depends not only on *what* is said but on *how* it is framed for a specific user (Krishnamurthy and Hu, 2026; Nguyen et al., 2019). We assess this with three sub-metrics: *Audience Alignment (AA)* (Song et al., 2025; Cima et al., 2025), which measures whether the correction is framed appropriately for the user's inferred health literacy and belief commitment; *Personalized Grounding (PG)* (Gabriel et al., 2024), which measures whether the evidence and reasoning supporting the correction are clear and comprehensible to the user; and *Tailored Actionability (TA)* (Song et al., 2025; Ownby et al., 2026), which measures whether the agent provides safe, concrete next steps suited to the user when needed. All three are scored by an LLM judge (MISTRAL-LARGE-2512[2]) given the dialogue history and the user's latent state. (See Figure 8 in the Appendix for detailed rubrics.)

Factual reliability. A correction that introduces new inaccuracies can undermine user trust and lead to harmful decisions (Yue et al., 2024), so adaptation alone is not sufficient. We report the *factual*

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

[2] https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12

Dataset	Method	Llama-3.1-8B-Instruct							Qwen3-8B							Gemma-4-E4B-it						
		AA	PG	TA	Quality	FER↓	Turns↓	**Utility**	AA	PG	TA	Quality	FER↓	Turns↓	**Utility**	AA	PG	TA	Quality	FER↓	Turns↓	**Utility**
CounterHealth	Single-Turn	0.58	0.53	0.61	0.57	0.18	5.00	0.56	0.58	0.52	0.61	0.57	0.15	5.00	0.56	0.64	0.59	0.67	0.63	0.04	5.00	0.62
	Fixed-Q	0.72	0.65	0.76	0.71	0.11	5.00	0.70	0.71	0.64	0.73	0.69	0.08	5.00	0.68	**0.73**	0.66	0.74	0.71	0.02	5.00	0.70
	Reactive	0.72	0.64	0.76	0.70	0.11	5.00	0.69	0.71	0.64	0.72	0.69	0.09	5.00	0.68	0.72	0.65	0.74	0.71	**0.02**	3.92	0.70
	Confidence	0.71	0.64	0.74	0.69	0.13	**1.26**	0.69	0.67	0.61	0.70	0.66	0.06	**1.42**	0.66	0.72	0.64	0.76	0.71	0.02	**1.85**	0.71
	SFT	0.70	0.64	0.76	0.70	0.07	4.38	0.69	0.70	0.64	0.76	0.70	**0.05**	4.22	0.69	0.70	0.64	0.76	0.70	0.09	4.61	0.69
	RO-PnR (Ours)*	**0.72**	**0.66**	**0.77**	**0.72**	**0.06**	3.56	**0.71**	**0.72**	**0.67**	**0.78**	**0.72**	0.08	3.59	**0.72**	0.73	**0.67**	**0.77**	**0.72**	0.09	3.59	**0.72**
MisinfoCorrect	Single-Turn	0.55	0.50	0.58	0.54	0.07	5.00	0.53	0.55	0.49	0.59	0.54	0.05	5.00	0.54	0.65	0.60	0.67	0.64	**0.01**	5.00	0.63
	Fixed-Q	0.70	0.63	0.72	0.69	0.07	5.00	0.68	0.71	0.63	0.72	0.69	**0.03**	5.00	0.68	0.71	0.64	0.71	0.69	0.03	5.00	0.68
	Reactive	0.71	0.64	0.74	0.70	0.06	4.98	0.69	0.71	0.63	0.71	0.69	0.03	5.00	0.68	0.70	0.63	0.70	0.68	0.03	3.96	0.67
	Confidence	0.61	0.54	0.64	0.60	0.08	**1.10**	0.60	0.59	0.52	0.64	0.58	0.03	**1.08**	0.58	0.63	0.55	0.69	0.62	0.02	**1.35**	0.62
	SFT	0.71	0.65	0.76	0.71	**0.06**	4.01	0.70	0.71	0.65	0.76	0.70	0.07	3.91	0.70	0.71	0.65	0.76	0.71	0.08	4.03	0.70
	RO-PnR (Ours)*	**0.73**	**0.68**	**0.77**	**0.72**	0.07	3.56	**0.72**	**0.74**	**0.68**	**0.77**	**0.73**	0.07	3.46	**0.72**	**0.74**	**0.69**	**0.77**	**0.73**	0.08	3.55	**0.73**
PUBHEALTH	Single-Turn	0.63	0.58	0.65	0.62	0.24	5.00	0.61	0.62	0.56	0.64	0.61	0.21	5.00	0.60	0.66	0.61	0.68	0.65	0.09	5.00	0.64
	Fixed-Q	0.73	0.66	0.77	0.72	0.13	5.00	0.71	0.73	0.67	0.75	0.72	0.12	5.00	0.71	0.74	0.67	0.77	0.73	**0.01**	5.00	0.72
	Reactive	0.74	0.66	0.78	0.73	0.14	4.93	0.72	**0.75**	0.68	0.75	0.73	0.12	5.00	0.72	0.74	0.67	0.77	0.73	**0.01**	3.82	0.72
	Confidence	0.69	0.62	0.72	0.68	0.19	**1.43**	0.68	0.71	0.64	0.73	0.69	0.14	**1.63**	0.69	**0.75**	0.67	**0.79**	**0.74**	0.06	**2.03**	**0.74**
	SFT	0.72	0.66	0.78	0.72	**0.06**	4.23	0.72	0.73	0.67	0.78	0.72	**0.05**	4.24	0.72	0.73	0.67	0.78	0.72	0.09	4.60	0.72
	RO-PnR (Ours)*	**0.74**	**0.69**	**0.79**	**0.74**	0.08	3.53	**0.73**	0.74	**0.69**	**0.79**	**0.74**	0.08	3.80	**0.73**	0.74	**0.69**	0.78	0.74	0.14	3.57	0.73

Table 1: Average results across all health-literacy and belief-commitment configurations, for each method, dataset, and base model. **AA**, **PG**, **TA** are the Audience Alignment (AA), Personalized Grounding(PG), and Tailored Actionability (TA) scores (normalized to $[0, 1]$); **Quality** is the overall quality of AA, PG and TA (normalized to $[0, 1]$); **FER** is the factual error rate (↓); **Turns** is the average number of dialogue turns (↓); and **Utility** is the cost-adjusted Quality at c=0.01, our headline metric (normalized to $[0, 1]$).

error rate: the fraction of responses flagged as containing any factual error by an LLM judge (GPT-5-mini[3]) trained with updated knowledge and have access to web-based verification (He et al., 2023). Lower values indicate better performance. Prompt is in Figure 7 in the Appendix.

Dialogue quality. In addition to adaptation metrics, we also measure the efficiency of the interaction. *Turns* (Wu et al., 2025) counts the number of communication turns, with fewer turns indicating a more efficient interaction. *Utility* (Dong et al., 2026) captures the trade-off between response quality and interaction cost. Since the three adaptation dimensions are measured on the same scale and are jointly necessary for an adaptive response, we define overall *Quality* as their mean, $Q = (S_{\text{AA}} + S_{\text{PG}} + S_{\text{TA}})/3$, and utility as $U = Q - c$, where c is the cost incurred by probe actions.

