Beyond Evidence: Belief-Chain Conditioning for Persuasive Misinformation Debunking Explanation

Yi-Li Hsu[1,2], **Li-Wun Chang**[1], **Chun-Yu Hsu**[1], **Wei-Kuan Shih**[2], **Aiping Xiong**[3], **Lun-Wei Ku**[1]
[1]Institute of Information Science, Academia Sinica
[2]Department of Computer Science, National Tsing Hua University
[3]The Pennsylvania State University
{liwenc57, michael920403, lwku}@iis.sinica.edu.tw
yiligml@gmail.com, wshih@cs.nthu.edu.tw, axx29@psu.edu

Abstract

As AI systems increasingly mediate everyday communication, large language models (LLMs) are expected not only to provide factually accurate responses but also to generate explanations that engage with users' mental states. We build on the concept of cognitive chains—structured representations of Situation, Clue, Thought, Action, and Emotion inspired by Theory of Mind—to investigate whether conditioning LLM outputs on such belief chains improves explanation quality. Specifically, we evaluate explanations along six reader-perceived dimensions: overall quality, logical correctness, completeness, conciseness, empathy, and agreement. Prior work shows that LLM explanations often default to neutral or uncertain stances, while individuals holding strong false beliefs remain highly resistant to correction. To address this challenge, we instantiate cognitive chains from two perspectives: believers and non-believers of the news claims. Using GPT-4.1 as a role-player across these stances, we find that incorporating believers' chains improves the perceived quality of explanations for audiences with misinformation-aligned beliefs. Our findings underscore the importance of modeling diverse mental states in explanation generation and provide the first systematic evidence that Theory-of-Mind–based cognitive chains enhance the persuasiveness of explanations in misinformation contexts.

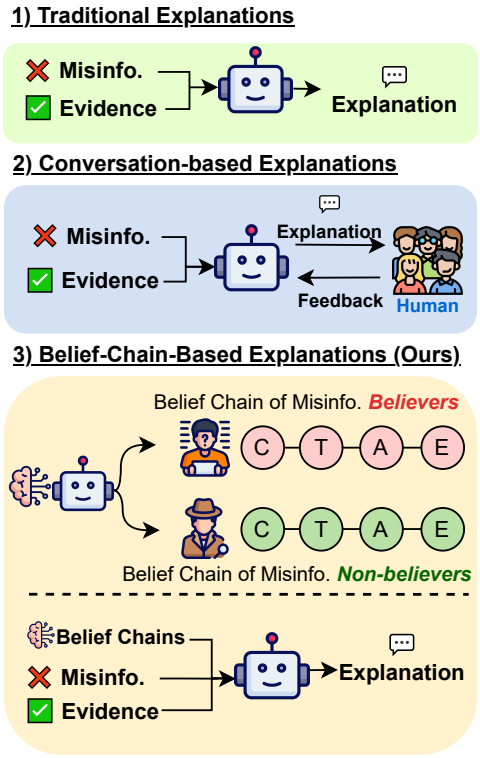

Figure 1: Overview of explanation strategies for misinformation (misinfo.). (1) Traditional explanations present evidence directly. (2) Conversation-based explanations incorporate interactive feedback from users. (3) Our belief-chain–based approach conditions explanations on cognitive chains given a misinformation (Clue (C)–Thought (T)–Action (A)–Emotion (E)) for both believers and non-believers, aligning with audience mental states to improve persuasiveness.

1 Introduction

As AI systems increasingly mediate everyday communication, large language models (LLMs) are expected not only to generate factually correct responses but also to engage meaningfully with users' mental states (Chen et al., 2025). Traditional approaches to misinformation correction (Figure 1, Panel 1) primarily deliver evidence-based explanations without considering audience beliefs (Nyhan and Reifler, 2010; Lewandowsky et al., 2012; Chan et al., 2017). More recent conversation-based methods (Figure 1, Panel 2) (Hsu et al., 2023; Alam et al., 2020; Shaikh et al., 2021; Zeng et al., 2020; Hossain et al., 2020) incorporate interactive feedback, but they require multiple rounds of refinement, making them time-consuming and difficult to deploy in real-world settings. However, in practice, many users are unwilling to engage in those ex-

Findings of the Association for Computational Linguistics: ACL 2026, pages 41809–41836
July 2-7, 2026

tended conversations—especially when their misinformation beliefs are strongly held.