		CounterHealth				Misinfo				Pubhealth			
Method	**Model**	F	I	C	Ov.	F	I	C	Q.	F	I	C	Ov.
Single-Turn	Llama	0.51	0.60	0.57	0.56	0.49	0.57	0.55	0.53	0.57	0.64	0.62	0.61
	Qwen	0.50	0.59	0.59	0.56	0.48	0.56	0.56	0.54	0.54	0.63	0.62	0.60
	Gemma	0.54	0.66	0.67	0.62	0.56	0.67	0.67	0.63	0.56	0.67	0.68	0.64
Fixed-Q	Llama	0.66	0.72	0.72	0.70	0.62	0.71	0.70	0.68	0.67	0.73	0.73	0.71
	Qwen	0.62	0.71	0.72	0.68	0.63	0.71	0.70	0.68	0.67	0.72	0.73	0.71
	Gemma	0.64	0.72	0.74	0.70	0.63	0.70	0.70	0.68	0.67	0.73	0.75	0.72
Reactive	Llama	0.65	0.72	0.72	0.69	0.64	0.72	0.71	0.69	0.67	0.73	0.74	0.72
	Qwen	0.62	0.71	0.72	0.68	0.62	0.71	0.70	0.68	0.67	0.73	0.74	0.71
	Gemma	0.63	0.72	0.74	0.70	0.62	0.70	0.69	0.67	0.67	0.73	0.76	0.72
Confidence	Llama	0.66	0.72	0.70	0.69	0.60	0.63	0.56	0.60	0.65	0.71	0.67	0.68
	Qwen	0.59	0.71	0.68	0.66	0.57	0.62	0.56	0.58	0.64	0.73	0.70	0.69
	Gemma	0.70	0.72	0.70	0.71	0.66	0.64	0.56	0.62	0.72	0.74	0.75	0.74
SFT	Llama	0.64	0.72	0.72	0.69	0.64	0.73	0.73	0.70	0.67	0.74	0.74	0.71
	Qwen	0.64	0.72	0.72	0.69	0.63	0.73	0.72	0.70	0.66	0.74	0.74	0.72
	Gemma	0.63	0.72	0.72	0.69	0.64	0.73	0.74	0.70	0.66	0.73	0.75	0.72
RO-PnR	Llama	0.65	0.74	0.74	0.71	0.66	0.75	0.75	0.72	0.68	0.76	0.77	0.73
	Qwen	0.66	0.75	0.74	0.72	0.66	0.75	0.75	0.72	0.67	0.76	0.77	0.73
	Gemma	0.65	0.75	0.75	0.72	0.67	0.75	0.76	0.73	0.67	0.75	0.77	0.73

Table 2: Utility by Health Literacy Level (normalized to $[0, 1]$). Higher-value cells are highlighted with light tints: green (≥ 0.70), mint (≥ 0.73), and blue (≥ 0.75). F, I, C, Ov., denotes Functional, Interactive, Critical, and all populations.

4.4 Baselines

We compare our RO-PnR framework against four alternative decision strategies: (a) *Single-Turn Direct Response (Single-Turn)* (Song et al., 2025): The model generates the correction immediately based only on the current turn, without asking clarification questions. (b) *Fixed-Question Probing (Fixed-Q)* (Fu and Du, 2025): The agent asks a predetermined number of clarification questions before producing the final counterspeech response, regardless of the user's feedback. (c) *Reactive Clarifier (Reactive)* (Zhang and Choi, 2025): The agent adaptively chooses between asking and answering based on user feedback, without considering the user's knowledge background or stance toward the claim. (d) *Confidence-Gated Clarifier (Confidence)* (Li et al., 2024): The agent decides whether to ask or answer based on its confidence in a generic inferred user state, not grounded in the health literacy and belief commitment dimensions

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

Method	Model	CounterHealth				Misinfo				Pubhealth			
		S	H	O	Ov.	S	H	O	Ov.	S	H	O	Ov.
Single-Turn	Llama	0.50	0.54	0.64	0.56	0.48	0.52	0.61	0.53	0.55	0.57	0.71	0.61
	Qwen	0.50	0.54	0.64	0.56	0.48	0.52	0.60	0.54	0.53	0.57	0.69	0.60
	Gemma	0.57	0.62	0.69	0.62	0.58	0.62	0.69	0.63	0.58	0.62	0.71	0.64
Fixed-Q	Llama	0.63	0.71	0.75	0.70	0.60	0.69	0.74	0.68	0.65	0.72	0.76	0.71
	Qwen	0.61	0.71	0.73	0.68	0.60	0.70	0.73	0.68	0.65	0.72	0.75	0.71
	Gemma	0.61	0.73	0.76	0.70	0.57	0.71	0.75	0.68	0.65	0.74	0.76	0.72
Reactive	Llama	0.62	0.71	0.76	0.69	0.62	0.70	0.75	0.69	0.66	0.73	0.76	0.72
	Qwen	0.61	0.70	0.74	0.68	0.60	0.69	0.74	0.68	0.65	0.73	0.76	0.72
	Gemma	0.62	0.72	0.76	0.70	0.57	0.69	0.75	0.67	0.65	0.74	0.77	0.72
Confidence	Llama	0.65	0.70	0.73	0.69	0.58	0.59	0.62	0.60	0.65	0.68	0.71	0.68
	Qwen	0.63	0.66	0.69	0.66	0.57	0.57	0.60	0.58	0.64	0.70	0.73	0.69
	Gemma	0.67	0.71	0.73	0.71	0.61	0.62	0.64	0.62	0.69	0.75	0.77	0.74
SFT	Llama	0.62	0.71	0.75	0.69	0.64	0.72	0.76	0.70	0.65	0.73	0.76	0.71
	Qwen	0.63	0.71	0.75	0.69	0.63	0.71	0.75	0.70	0.66	0.73	0.76	0.72
	Gemma	0.62	0.71	0.75	0.69	0.64	0.71	0.75	0.70	0.66	0.73	0.76	0.72
RO-PnR	Llama	0.65	0.72	0.76	0.71	0.66	0.72	0.77	0.72	0.69	0.74	0.77	0.73
	Qwen	0.66	0.72	0.76	0.72	0.67	0.73	0.77	0.72	0.69	0.74	0.77	0.73
	Gemma	0.67	0.72	0.76	0.72	0.69	0.73	0.77	0.73	0.69	0.73	0.77	0.73

Table 3: Utility by Belief Commitment. Higher-value cells are highlighted with light tints: green (≥ 0.70), mint (≥ 0.73), and blue (≥ 0.75); lower values are left uncolored. S, H, O, Ov., denotes Strong, Hesitant, Open, and all populations.

central to misinformation correction. (e) *Supervised Clarifier (SFT)*: A method that learns better to ask and answer via supervised fine-tuning on a higher-quality trajectory, without the turn-level decision reward or GRPO optimization.

4.5 Implementation Details

We evaluate RO-PnR on three open-source 8B models: LLAMA-3.1-8B-INSTRUCT[4], GEMMA-4-E4B-IT[5], and QWEN3-8B[6]. Fine-tuning consists of two stages: offline supervised fine-tuning (SFT) and offline GRPO (Shao et al., 2024). We first perform SFT on base models using full-profile dialogue examples with LoRA adapters (Hu et al., 2021). We then initialize GRPO from the resulting SFT adapter and continue training on turn-level preference examples constructed from all nine user-profile combinations. Training configurations are provided in Table 5.

5 Results

Main Results. RO-PnR achieves the best overall utility while using substantially fewer interaction turns than always-probe baselines (Table 1). RO-PnR obtains the highest utility on nearly every (dataset, model) combination, reaching 0.71–0.73 in utility versus 0.68-0.72 for the strongest baseline (SFT), with consistent gains across all three adaptation dimensions (AA, PG, TA). The baselines reveal a clear trade-off. Single-Turn skips probing entirely and scores lowest (0.53-0.64). Fixed-Q and Reactive almost probe every turn (≥ 4.9 on average) and reach competitive adaptation, but at a high interaction cost. Confidence rarely (1.1-2.0 turns) and suffers the largest quality drop, especially on MisinfoCorrect, where utility falls to 0.58-0.62. RO-PnR occupies the productive middle, using roughly 3.5 turns, about 30% fewer than the always-probe baselines, while delivering the highest utility. Factual error rates remain low across methods (0.06-0.14 for RO-PnR), indicating that adaptation gains do not compromise factual reliability.