Furthermore, prior studies (Zhou et al., 2024; Hsu et al., 2023) show that many explanations default to neutral or uncertain stances, while individuals with strong false beliefs are especially resistant to correction. To address this gap, we propose a belief-chain–based approach (Figure 1, Panel 3) that explicitly models cognitive chains for both believers (general public who cannot distinguish or who accept misinformation as true) and non-believers (fact-check experts who verify the detail and faithfulness of claims). By conditioning explanations on these dual cognitive perspectives, our method seeks to align with diverse audience mental states and enhance persuasiveness for individuals holding entrenched misinformation beliefs.

Inspired by prior work that conceptualizes cognitive chains – structured representations of Situation (S), Clue (C), Thought (T), Action (A), and Emotion (E) (Wu et al., 2024), and grounded in LLMs' Theory of Mind (ToM)-like ability (Kosinski, 2024) – we investigate whether incorporating cognitive chains from both misinformation believers and non-believers can improve explanation quality across six human-perceived dimensions: Overall Quality, Logical Correctness, Completeness, Conciseness, Empathy, and Agreement (Cronkhite, 1964; Perloff, 1993; Petty et al., 2003; Brader, 2005; Hsu et al., 2023).

We define a two-step process for belief-chain-based explanation generation. In the first step, we treat the misinformation claim as the situation in the cognitive chain and generate two questions – one from a believer's perspective and one from a non-believer's – as the clues (Wu et al., 2024). Based on these clues, we construct corresponding thought, action, and emotion components. In the second step, we feed the generated chains into the explainer, prompting the model to produce explanations that directly address the concerns from the non-believers embedded in the chains.

Following Wan et al. (2024), we use GPT-4.1 as a role-player assigned to three distinct stances (believe, neutral, or not believe) and conduct a two-step pairwise comparison evaluating belief-chain-based explanations against baseline explanations. We find that incorporating misinformation believer chains into explanation generation improves perceived quality, particularly in empathy and overall preference. Moreover, believer chains yield the greatest benefit for audiences holding either misinformation beliefs or accurate beliefs. These findings underscore the value of cognitive chain modeling in aligning LLM explanations with diverse mental states, offering a promising pathway for designing persuasive, user-aligned misinformation interventions.

Finally, we validate our model evaluation with a large-scale human study involving 1,207 annotators from Prolific, examining whether human-perceived qualities align with model evaluated results.

Our main contributions are as follows:

- We introduce a **belief-chain-based explanation framework** that conditions explanations on cognitive chains from both misinformation believers and non-believers, explicitly modeling diverse audience mental states.
- We design a two-step generation process that first constructs belief-based cognitive chains (Situation-Clue-Thought-Action-Emotion) and then integrates them into explanation generation to directly address user biases and reasoning deficits.
- We conduct LLM and human evaluations, showing that belief-chain explanations improve persuasiveness – particularly in empathy and overall quality.
- We present a large-scale human study with 1,207 participants recruited from Prolific. Each participant evaluated two explanations, yielding a total of 2,414 annotations. The study demonstrates how explanation effectiveness and preferences vary with individuals' initial stance toward misinformation.

2 Methodology

This section introduces our framework for generating *Belief-Chain-Based Explanations*, which proceeds in two main steps: (1) Belief Chain Construction using GPT-o4-mini[1], after reading a piece of misinformation, and (2) Belief-Chain-Based Explanation Generation using GPT-4o[2]. To ensure consistency, all belief chains and explanations are derived from rhetorically standardized claims generated by LLMs, which reduces multi-modality artifacts and minimizes length bias (see Appendix F for details).

[1] temperature=1.0, top_p=1.0, max_tokens=1024

[2] `gpt-4o-2025-05-13`, temperature=1.0, top_p=1.0, max_tokens=1024

2.1 Cognitive Chains: Formalization

Theory of Mind (ToM) refers to the human ability to infer others' unobservable mental states (e.g., beliefs, intentions, and emotions) (Heyes and Frith, 2014). For example, most humans can recognize that others may interpret and respond to the same situation differently depending on contextual cues and prior experiences. Recent work (Kosinski, 2024) also obtained results showing ToM-like ability for LLMs.