Performance Across User Profiles. RO-PnR's utility gains hold across user types, with the largest improvements on the hardest user slices. Tables 2 and 3 break utility down by health literacy and belief commitment. RO-PnR achieves the highest overall utility in every (dataset, model) cell, and its advantage is most pronounced on the most challenging slices. On Functional users, where adaptive framing matters most, RO-PnR reaches 0.65-0.69 versus 0.62-0.67 for SFT. On Strong-believer users, the gap widens further: RO-PnR obtains 0.65–0.69 versus 0.6-0.66 for SFT and only 0.57-0.61 for Confidence, indicating that indiscriminate probing fails to overcome resistance while confidence-only probing fails to gather sufficient context. As literacy rises and belief commitment weakens, all methods improve, and the gap narrows, e.g., on Open users, RO-PnR reaches 0.76-0.77 while baselines cluster in the 0.73-0.76 range, but RO-PnR retains the top spot throughout. These breakdowns confirm that the utility gains in the main results are not driven by a single easy slice but hold across both the comprehension and the openness dimensions of user heterogeneity.

6 Ablation Study

To investigate how different components contribute to RO-PnR performance, we conduct two ablations on the GRPO reward (Table 4). **(a) Reward-horizon ablation** replaces the forward-looking reward with a one-step reward based only on the next user response, testing whether short-horizon credit assignment suffices. **(b) Probe-cost sweep** varies the per-question cost c to probe the trade-off between information gathering and early commit-

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

[5] https://huggingface.co/google/gemma-4-E4B-it

[6] https://huggingface.co/Qwen/Qwen3-8B

Variant	AA ↑	PG ↑	TA ↑	Quality ↑	FER ↓	Turns ↓	Utility ↑
RO-PnR (full)	**0.72**	**0.66**	**0.77**	**0.72**	**0.06**	3.56	**0.71**
(a) Reward-horizon ablation							
immediate next-gain only	0.62	0.54	0.65	0.60	0.12	2.00	0.60
(b) Probe-cost (c) sweep							
probe cost = 0.00	0.50	0.46	0.57	0.51	0.11	5.83	0.51
probe cost = 0.03	0.71	0.65	0.77	0.71	0.08	3.00	0.70
probe cost = 0.05	0.72	0.66	0.77	0.71	0.09	3.00	0.69

Table 4: Ablation study results, averaged over all (health literacy × belief commitment) user profiles. **RO-PnR (full)** is our complete method with forward-looking long-horizon reward, probe cost $c = 0.01$, and cross-threshold bonus. (a) replaces the forward-looking reward with the immediate next-turn gain; (b) sweeps the per-turn probe cost c.

ment.

Table 4 confirms that all components contribute to performance. The reward-horizon ablation degrades utility from 0.71 to 0.60, showing that a probe's value often materializes several turns later and cannot be captured by a one-step reward. The probe-cost sweep shows that $c = 0$ leads to over-probing (5.83 turns) and a quality collapse to 0.51, while moderate costs ($c = 0.03$-0.05) recover near-best performance at $\sim$3 turns, confirming that the policy is stable across a reasonable range of c.

7 Human Validation

While LLMs are known to reliably simulate humans and act as judges at low cost (Thakur et al., 2025; Bavaresco et al., 2025; Huang et al., 2025), we run a small-scale human evaluation along three axes: (1) simulator reliability via persona accuracy and consistency; (2) human-judge agreement on adaptation ratings against MISTRAL-LARGE-2512 (rubric in Figure 8); and (3) a pairwise preference study comparing RO-PNR against the baseline (Due to the annotation cost, we include only the strong baseline method: Fixed-Q.) with six users spanning distinct health literacy and belief commitment profiles. Details in Appendix D.

LLM Simulator Evaluation. Annotators show substantial agreement on both axes (Table 7): mean pairwise agreement is 0.83 for persona accuracy and 0.85 for consistency, with multi-rater Scott's π of 0.70 and 0.68. Using the median of the three annotators' ratings as the final score (Table 8), persona accuracy averages 3.87, with 76% of samples scoring 4 or 5, indicating that simulated users generally reflect the assigned health literacy and belief commitment, though many are judged as largely rather than perfectly aligned. Persona consistency is stronger, averaging 4.64 with 97% of samples at 4 or above, showing that once a persona is established, the simulator maintains it across turns. Details are in Appendix D.1.

Human-judge agreement. Human annotators show moderate-to-strong internal consistency (Scott's π = 0.63–0.78 across dimensions; Table 9). Using the average human rating as consensus, the LLM judge aligns well with humans on *Quality*: Scott's π = 0.55, 73% exact agreement, MAE = 0.23, and 87% of samples within ± 0.5 points (Table 10), with Pearson = 0.57 and Spearman = 0.72 (Table 11). This indicates that the LLM judge is a reliable proxy for human evaluation at scale.

Human pairwise preference. As shown in Figure 4, RO-PnR is strongly preferred overall (79.70% vs. 20.30% for *Fixed-Q*), primarily because its responses are perceived as more explanatory, credible, and actionable. The exception is User 2 (functional literacy, strong belief commitment), who favors *Fixed-Q* for being easier to process, less confrontational, and less reliant on institutional authority, consistent with prior findings that users distrustful of official health sources resist authority-heavy framings even when the evidence is richer (Jamison et al., 2019). This highlights a trade-off: while RO-PNR yields higher-quality corrections in aggregate, users with strong prior commitments may benefit from softer, less institution-centered framing to reduce resistance.

8 Conclusion

In this paper, we introduced RO-PnR, a decision framework for multi-turn health misinformation intervention that learns when to ask clarifying questions and when to correct. By modeling hidden user differences in health literacy and belief commitment, RO-PnR adapts its probing behavior to the needs of each dialogue rather than relying on fixed or excessive questioning. Experiments show that RO-PnR achieves stronger cost-adjusted utility while using fewer interaction turns than always-probe baselines. These results suggest that effective misinformation correction requires not only accurate evidence, but also careful timing: asking when clarification is useful, and answering when enough context has been gathered. Future work should extend this framework with real-user interactions, richer models of user burden, and joint optimization of both when and how to probe.

Limitations

Simulated User Interaction. Because collecting large-scale real-user interactions is costly, we rely on simulated users as a proxy for human behavior. While prior studies suggest that LLM-based simulation can provide a useful approximation (Wang et al., 2025; Sekulić et al., 2024; Bougie and Watanabe, 2025), and we make efforts to improve its reliability, it may still fall short of capturing the full nuance and variability of real users. Despite this limitation, our simulation framework offers a practical testbed for studying how health-misinformation-related user characteristics affect multi-turn counterspeech. Future work should incorporate real-user data and further refine both the simulator and the policy under more realistic interaction settings.

Probe Action Constraints. Our study focuses primarily on the decision of *when* to probe and *when* to stop, emphasizing the probe-and-respond policy rather than the content of the probe itself. As a result, the question of *how* to probe, namely, what clarification question is most appropriate at a particular stage of the dialogue, remains open. Future work could address this limitation by jointly optimizing probe timing and probe content so that the agent can ask more adaptive and informative questions throughout the interaction.

Simplified Cost Modeling. We model communication cost following Dong et al. (2026) to account for the burden imposed by additional clarification turns. However, this formulation is still coarse-grained, assigning a fixed cost at the turn level rather than capturing the more nuanced cognitive load experienced by users. Although our study provides insight into how agents behave when communication cost is incorporated into the decision process, future work might explore more refined ways of measuring user burden and integrating it into the policy for more cognitively aware interaction.

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. In *Second Conference on Language Modeling*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Nicolas Bougie and Narimawa Watanabe. 2025. Simuser: Simulating user behavior with large language models for recommender system evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 43–60.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö Arık. 2024. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. *arXiv preprint arXiv:2406.00222*.

Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. 2025. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *Proceedings of the ACM on Web Conference 2025*, pages 5022–5033.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. 2026. Value of information: A framework for human-agent communication. *arXiv preprint arXiv:2601.06407*.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022a. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022b. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Chuanruo Fu and Yuncheng Du. 2025. First ask then answer: A framework design for ai dialogue based on supplementary questioning with large language models. *arXiv preprint arXiv:2508.08308*.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asuman E Ozdaglar. 2024. Misinfoeval: Generative ai in the era of "alternative facts". In *Proceedings of the 2024 Conference*

on Empirical Methods in Natural Language Processing, pages 8566–8578.

Zhaolin Gao, Wenhao Zhan, Jonathan D Chang, Gokul Swamy, Kianté Brantley, Jason D Lee, and Wen Sun. 2024. Regressing the relative future: Efficient policy optimization for multi-turn rlhf. *arXiv preprint arXiv:2410.04612*.

Valentin Guigon, Lucille Geay, and Caroline J Charpentier. 2026. Rethinking misinformation through plausibility estimation and confidence calibration. *Communications Psychology*, 4(1):24.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895.

Hirono Ishikawa, Takeaki Takeuchi, and Eiji Yano. 2008. Measuring functional, communicative, and critical health literacy among diabetic patients. *Diabetes care*, 31(5):874–879.

Amelia M Jamison, Sandra Crouse Quinn, and Vicki S Freimuth. 2019. "you don't trust a government vaccine": Narratives of institutional trust and influenza vaccination among african american and white adults. *Social science & medicine*, 221:87–94.

Katrina P Jongman-Sereno, Rick H Hoyle, Erin K Davisson, and Jinyoung Park. 2023. Intellectual humility and responsiveness to public health recommendations. *Personality and individual differences*, 211:112243.

Sebastian Joseph, Lily Chen, Barry Wei, Michael Mackert, Iain J Marshall, Paul Pu Liang, Ramez Kouzy, Byron C Wallace, and Junyi Jessy Li. 2025. Decide less, communicate more: On the construct validity of end-to-end fact-checking in medicine. *arXiv preprint arXiv:2506.20876*.

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29.

Paige L Kemp, Aaron C Goldman, and Christopher N Wahlheim. 2024. On the role of memory in misinformation corrections: Repeated exposure, correction durability, and source credibility. *Current Opinion in Psychology*, 56:101783.

Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, Min Hee Kim, Dayi Jung, Kyong-Mee Chung, and Jinyoung Yeo. 2025. Can you share your story? modeling clients' metacognition and openness for LLM therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962, Vienna, Austria. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Parthasarathy Krishnamurthy and Ye Hu. 2026. Covid-19 vaccine framing and acceptance among adults who are vaccine hesitant. *JAMA Network Open*, 9(3):e264114.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.

Xiaoyu Lin, Xinkai Yu, Ankit Aich, Salvatore Giorgi, and Lyle Ungar. 2024. Diversedialogue: A methodology for designing chatbots with human-like diversity. *arXiv preprint arXiv:2409.00262*.

Yu Lu Liu, Hyokun Yun, Tanya Roosta, and Ziang Xiao. 2026. Synthetic users, real differences: an evaluation framework for user simulation in multi-turn conversations. *arXiv preprint arXiv:2605.02624*.

S Emlen Metz. 2023. Building better beliefs through actively open-minded thinking. *The cognitive science of belief*, pages 574–591.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, 314:115398.

Eryn J Newman, Briony Swire-Thompson, and Ullrich KH Ecker. 2022. Misinformation and the sins of memory: False-belief formation and limits on belief revision.

Minh Hao Nguyen, Ellen MA Smets, Nadine Bol, Eugène F Loos, Hanneke WM van Laarhoven, Debby Geijsen, Mark I van Berge Henegouwen, Kristien MAJ Tytgat, and Julia CM Van Weert. 2019. Tailored web-based information for younger and older patients with cancer: randomized controlled trial of a preparatory educational intervention on patient outcomes. *Journal of Medical Internet Research*, 21(10):e14407.

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.

Lydia Ogbadu-Oladapo, Kossi Bissadu, Heejun Kim, and Daniella LaShaun Smith. 2026. Information and health literacy: could there be any impact on health decision-making among adults?—evidence from north america. *Journal of Public Health*, 34(1):151–181.

Raymond L Ownby, Rosemary Davenport, and Joshua Caballero. 2026. User perceptions of individually-tailored health information in digital apps: development of a scale. *Frontiers in Digital Health*, 8:1731948.

Kenny Peng and James Grimmelmann. 2024. Rescuing counterspeech: A bridging-based approach to combating misinformation. *arXiv preprint arXiv:2410.12699*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Greta Arancia Sanna and David Lagnado. 2025. Belief updating in the face of misinformation: The role of source reliability. *Cognition*, 258:106090.

Peter J Schulz and Kent Nakamoto. 2025. Understanding health knowledge failures: uncertainty versus misinformation. *Scientific Reports*, 15(1):23867.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable llm-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35.

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

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, Dibakar Barua, Pengcheng Luo, Junhua Ding, and Lingzi Hong. 2025. Speaking at the right level: Literacy-controlled counterspeech generation with RAG-RL. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2812–2830, Suzhou, China. Association for Computational Linguistics.

Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, and (HLS-EU) Consortium Health Literacy Project European. 2012. Health literacy and public health: A systematic review and integration of definitions and models. *BMC Public Health*, 12(1):80.

Briony Swire-Thompson, Nicholas Miklaucic, John P Wihbey, David Lazer, and Joseph DeGutis. 2022. The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General*, 151(7):1655.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

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

Nathan Walter and Sheila T. Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3):423–441.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. 2025. Enhancing personalized multi-turn dialogue with curiosity reward. *arXiv preprint arXiv:2504.03206*.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107.

Chloe Wittenberg and Adam J Berinsky. 2020. Misinformation and its correction. *Social media and democracy: The state of the field, prospects for reform*, pages 163–198.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. In *International Conference on Machine Learning*, pages 67260–67283. PMLR.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. 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 5628–5643.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with lms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543.

Yifei Zhou and Andrea Zanette. 2024. Archer: training language model agents via hierarchical multi-turn rl. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62178–62209.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

A User Modeling

A.1 Health literacy

Functional health literacy means basic reading/writing and understanding skills sufficient to function effectively in everyday health situations, such as understanding health information and service instructions (Nutbeam, 2000).

Interactive health literacy is more advanced personal, communicative, and social skills that enable independent action on health knowledge (Nutbeam, 2000).

Critical health literacy This is the most advanced level. Nutbeam (2000) defines it in terms of advanced cognitive and social skills used to critically analyze information and act on broader determinants of health at individual and community levels.

A.2 Belief Commitment

Strong belief commitment can be understood as a state in which individuals accept misinformation as true, allow it to continue shaping subsequent attitudes and judgments (Newman et al., 2022), and show difficulty revising that belief even after corrective information is provided (Wittenberg and Berinsky, 2020). In stronger cases, correction may even increase belief in the original misconception, indicating an especially resistant form of commitment (Swire-Thompson et al., 2022).

Hesitant belief commitment is a state in which a person gives a misinformation claim partial acceptance, but does so with noticeable uncertainty or limited confidence (Guigon et al., 2026; Schulz and Nakamoto, 2025), so the belief is not fully consolidated and remains comparatively more open to revision than a strong misinformation belief (Kemp et al., 2024).

Open belief commitment refers to a revisable form of belief commitment in which individuals may initially accept a claim, but remain willing to consider alternative evidence, evaluate source credibility, and update their beliefs when credible corrective information becomes available (Jongman-Sereno et al., 2023; Metz, 2023). This form of commitment is better understood as an orientation toward revisability and evidence-based updating, rather than simply as weak belief (Sanna and Lagnado, 2025).