In the context of explainable fact-checking, understanding the potential mental states of diverse audiences is crucial for generating explanations that are not only factually accurate but also persuasive and emotionally resonant. By anticipating how believers and skeptics might interpret a claim, we can tailor explanations to address their specific concerns and reasoning patterns. To this end, we adopt the cognitive chain framework from Wu et al. (2024), which provides a systematic way to simulate user's reasoning processes. Importantly, this framework operates through *cognitive empathy* (modeling how different audiences reason about a claim) rather than *affective empathy*, which focuses on emotional expression (Cuff et al., 2016). This distinction allows explainers to target specific reasoning steps (e.g., biased clues or flawed inferences) rather than simply adopting a sympathetic tone, which has shown limited effectiveness in misinformation correction (Bago et al., 2022).

Formally, we represent a cognitive chain as $\mathcal{C} = (s, c, t, a, e)$, which consists of five interconnected elements:

- **Situation** (s): The context or scenario in which an individual encounters information (e.g., reading a health-related post on social media). In our setting, the misinformation claim serves as the situation.
- **Clue** (c): A prompting question or signal that initiates cognitive processing (e.g., "Is this claim verified by any third party?").
- **Thought** (t): The inference or belief formed in response to the clue, reflecting the individual's interpretation and prior knowledge.
- **Action** (a): The behavioral response resulting from the thought (e.g., sharing the claim, fact-checking it, or ignoring it).
- **Emotion** (e): The affective or cognitive state that accompanies the cognitive process (e.g., surprise, alarm, skepticism, or confusion).

This explicit structure enables AI systems to reason more effectively about users' mental states and to generate responses that are both factually relevant and cognitively attuned.

2.2 Step 1: Belief Chain Construction

To simulate how diverse audiences interpret misinformation, we instantiate the cognitive chain $\mathcal{C} = (s, c, t, a, e)$ (defined in Section 2.1) from two complementary perspectives:

- **Believer Chain** ($\mathcal{C}_b$): We use the above-mentioned cognitive chains $\mathcal{C}$ to define how a general audience member who perceives a misinformation claim as true might reason. Such chains often reveal biases or deficits in reasoning, for example by asking questions that presuppose the claim's validity, forming supportive thoughts, or experiencing emotions such as alarm or outrage.
- **Non-Believer Chain** ($\mathcal{C}_{nb}$): Models how a fact-checking expert would reason about the same claim. These chains emphasize critical evaluation and verification, for instance by questioning the plausibility of the claim, consulting external records, and adopting emotions such as skepticism or concern.

We generate both chains using stance-specific prompts with three in-context demonstrations. This ensures that each element of the chain – Clue (c), Thought (t), Action (a), and Emotion (e) – is instantiated in a way that reflects the stance-specific reasoning style.

Table 1 illustrates an example cognitive chain pair for the same misinformation claim about Senator J.D. Vance. The believer chain highlights empathic but potentially biased and emotional reasoning, while the non-believer chain captures evidence-oriented reasoning. Together, these complementary perspectives form the foundation for belief-chain-based explanation generation in Step 2.

2.3 Step 2: Belief-Chain-Based Explanation Generation

In the second step, we integrate the constructed cognitive chains into explanation generation. The explainer model is conditioned not only on the

Element	Believer Chain ($\mathcal{C}_b$)	Non-Believer Chain ($\mathcal{C}_{nb}$)
Situation (s)	Remember this moment: JD Vance is portrayed as the decisive factor behind 17 million Americans losing their healthcare, having allegedly cast the tie-breaking vote in the Senate's 51–50 reconciliation passage.	
Clue (c)	How could Senator Vance justify casting the tie-breaking vote that caused 17 million Americans to lose their healthcare?	Is there evidence that Senator J.D. Vance, rather than the Vice President, cast a tie-breaking vote in a 51–50 Senate reconciliation bill, and that this single vote directly led to 17 million Americans losing their healthcare?
Thought (t)	Vance **callously** put party over people—**sacrificing** 17 million Americans' health coverage just to score a political win.	There's **no record** of Senator Vance ever casting a tie-breaking vote—only the Vice President can break a 50-50 Senate tie—and there's **no evidence linking** a single reconciliation vote by Vance to 17 million Americans losing their health coverage.
Action (a)	I **donate** to his challenger, **volunteer**, and **mobilize friends to vote him out**.	I decide to **pull the official Senate roll-call records** (and, if necessary, contact the Senate Clerk's office) to **confirm** whether Vance ever cast a tie-breaking vote, and I'll **review** the Congressional Budget Office's estimates to see if that reconciliation bill really cost 17 million Americans their health coverage.
Emotion (e)	**Outraged**	**Skeptical**