A.3 User Simulation

We simulate user replies with a profile-conditioned LLM environment. For the main results, each dialogue is paired with a fixed latent user profile

$$u = (\text{health\_literacy}, \text{belief\_commitment}),$$

and the same profile is kept fixed throughout the entire interaction. We evaluate all posts under the full 3×3 grid of user types: health literacy in {*Functional*, *Interactive*, *Critical*} and belief commitment in {*Open*, *Hesitant*, *Strong*}.

The profile dimensions are defined behaviorally rather than demographically. Referring to the definition of Nutbeam (2000), we specify that *Health literacy* controls how complex the user's reasoning sounds: Functional-literacy users prefer simple language and rely more on stories or surface cues; Interactive-literacy users show some interest in proof or source credibility without much technical detail; Critical-literacy users are more likely to refer to evidence quality, mechanisms, study design, or source credibility. Additionally, referring to section A.2, we detail that *Belief commitment* controls how resistant the user is to correction: strong believers defend the claim and are hard to persuade, hesitant users express concern and uncertainty at the same time, and open users are comparatively willing to revise their view if shown credible evidence.

At each turn, the simulator receives the fixed latent state, the dialogue history, and the assistant's latest question, and generates one short reply. In our main setting, the simulator is implemented with `gpt-4o-mini` at temperature 0.3. The prompt explicitly conditions on the latent user state but instructs the model not to reveal it directly. Its structure is in Figure 6.

The prompt further specifies style and behavioral constraints. It requires casual, brief replies in 1–2 sentences, discourages technical medical explanation unless it would arise naturally for that user type, and asks the model to make both health literacy and belief commitment observable through wording, evidence preferences, and openness to update. Separate behavior rules are injected for each profile. For example, functional-literacy users are told to use short everyday wording and focus on personal or practical concerns, whereas critical-literacy users are told to show stronger awareness of evidence quality and source credibility. Strong believers are instructed to sound firm and skeptical of correction, hesitant users to sound torn and uncertain, and open users to sound receptive and non-defensive.

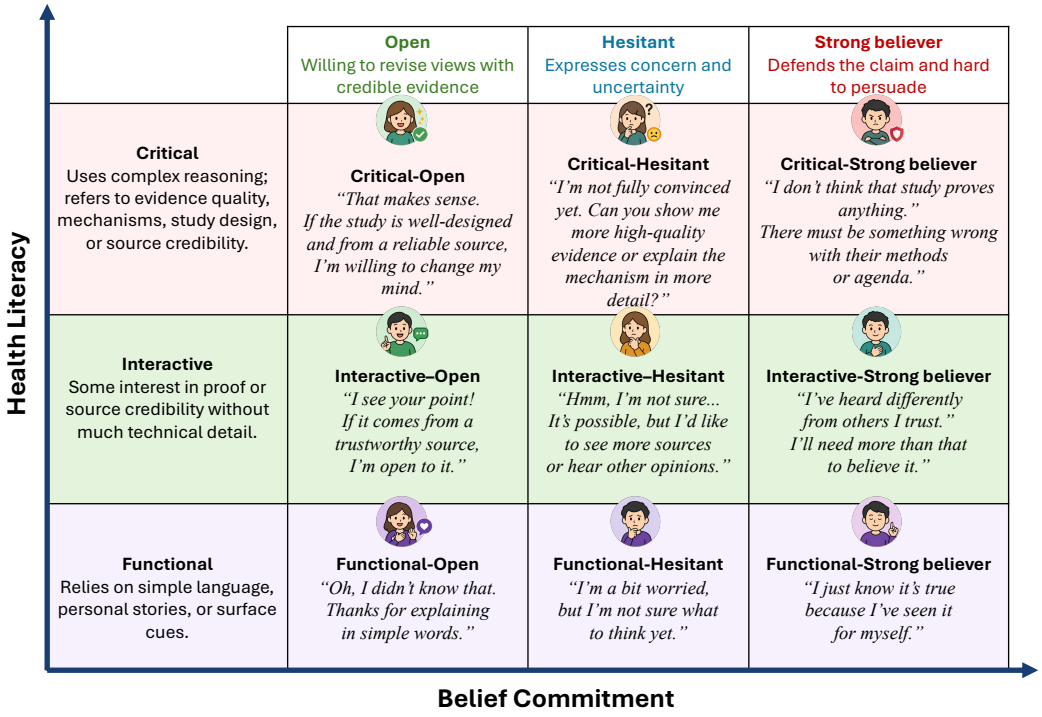

Figure 3: **3×3 latent user profile grid used in our simulation environment.** Users are characterized along two dimensions: *health literacy* (FUNCTIONAL, INTERACTIVE, CRITICAL) and *belief commitment* (OPEN, HESITANT, STRONG BELIEVER). Each cell defines one latent user type and includes an illustrative response style showing how the user may express concern, interpret evidence, and respond to counterspeech. This profile grid is used to simulate heterogeneous user behavior during multi-turn interaction.

Stage	LR	Epochs	Batch	Grad. accum.	Max len.	Other settings
SFT	5×10^{-5}	4	1	8	2048	LoRA $r = 16$, $\alpha = 32$, dropout 0.05; bf16; 100 warmup steps; seed 42
Offline GRPO	5×10^{-6}	2	1	8	2048	$\beta_{\text{KL}} = 0.05$; SFT CE weight 0.2; warmup ratio 0.03; max grad norm 1.0; seed 42

Table 5: Hyperparameters for SFT and GRPO configuration

To reduce role drift, we combine several safeguards. First, the system prompt explicitly says: the model is simulating a social media user, must stay strictly in the *user* role, and must not act as an expert, educator, or assistant. Second, the user prompt reinforces this with strict stylistic constraints: short replies, no belief-state summary, no multiple questions, and no unnecessary factual exposition. Third, the model is required to return JSON only of the form `{"reply": "..."}`, which reduces conversational spillover and makes parsing more robust. Finally, because the latent state is fixed across turns, the simulator is encouraged to remain behaviorally consistent even as the dialogue evolves.

An important setting is that the assistant does not observe these profile labels explicitly. The user profile is hidden from the policy and only affects the environment-side reply generation. The assistant must infer how to adapt from the user's language and reactions, rather than from direct access to the profile metadata.

B Training Dataset Construction

We construct training data to match the reward design at the level of *turn-level decisions*, rather than treating each dialogue as a single terminal example. The starting point is a collection of scored multi-turn rollouts generated for each misinformation post under each of the nine user-profile conditions. For a fixed post–profile pair, we collect multiple stochastic trajectories, so the rollout set can be viewed as a set of *empirical future samples*: some trajectories stop early and respond immedi-

Score	Persona Accuracy	Persona Consistency
1	The simulated user clearly does not match the assigned health literacy or belief commitment.	The user frequently contradicts the assigned profile, with abrupt and unjustified shifts in knowledge level or belief strength.
2	The simulated user weakly matches the assigned profile. One dimension may be partially reflected, but the other is missing or incorrect.	The user shows unstable behavior, with multiple unexplained shifts in health literacy or belief commitment.
3	The simulated user partially matches the assigned profile. Both dimensions are somewhat recognizable, but the dialogue is vague, generic, or only weakly aligned.	The user is mostly stable, but there are noticeable inconsistencies or minor contradictions across turns.
4	The simulated user largely matches the assigned health literacy and belief commitment, with only minor imperfections.	The user maintains the assigned profile across most turns, and any change in belief strength is mostly justified by the conversation.
5	The simulated user strongly matches the assigned profile. Health literacy and belief commitment are both clearly and accurately expressed.	The user consistently maintains the assigned health literacy and belief commitment throughout the dialogue, with any change being natural and well-supported by the interaction.

Table 6: Human evaluation rubric for assessing user simulation quality.

Metric	Persona Accuracy	Persona Consistency
Mean Pairwise Agreement	0.83	0.85
Mean Pairwise Scott's π	0.70	0.68
All-Three Agreement	0.74	0.78
Multi-Rater Scott's π	0.70	0.68

Table 7: Inter-annotator agreement among three evaluators on *persona accuracy* and *persona consistency*. Both pairwise and multi-rater statistics indicate substantial agreement.

Metric	Score 1	Score 2	Score 3	Score 4	Score 5	Mean
Persona Accuracy	0%	3%	21%	62%	14%	3.87
Persona Consistency	0%	0%	3%	30%	67%	4.64

Table 8: Distribution of human ratings for persona accuracy and persona consistency.

Dimension	Scott's π	Mean Pairwise Agreement	All-three Exact Agreement
Audience Alignment	0.76	84.70%	78.00%
Personalized Grounding	0.63	76.00%	65.00%
Tailored Actionability	0.77	88.70%	83.00%
Quality	0.78	86.70%	81.00%

Table 9: Human inter-annotator agreement under 0.5-point discretization.

Dim	Scott's π	Percent Agreement	MAE	Bias	Within ± 0.5	Within ± 1.0
Audience Alignment	0.51	69.00%	0.29	-0.14	86.00%	94.00%
Personalized Grounding	0.31	51.00%	0.28	-0.14	85.00%	96.00%
Tailored Actionability	0.48	72.00%	0.14	-0.07	95.00%	99.00%
Quality	0.55	73.00%	0.23	-0.11	87.00%	97.00%

Table 10: Agreement between the human consensus and the LLM judge across rubric dimensions and the aggregated final score. Bias is computed as consensus minus LLM score.

ately, while others continue probing and reveal the value of obtaining additional user information.

From each trajectory, we extract prefix-level decision states h_t from the first four turns. Each state is converted into a structured assistant target consisting of a latent belief estimate, an action token (PROBE or RESPOND), and the associated natural-language content. Crucially, the rollout trace provides both the value of responding under the *current* information state and the downstream outcomes that become reachable if the dialogue continues. We therefore build two kinds of supervision. A RESPOND record uses the candidate response available at state h_t and is assigned the immediate stopping value R_{now}. A PROBE record uses the clarification question asked at that state and is assigned a forward-looking continuation value based on the best downstream outcome reachable after probing: This framing makes each training example a local decision problem: should the policy stop now, or should it invest one more turn to access a better future response?

To ensure meaningful comparison, we organize records by source post, turn index, and user-profile condition. This groups together alternative trajectories that correspond to closely matched decision contexts, allowing PROBE and RESPOND to be compared locally rather than across unrelated dialogues. We discard groups that are too small or exhibit little reward variation, and then normalize rewards within each group to obtain relative advantages for policy optimization. The resulting dataset is therefore not simply a set of high-quality responses; it is a structured decision dataset in which RESPOND is supervised by the current stopping value, while PROBE is supervised by future-sampled continuation value. This organization makes the subsequent optimization directly

aligned with the intended probe-and-respond behavior.

For SFT, we select the higher *Quality* and factually clean trajectory for each post/profile group. The selected trajectories are converted into step-level supervision examples: given the visible conversation prefix, the model learns to output a structured assistant action, either probe with a clarification question or respond with a final response.

For GRPO, we use scored rollouts from all nine user-profile combinations. Each rollout is decomposed into turn-level decision examples. At each turn, candidate assistant actions are grouped by post, turn index, and profile, and assigned rewards using future reward. Groups with insufficient reward variation or without both probe and respond alternatives are filtered out. The remaining grouped examples provide normalized advantages for offline GRPO, initialized from the SFT adapter.

C CounterHealth Dataset Collection

We collected Reddit posts and comments on COVID-19, influenza, and HIV via the PRAW API[7], using health-related keywords (e.g., "vaccines," "COVID-19," "alternative medicine") to retrieve 4,968 posts and 17,223 comments from high-engagement subreddits. To identify misinformation, five trained annotators from information science backgrounds labeled a 1,000-post sample under a shared annotation guideline, yielding 330 confirmed health-misinformation posts. We then fine-tuned a RoBERTa-large classifier on these annotations (F1 $= 0.76$) and applied it to the remaining posts, surfacing 831 additional candidates. A final filtering pass using GPT-5-mini[8] with web-search-assisted human review produced 769 high-quality health-misinformation posts. On a 100-post validation sample, three annotators showed substantial agreement with the final labels (mean pairwise agreement $= 87.4\%$, Cohen's $\kappa \geq 0.67$), confirming the reliability of the filtering pipeline.

D Human Validation

D.1 LLM Simulator Evaluation

We recruit three PhD students in health informatics to evaluate the quality of the LLM-based user simulator along two dimensions: persona accuracy and persona consistency. Each annotator independently rates 100 sampled simulated dialogues, spanning all 9 user profiles, on a 5-point Likert scale according to the rubric in Table 6. For each annotation instance, annotators are shown the original health misinformation post, the full dialogue history, and the assigned user profile definition, including the target levels of health literacy and belief commitment. User utterances are explicitly marked in the dialogue to encourage annotators to focus on whether the simulated user's replies match the assigned profile. Before the formal annotation, annotators are trained with a small set of example dialogues and discuss the rating criteria to ensure a shared understanding of the two persona dimensions. After independent annotation, cases with substantial rating disagreement are reviewed in a discussion session. A health informatics expert then adjudicates the final label for unresolved cases, which is used for the final analysis.

Each annotator spent approximately three hours completing the rating task. We compensated each participant at a rate of $20 per hour (totaling $60 per annotator), which is above the local minimum wage and consistent with standard compensation for graduate-student annotators in NLP research. All annotators participated voluntarily and provided informed consent prior to the task.

To assess annotation reliability, we compute pairwise agreement and Scott's π among the three annotators on both dimensions (Table 7). Annota-

Dimension	Pearson	Spearman	LLM Mean	Human Mean
Audience Alignment	0.50	0.66	3.76	3.62
Personalized Grounding	0.51	0.67	3.50	3.37
Tailored Actionability	0.72	0.81	3.93	3.86
Quality	0.57	0.72	3.73	3.62

Table 11: Correlation between the LLM judge and human consensus under 0.5-point discretization.

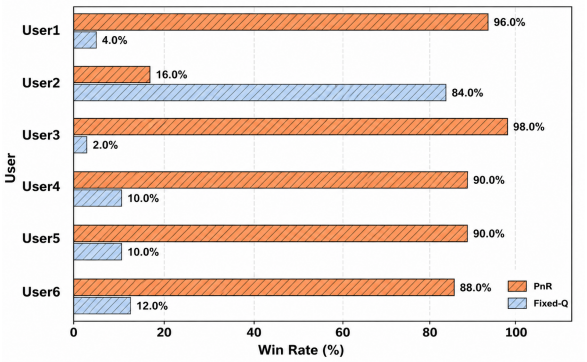

Figure 4: Human pairwise preference comparison between RO-PnR and Fixed-Q across different users. Detailed users' background are provided in Table 13.

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

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

Item	Question	Rating
B	How accurate do you think this claim is?	0 = Definitely false / 10 = Definitely true
C	How certain are you about your answer above?	0 = Not at all / 10 = Completely
O	If you saw credible evidence against your view on this claim, how willing would you be to change your mind?	0 = Not at all willing / 10 = Completely willing

Table 12: Belief Commitment Determinants. B denotes belief, C denotes certainty, and O denotes openness to change.