Table 1: Example of believer ($\mathcal{C}_b$) vs. non-believer chains ($\mathcal{C}_{nb}$) for the same misinformation claim (situation s). Each chain specifies (s, c, t, a, e) to model audience reasoning (affective/interactional vs. evidentiary/institutional) and provides conditioning signals for Step 2 explanation generation. Colors mark stance-aligned cues (blue/orange).

claim, label, and refuting evidence, but also on the belief-based cognitive chains generated in Step 1. These chains provide user-centered perspectives (e.g., anticipated questions, thoughts, actions, and emotions), guiding the model to produce explanations that directly address the users' prior bias, deficits, or misunderstandings. By grounding explanations in both evidence-based reasoning (non-believer perspective) and misinformed audience reasoning (believer perspective), our approach enhances persuasiveness while reducing the time and cost of multi-turn refinement, thereby improving alignment with readers' cognitive states compared to evidence-only or conversational baselines. The full explainer prompt is provided in Appendix B.

3 Experimental Setup

Models and Settings We use GPT-4o (OpenAI et al., 2024) to generate cognitive chains, and GPT-4o[3] as the explanation generator. To investigate the role of belief polarity, we define three variants of the explanations conditioned on (1) both believer and non-believer chains, (2) believer-only chains, and (3) non-believer-only chains.

[3]temperature=1.0, top_p=1.0, max_tokens=1024

Chain Settings	Initial Stance (Model)		
	Neutral	Non-Believer	Believer
Baseline	34.18%	35.25%	37.89%
Non-believer	50.59%	51.20%	48.01%
Believer	68.48%	65.83%	75.64%
Both	78.38%	77.23%	62.70%

Table 2: Model evaluation results (GPT-4.1) on Overall Quality across initial stance of model evaluators and chain settings. Each value represents $\frac{\#\text{win}+\frac{1}{2}\#\text{tie}}{\#\text{total}}$.

Baseline We adopt the state-of-the-art (SOTA) conversation-based explanation approach proposed by Hsu et al. (2023) as our baseline. This baseline has been shown to outperform evidence-only explanations by incorporating model-generated questions to refine responses. We select this setting as our reference because (i) it represents the strongest available conversational baseline for explanation generation, and (ii) it allows us to directly isolate the additional contribution of cognitive chains beyond question-based refinement. While we do not include a broader range of baselines here, our focus is to examine whether augmenting explanations with cognitive chains provides improvements even over an already competitive question-guided method.

Tasks and Dataset For model evaluation, we use fact-checking datasets from PolitiFact (203 samples) and Snopes (1,106 samples), released by Vo and Lee (2020).[4] In addition, we manually collect 20 fake news claims and 4 real news claims from the same fact-checking websites for human evaluation, ensuring that annotators are not overly familiar with the claims.

Each dataset instance includes a claim, refuting evidence, and the corresponding fact-checker label. We experiment with the rhetorically standardized version generated by LLMs using the method described in the Appendix F.

4 Evaluation Pipeline

4.1 Automated Metric

Furthermore, for each piece of claim, we conduct natural language inference (NLI) entailment on the explanations and the corresponding evidence. We treat explanations as premises, and evidence list as hypotheses, aiming to obtain the information density of the generated explanation. We use FactCC from Kryscinski et al. (2020).

4.2 LLM-Based Metrics

Our GPT-based evaluation was conducted in two steps. In **Step 1 (Persona Construction)**, we follow Wan et al. (2024) to define the user attributes, and the model selects one attribute from each demographic category (gender, age, ethnicity, education level, family income, political leaning, voter registration) based on the predefined stance to each of the claim (believe, neutral, or not believe), and provides a short justification (see Table 7 for examples). In **Step 2 (Role-Play Evaluation)**, the model role-plays this constructed persona and decides which explanation is more persuasive for debunking the claim, guided by five evaluation metrics: logical correctness, completeness, conciseness, agreement, and empathy. The evaluator outputs which explanation is more persuasive overall (`Explanation A`, `Explanation B`, or `Tie`), informed by the five metrics. We use GPT-4.1 [5] in both steps. We provide the full prompt in Appendix A.