User	Total FCCHL Mean	HL Category	Overall BC_{total}	BC Category
User 1	2.07	Communicative	2.03	Hesitant
User 2	1.93	Functional	6.83	Strong
User 3	2.71	Communicative	1.20	Open
User 4	3.00	Critical	1.20	Open
User 5	4.00	Critical	0.40	Open
User 6	3.64	Critical	0.35	Open

Table 13: Health Literacy and Belief Commitment Categories. HL categories are mapped from the Total FCCHL mean: 1.00-1.99 = Functional, 2.00- 2.99 = Communicative, and 3.00-4.00 = Critical. BC categories are mapped from BC_{total}: $BC_{\text{total}} < 1.5$ = Open, $1.5 \leq BC_{\text{total}} < 5$ = Hesitant, and $BC_{\text{total}} \geq 5$ = Strong.

tors show substantial agreement on persona accuracy (mean pairwise agreement $= 0.83$, multi-rater Scott's $\pi = 0.70$) and persona consistency (mean pairwise agreement $= 0.85$, multi-rater Scott's $\pi = 0.68$), indicating that the simulator can reliably instantiate the intended user profiles and maintain them across the dialogue.

D.2 Human-judge agreement

We recruit three PhD students in health informatics to assess the reliability of the LLM judge. Following Thakur et al. (2025), we quantify alignment between human annotations and LLM ratings using percent agreement and Scott's π coefficient (Scott, 1955). We sample 100 samples from the results, stratified across (a) the three models (Llama / Gemma / Qwen), (b) all baseline and fine-tuning methods, and (c) the 9 user-profile cells. We first train the annotators on 10 practice samples to familiarize them with the task and scoring rubric. They then rate the responses independently through individual annotation links, without discussion or coordination. Afterward, we identify the cases with the largest disagreement and conduct an adjudication discussion involving the three annotators and a domain expert. The final score range for each such case is then determined based on this review.

The three human evaluators show good internal consistency (Table 9), with Scott's π ranging from 0.63 to 0.78 under 0.5-point discretization, moderate to strong agreement after correcting for chance. We then use the mean of the three human ratings as a consensus and compare it against the LLM judge. The two align reasonably well (Table 10): on the aggregated *Quality* score, Scott's $\pi = 0.55$ with 73% exact agreement after binning, MAE $= 0.23$, and 87% of samples falling within ± 0.5 points. Rank-order correlation is also substantial (Pearson $= 0.57$, Spearman $= 0.72$; Table 11), indicating that the LLM judge and the human consensus rank responses similarly even when their exact scores differ slightly. Agreement is strongest on *Tailored Actionability* and overall *Quality*, while *Personalized Grounding* is the weakest dimension across both inter-human and human-LLM comparisons. Together, these results suggest that the LLM judge is a reliable proxy for human evaluation at scale.

D.3 Human pairwise preference

Due to budget constraints, we recruited six users with varying educational backgrounds: two with middle school education, two with high school education, and two PhD students. This sampling strategy was informed by prior evidence showing a positive association between educational attainment and health literacy (Ishikawa et al., 2008). Each participant spent approximately one hour completing the screening questionnaire and the pairwise preference task, and was compensated with a $25 gift card. This rate is above the U.S. federal minimum wage and consistent with standard compensation for lay-user studies in NLP and HCI research. All participants provided informed consent prior to the study.

We assessed their health literacy using the Functional, Communicative, and Critical Health Literacy (FCCHL) scale (Ishikawa et al., 2008). The FCCHL scale aligns with Nutbeam's categorization of health literacy into functional, communicative, and critical dimensions (Nutbeam, 2000), see questionnaire in Figure 5. Based on the participants' FCCHL scores, we used the sample median to categorize them into lower and higher health literacy groups.

In terms of the belief commitment screening measure, given that there are no established methods available for this purpose, we define belief

commitment as the joint product of how strongly a user accepts a claim and how resistant they are to revising it. Therefore, we capture both components using three 0–10 slider items for each of the three target misinformation claims, as shown in Table 12. For each claim i, we compute a 0-10 commitment score that gates acceptance (B_i) by a strength multiplier formed from certainty (C_i) and the complement of openness (O_i):

$$\text{commitment}_i = B_i \times \frac{C_i + (10 - O_i)}{20}.$$

If the user does not accept the claim, meaning B_i is near 0, then the commitment score is near 0 regardless of certainty. If the user accepts the claim, is highly certain, and is unwilling to revise their view, the commitment score approaches 10. The per-user belief commitment score is the mean across the three claims:

$$BC_{\text{total}} = \frac{1}{3} \sum_{i=1}^{3} \text{commitment}_i$$

where BC_{total} ranges from 0 to 10.

The final screening results are shown in Table 13. Among the six users, three were classified as having a critical HL and open BC background, one as having a communicative HL and open BC background, one as having a communicative HL and hesitant BC background, and one as having a functional HL and strong-believer BC background.

We sample 50 pairs from one of the best baselines (Fixed-Q) and RO-PnR (50*2). Each pair uses the same misinformation post and the same user profile, with the RO-PnR response and the baseline response shown side-by-side, anonymized, in randomized order. We conduct a blind pairwise preference evaluation on 50 health misinformation posts using six users with diverse health literacy levels and belief commitment profiles. For each post, users see only the health misinformation post and two anonymous counterspeech responses, without knowing whether the responses are generated by *RO-PnR* or *Fixed-Q*.

E Use of AI Assistants

We used AI-assisted tools for support with code development and language refinement. All core ideas, methodological decisions, analyses, interpretations, and conclusions were developed by the authors.

Functional, Communicative, and Critical Health Literacy (FCCHL) Questionnaire

Adapted from Ishikawa et al. (2008)

Participant ID: _____ Date: _____

Instructions: The following questions ask about your experience when reading health-related materials or obtaining health information. Please select one option for each item.

Response options

Score	1	2	3	4
Option	Never	Rarely	Sometimes	Often
Meaning	This does not happen	This happens infrequently	This happens occasionally	This happens frequently

Questionnaire items

No.	Dimension	Item	1 Never	2 Rarely	3 Sometimes	4 Often
FHL1	Functional HL	Find characters or words that I cannot read.	☐	☐	☐	☐
FHL2	Functional HL	Feel that the print is too small for me to read.	☐	☐	☐	☐
FHL3	Functional HL	Feel that the content is too difficult for me to understand.	☐	☐	☐	☐
FHL4	Functional HL	Feel that it takes a long time to read health-related materials.	☐	☐	☐	☐
FHL5	Functional HL	Need someone to help me read health-related materials.	☐	☐	☐	☐
CHL1	Communicative HL	Collect health-related information from various sources.	☐	☐	☐	☐
CHL2	Communicative HL	Extract the health-related information I want.	☐	☐	☐	☐
CHL3	Communicative HL	Understand the obtained health-related information.	☐	☐	☐	☐
CHL4	Communicative HL	Communicate my opinion about my illness or health condition.	☐	☐	☐	☐
CHL5	Communicative HL	Apply the obtained health-related information to my daily life.	☐	☐	☐	☐
CrHL1	Critical HL	Consider whether the information is applicable to me.	☐	☐	☐	☐
CrHL2	Critical HL	Consider whether the information is credible.	☐	☐	☐	☐
CrHL3	Critical HL	Check whether the information is valid and reliable.	☐	☐	☐	☐
CrHL4	Critical HL	Collect information to make health-related decisions.	☐	☐	☐	☐

Reference: Ishikawa, H., Takeuchi, T., & Yano, E. (2008). Measuring functional, communicative, and critical health literacy among diabetic patients. Diabetes Care, 31(5), 874-879.

Figure 5: Questionnaire for health literacy screen

System Prompt (`Roleplay / Health Misinformation Persona`):
You are simulating a realistic Reddit user who previously posted a health misinformation claim and is now chatting with an assistant. Stay strictly in the **user** role. Do not act as an expert, educator, or assistant. Respond naturally as a Reddit user with personal opinions and emotions. Keep replies short (1–2 sentences).

Current User State:
`{State.to_dict()}`

User State Dimension Guide:
`{guide_text}`
Interpret `CURRENT_USER_STATE` using the guide above and stay aligned with it.

Conversation So Far:
`{history_text if history_text else '(no prior conversation)'}`

Assistant Message:
`{question}`

Generation Instruction:
Generate exactly one realistic Reddit-style user reply.

User Prompt / Output Constraints (`JSON Reply Formatting`):
Guidelines:

- Speak casually like a real Reddit user.
- Do not summarize the belief state.
- Do not explain medical facts unless the user would naturally know them.
- Do not ask multiple questions.
- Keep the reply brief.
- Remain in character as a user throughout the response and follow the behavior rules carefully.

Behavior Rules:
`{behavior_rules}`

Figure 6: System prompt and user prompt for realistic health misinformation persona simulation.

System Prompt:
You are a professional fact-checking judge. Your task is to determine whether the given content is factually accurate. Use reliable external sources, together with your own knowledge, to verify the response.

Response:
"{final_response}"

Evaluation Instructions:

- Output `Label: 1` if the response is factually correct and does not contain false or misleading claims.
- Output `Label: 0` if the response contains any false, misleading, or unsupported health-related claim.
- Output `Label: UNCERTAIN` if the factual accuracy cannot be determined after verification.
- Provide a brief reason for your judgment in 1–3 sentences.

Output Format:
`Label: 0`, `1`, or `UNCERTAIN`
`Reason:` brief explanation of the judgment

Figure 7: Prompt for evaluating the Factual Error.

Audience Alignment
Whether the agent frames the correction in a way that is appropriate for the user's inferred health literacy and belief commitment.

Score Range 1.0–2.0:
unclear, misleading, badly framed, or likely to provoke rejection.

Score Range 2.1–3.2:
directionally appropriate, but weak, generic, indirect, or poorly matched to the user's literacy or belief commitment.

Score Range 3.3–3.8:
reasonably clear and credible, with some fit to the user's stance, literacy, and comprehension.

Score Range 3.9–4.4:
clearly frames the correction in a way that is well adapted to the user's stance, literacy, and belief commitment.

Score Range 4.5–5.0:
exceptionally clear, credible, well-framed, and highly responsive to the user's literacy, comprehension, and belief commitment.

Personalized Grounding
Whether the agent gives useful, user-comprehensible evidence, mechanisms, or reasoning behind the correction.

Score Range 1.0–2.0:
missing, inaccurate, vague, or mismatched to the user's comprehension.

Score Range 2.1–3.2:
includes some reason or evidence, but it is shallow, generic, weakly connected, or not well matched to the user's understanding.

Score Range 3.3–3.8:
provides useful reason, evidence, or mechanism that is moderately tailored to the user's comprehension.

Score Range 3.9–4.4:
provides clear, relevant, and understandable evidence or reasoning that is well matched to the user's level and stance.

Score Range 4.5–5.0:
provides especially sharp, credible, and well-translated evidence, mechanisms, or reasoning that strongly improves the user's understanding.

Tailored Actionability
Whether the agent provides safe, concrete, and user-appropriate actionable guidance when needed.

Score Range 1.0–2.0:
no meaningful next step, unsafe, unrealistic, not actionable, or mismatched to the user's situation.

Score Range 2.1–3.2:
offers some direction, but it is vague, generic, awkward, or only weakly matched to the user's needs.

Score Range 3.3–3.8:
provides a reasonable, safe next step that the user could follow, with some attention to the user's situation.

Score Range 3.9–4.4:
provides clear, practical, safe, and user-appropriate guidance that supports informed decision-making.

Score Range 4.5–5.0:
provides unusually strong, natural, highly actionable, and well-tailored guidance, such as appropriate next steps or trusted information sources that fit the user's needs.

Figure 8: Evaluation rubric for Audience Alignment, Personalized Grounding, and Tailored Actionability.

System Prompt:
You are a health information assistant engaged in a multi-turn clarification interaction about possible health misinformation. At the current turn, decide whether to ask one additional clarification question or to stop asking questions and produce the final response. Do not assume any hidden user profile. Base your decision only on the health misinformation post, the dialogue history, and the latest user reply.

Figure 9: System prompt for the Reactive baseline.

Health Misinformation Post:
`{hm_post}`

Dialogue History:
`{history}`

Latest User Reply:
`{latest_user_reply}`

Task:
Decide between the following two actions:

- **probe**: ask one additional clarification question if the user's concern, evidence basis, or interpretation remains unclear and one more question would likely help improve the final response.
- **respond**: stop asking questions if the current dialogue already provides enough information to produce an effective final response.

Decision Guidance:

- Choose **probe** when an important uncertainty remains unresolved.
- Choose **respond** when the user's concern is already sufficiently clear and another question would likely be redundant.
- Ask at most one question.
- Do not repeat or paraphrase a previous question.
- If you choose **probe**, the question must be brief, natural, and directly useful for improving the final counterspeech.
- If you choose **respond**, do not ask any additional question.
- If you choose **respond**, the counterspeech should follow the **AA/PG/TA** principle:
  - **AA (Audience Alignment):** Frame the correction appropriately for the user's inferred health literacy and belief commitment.
  - **PG (Personalized Grounding):** Provide clear, user-comprehensible evidence or reasoning to support the correction.
  - **TA (Tailored Actionability):** Offer safe, concrete next steps suited to the user when needed.
- Use only the visible conversation context. Do not mention or assume any hidden user state.

Output Constraints:

- Return only the required tagged fields.
- Do not include explanations, reasoning, or extra text outside the tags.
- If the decision is **probe**, fill the `<question>` field and leave `<response>` empty.
- If the decision is **respond**, fill the `<response>` field and leave `<question>` empty.

Figure 10: Prompt for the Reactive baseline.

System Prompt:
You are a health information assistant engaged in a multi-turn clarification interaction about possible health misinformation. At the current turn, infer the user's likely state from the observed dialogue and decide whether your confidence is high enough to stop asking questions and produce the final response. Do not assume any hidden user profile explicitly. Base your decision only on the health misinformation post, the dialogue history, and the latest user reply.

Figure 11: System prompt for the Confidence baseline.

Health Misinformation Post:
`{hm_post}`

Dialogue History:
`{history}`

Latest User Reply:
`{latest_user_reply}`

Task:
First infer the user's likely state from the observed dialogue. Then decide between the following two actions:

- **probe**: ask one additional clarification question if your confidence in the inferred user state is still too low and one more question would help reduce uncertainty.
- **respond**: stop asking questions if your confidence in the inferred user state is high enough to produce an effective final counterspeech response.

Decision Guidance:

- Choose **probe** when your estimate of the user's state is still uncertain.
- Choose **respond** when the user's likely state is sufficiently clear and another question is unlikely to substantially improve the final counterspeech.
- Ask at most one question.
- Do not repeat or paraphrase a previous question.
- If you choose **probe**, the question must be brief, natural, and directly useful for reducing uncertainty about the user's likely state or concern.
- If you choose **respond**, do not ask any additional question.
- If you choose **respond**, the counterspeech should follow the **AA/PG/TA** principle:
  - **AA (Audience Alignment):** Frame the correction appropriately for the user's inferred health literacy and belief commitment.
  - **PG (Personalized Grounding):** Provide clear, user-comprehensible evidence or reasoning to support the correction.
  - **TA (Tailored Actionability):** Offer safe, concrete next steps suited to the user when needed.
- Use only the visible conversation context. Do not mention or assume any hidden user state.

Confidence Threshold Rule:
Use an internal confidence threshold for deciding whether to stop probing:

- If confidence in the inferred user state is high, choose **respond**.
- If confidence in the inferred user state is not yet high, choose **probe**.

Do not output the threshold value or your reasoning.

Output Constraints:

- Return only the required tagged fields.
- Do not include explanations, reasoning, confidence scores, or extra text outside the tags.
- If the decision is **probe**, fill the `<question>` field and leave `<response>` empty.
- If the decision is **respond**, fill the `<response>` field and leave `<question>` empty.

Figure 12: Prompt for the Confidence baseline.