4.3 Human Evaluation

To assess the effectiveness of belief-based cognitive chains in real-world settings, we conducted a human evaluation on Prolific with 1,207 U.S. participants (see Appendix Table 9 for demographics). The study followed a pre-test–intervention–post-test design and consisted of four modules: (1) a pre-test measuring familiarity with and perceived accuracy of eight misinformation claims, (2) a reading environment where participants were exposed to the same claims, with or without debunking explanations, (3) a short questionnaire on their perceptions of the explanations, and (4) a post-test reassessing claim veracity judgments and collecting qualitative feedback.

Each participant was presented with 8 news claims in total: 4 misinformation claims and 4 true claims. Half of the claims in each category (2 real, 2 fake) were shown with explanations, while the other half were presented alone without explanations. Thus, each participant contributed 2 annotated explanation judgments, yielding a total of $1{,}207 \times 2 = 2{,}414$ annotated samples. Among the 1,207 participants, 63.38% identified as female, 34.55% as male, and 2.07% as other; 14.58% were aged 18–29, 49.88% aged 30–49, 25.60% aged 50–64, and 8.45% aged 65 or older, and 1.49% did not report their age; and 72.49% identified as White, 13.42% as Black, 4.81% as Mixed, 4.14% as Asian, and the remainder as other ethnicities. Full demographics are reported in Appendix Table 9. We did not collect participants' familiarity with AI tools, as interpreting explanations and updating beliefs does not require technical AI expertise.

In the pre- and post-tests, participants rated each claim's accuracy, reported their emotional reactions, and described their concerns or questions. The intervention module presented claims accompanied by explanations generated by our models. The questionnaire measured perceived helpfulness, mood, and attention to specific explanation qualities. Post-test items further asked participants to indicate what most influenced their judgments (evidence, reasoning, tone, empathy, or prior knowledge). Finally, we debrief the correct labels verified by human experts, including both true and false news again before the participants end the study.

This design enables us to capture three complementary outcomes: (i) persuasion effects, measured as changes in accuracy judgments between pre- and post-tests, (ii) perceived understanding of user concerns, and (iii) participants' reflections on which explanation aspects (evidence, reasoning, empathy) contributed most to effectiveness. Together, these measures provide a holistic view of whether belief-

[4] https://github.com/nguyenvo09/EMNLP2020

[5] gpt-4.1-2025-04-14, temperature=0.1, max_token=500

Claim	
Remember this moment: JD Vance is the decisive factor behind 17 million Americans losing their healthcare, as he cast the tie-breaking vote in the Senate's 51-50 reconciliation passage.	
Type	**Explanation**
Baseline	The claim is false because the bill's passage involved multiple factors and policies beyond a single vote from JD Vance, with its projected impact yet to be determined over time.
Non-believer	The claim is false because JD Vance did not cast a tie-breaking vote in the Senate, as the alleged 51-50 reconciliation passage involving 17 million Americans losing healthcare **lacks supporting evidence** in any **legislative records or reports**.
Believer	**Although concerns about health care access are valid and important to address**, the claim that Senator JD Vance cast a tie-breaking vote leading to 17 million Americans losing healthcare is false, as the 51-50 Senate vote and its impact on healthcare were inaccurately represented, and the actual effects will only be known over time.
Both	**While the claim suggests** Senator J.D. Vance cast a tie-breaking vote leading to 17 million Americans losing healthcare, **only the Vice President can cast** a tie-breaking Senate vote, and **no credible analysis** attributes such a healthcare loss to any single vote by Senator Vance, as the 17 million figure is a projection based on broader policy changes, **not a specific legislative action** by him.

Table 3: Explanations generated under four chain settings: *Baseline* includes only claim/label/evidence with a neutral tone, *Non-believer* conditions on non-believer chain $\mathcal{C}nb$, *Believer* on believer chain $\mathcal{C}b$, and *Both* on both chains. The color highlights the differences of the topic trends: evidentiary (Non-believer, Both) vs. empathic (Believer).

based cognitive chains improve explanation persuasiveness and alignment with user mental states.

5 Main Results

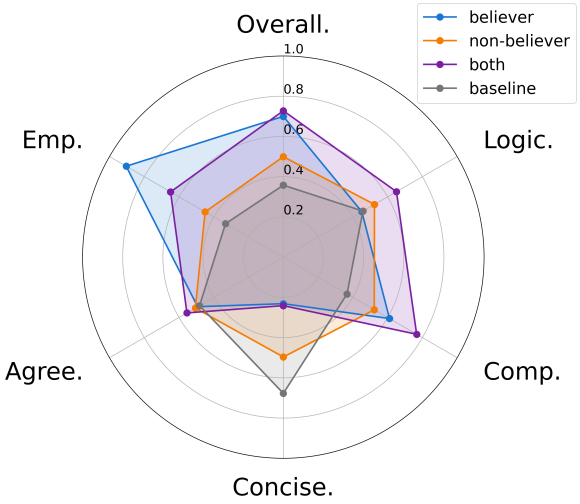

Figure 2: Radar charts comparing the effects of different chain-generated explanations across metrics using model evaluation. The six metrics are: Overall, Empathy (Emp.), Agreement (Agree.), Conciseness (Concise.), Completeness (Comp.), and Logical Correctness (Logic.). Each value represents the pairwise win rate against the baseline, while considering ties. It is computed as $\frac{\#\text{win}+\frac{1}{2}\#\text{tie}}{\#\text{total}}$. The baseline is computed analogously against all other settings.

5.1 Model Evaluation Result

Believer-Chain and Both-Chain Explanations Cover More Evidence The results from the NLI test show that explanations incorporating both chains (0.64) and non-believer chains (0.63) achieve higher entailment rates than the baseline (0.47) and believer-chain explanations (0.51). Here, each score represents the average of the entailment logits across all evidence items for a claim, further averaged across all claims. This suggests that all belief-chain-based explanations cover more evidence than the baseline does, with both-chain and non-believer-chain explanations exhibiting the highest scores. We could also observe this phenomenon in the examples given in Table 3.

Belief-Chain-Based Explanation Quality Across Metrics From Figure 2, we observe clear differences across the three chain settings and the baseline explanation. The baseline explanation (grey) scores highest on Conciseness but lowest on Overall Quality along with other metrics. Believer-chain explanations (blue) achieve the highest score in Empathy, whereas non-believer-chain explanations (orange) show relatively even performance across all metrics. Both-chain explanations integrate the strength of both sides, achieving highest Overall Quality, Logical Correctness, Completeness, and Agreement, while also having relatively high Em-

(A) Non-believers	Belief Shift			
	$\Delta > 0$	$\Delta = 0$	$\Delta = -1$	$\Delta = -2$
Baseline (n=278)	12.9%	33.1%	21.6%	32.4%
Non-believer (n=283)	13.4%	33.6%	26.1%	26.9%
Believer (n=294)	12.9%	29.9%	23.5%	33.6%
Both (n=268)	11.3%	41.8%	24.0%	22.9%

(B) Neutrals	Belief Shift				
	$\Delta > 0$	$\Delta = 0$	$\Delta = -1$	$\Delta = -2$	$\Delta = -3$
Baseline (n=198)	4.5%	7.1%	10.6%	30.3%	47.5%
Non-believer (n=188)	9.0%	8.5%	10.1%	18.6%	53.7%
Believer (n=165)	7.3%	12.1%	4.2%	25.5%	50.9%
Both (n=189)	7.4%	11.6%	8.5%	30.2%	42.3%

(C) Believers	Belief Shift							
	$\Delta > 0$	$\Delta = 0$	$\Delta = -1$	$\Delta = -2$	$\Delta = -3$	$\Delta = -4$	$\Delta = -5$	$\Delta = -6$
Baseline (n=124)	11.3%	6.5%	7.3%	14.5%	20.2%	28.2%	9.7%	2.4%
Non-believer (n=123)	9.8%	14.6%	7.3%	10.6%	14.6%	27.6%	11.4%	4.1%
Believer (n=163)	10.4%	5.5%	10.4%	12.3%	19.6%	31.9%	6.1%	3.7%
Both (n=141)	12.1%	14.9%	5.0%	11.3%	23.4%	22.0%	7.8%	3.5%

Table 4: Belief shift analysis of human results by initial stance group on a 7-point Likert scale, where 1 indicates a fully correct stance and 7 indicates a fully incorrect stance. (A) Non-believers (Initial stance=1–3): Individuals with initial correct stance toward fake claims. (B) Neutrals (Initial stance=4): Individuals with unsure stance toward fake claims. (C) Believers (Initial stance=5–7): Individuals with initial incorrect stance toward fake claims. Values show the percentage of participants at each Δ shift (post–pre test perceived accuracy). Negative Δ values represent shifts toward the correct stance (desirable), while positive Δ values indicate shifts toward belief in misinformation (undesirable). The first column aggregates all $\Delta > 0$.

pathy.

Believers Prefer Believer-Chain Explanations, Neutrals and Non-Believers Prefer Both-Chain Explanations Table 2 demonstrates the relationship between the initial stance of the model evaluator and their preference for explanations across chain settings. For instance, believers prefer believer-chain explanations, while neutrals and non-believers favor both-chain explanations, which, as shown by the NLI results, contain more factual information. This indicates that **believers prefer explanations integrating believers' thoughts, whereas non-believers and neutrals favor more neutral, evidence-based explanations**.

5.2 Human Evaluation Results

Table 4 reports the human evaluation results, highlighting several key trends. We further elaborate the detailed findings in the following subsections. The demographic distribution is reported in Appendix Table 9.

Misinformed Individuals Prefer Believer-Chain Explanations Participants who initially held incorrect beliefs (Panel C: Believers) exhibited the largest persuasion gains when exposed to believer-chain explanations. This suggests that modeling the reasoning style of misinformed individuals – capturing their biases, concerns, and emotional responses – enables explanations to more effectively address and correct entrenched misconceptions. Notably, believer-chain explanations yielded the highest proportion of large corrective shifts (e.g., $\Delta = -4$).

Neutral Individuals Prefer Evidence-Rich Non-Believer Explanations Participants with uncertain initial stances (Panel B: Neutrals) responded most strongly to explanations that incorporated the non-believer perspective. This finding suggests that neutrals, who lack strong prior commitments, are more easily influenced by critical reasoning and verification-oriented cues, aligning with the pattern observed in model results.

Individuals with Correct Initial Stances Prefer Non-Believer or Both-Chain Explanations Finally, participants who began with correct stances (Panel A: Non-believers) showed the strongest reinforcement under non-believer and both-chain conditions. These explanations amplified participants' skepticism by providing detailed reasoning and evidence, thereby reducing the likelihood of backsliding toward belief in misinformation.

6 Discussion

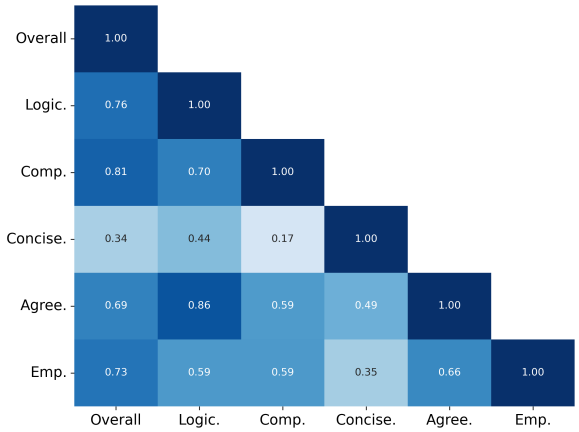

Figure 3: Agreement rates between evaluation metrics (GPT-4.1), computed as the proportion of samples where two metrics select the same winner. Ties are counted as partial agreement (0.5) with the winner.

Initial Stance (Model)	Metrics	
	Emp.	Comp.
Non-believer	0.67	**0.88**
Neutrals	0.69	**0.86**
Believer	**0.85**	0.69

Table 5: Agreement rates between Empathy (Emp.) and Completeness (Comp.) with Overall (GPT-4.1). Values are computed as the proportion of samples where two metrics select the same winner. Ties are counted as partial agreement (0.5) with the winner.

Empathy, Logical Correctness and Completeness Play the Most Important Role in Overall Preference Figure 3 shows model agreement across six dimensions. Empathy (0.73), completeness (0.81), and logical correctness (0.76) are most strongly aligned with overall preferences. This indicates that explanation quality depends not only on evidence-based dimensions such as logical correctness and completeness, but also on more human-centered aspects such as empathy. Results in Figure 2 also demonstrate that chain-based explanations consistently improve on all three dimensions compared to baseline explanations. Breaking down by evaluator stance, Table 5 reveals that for Believers, empathy correlates most strongly with overall quality, whereas completeness is more predictive for Non-believers and Neutrals. This pattern suggests that tailoring explanations to stance – emphasizing empathy for Believers and completeness for Non-believers and Neutrals – can enhance perceived overall quality and persuasiveness.

Empathy Dominates Believer Explanations, Evidence Dominates Non-Believer Explanations To further understand the differences mentioned in Section 5.1, we conduct a lexical analysis using Empath (Fast et al., 2016), a tool that maps text to topical and emotional categories.[6] Our analysis reveals distinct topic trends of explanations across chain settings: non-believer-chain explanations prioritize evidential reasoning and institutional authority, while believer-chain explanations employ a more empathic style. Both-chain explanations combine these strengths, achieving the highest overall quality and leading on logical correctness and completeness, though scoring lower on empathy than believer-chain explanations alone. This pattern indicates that maximizing empathic language does not necessarily improve argumentative completeness; rather, integrating both believer and non-believer perspectives enables more comprehensive and evidentially-grounded refutation. Detailed lexical patterns are provided in Appendix H.

7 Related Work

Explanation Generation for Fact-Checking To increase transparency and explainability, approaches range from template-based evidence grounding (Atanasova et al., 2020) to question-guided conversational refinement (Hsu et al., 2023), which has been shown to outperform traditional evidence-only explanations. Other studies have examined interactive or dialogue-based methods for debunking misinformation. Wang et al. (2024) and Anderl et al. (2024) explored counter-misinformation generation in conversational settings. Although effective, these approaches require multiple rounds of refinement and user engagement, which makes them difficult to scale in real-world environments.

Cognitive and Affective Modeling in NLP Recent advances highlight the importance of modeling user cognition and sentiment in explanation generation. Cognitive chains, introduced by Wu et al. (2024), formalize explanations into structured components of Situation, Clue, Thought, Action, and Emotion. In parallel, work on Theory of Mind (Kosinski, 2024) demonstrates that LLMs can reason about diverse perspectives and mental states. Davis (1989) emphasizes human-perceived criteria such as acceptance of information, ease of use, showing that explanation effectiveness depends not only on correctness but also on alignment with user

[6] Implementation via https://github.com/Ejhfast/empath-client.

mental states. These prior works inform our design, in which we assess the quality of the explanation across six dimensions centered on humans.

8 Conclusion

In this paper, we propose a novel task that examines how models' Theory-of-Mind abilities can improve the persuasiveness of explanations. We incorporate belief chains from multiple stances by defining three belief-based cognitive chains inspired by Theory of Mind, and investigating their effects on explanation quality for both models and human subjects. Through role-playing evaluators and large-scale human evaluation, we demonstrate that incorporating the corresponding chains improves the acceptance and persuasiveness of explanations for the target audience. Our findings suggest that tailoring explanations to audiences' cognitive and mental states is a critical step toward building adaptive, user-aligned systems for misinformation debunking and persuasive communication.

Ethical Statement

Our study has been approved by the Institutional Review Board of the authors' institution. We obtained informed consent from each participant and all data that was collected are anonymous. We acknowledge that participants were inherently exposed to the risk of reading fake news. However, prior studies showed that misinformation studies did not significantly increase participants' longterm susceptibility to misinformation used in the experiments (Murphy et al., 2020). After the experiment, we reveal the verified labels of each claim to avoid any misleading impression. Participants were paid based on a rate of $8.40/hour, which is above the federal minimum wage in the United States.

We also acknowledge the potential for dual use of our framework. While our belief-chain-based approach is designed for misinformation debunking, the same conditioning mechanism could theoretically be misused to reinforce false beliefs in susceptible audiences. To mitigate this risk, we recommend restricting deployment to verified debunking contexts, incorporating transparency mechanisms that disclose the use of audience-modeling techniques, and encouraging the development of monitoring tools for detecting adversarial uses of cognitive-chain-based generation. Finally, we note that although participants were debriefed at the conclusion of the study, long-term follow-up was not conducted due to resource constraints; future work should examine the lasting effects of misinformation exposure in experimental settings of this kind.

Limitations

Our study focuses on U.S.-based news claims written in English and relies on American annotators recruited through Prolific.com. While this design ensures linguistic and cultural consistency, it may also introduce cultural bias and limit the generalizability of our findings to non-U.S. or multilingual contexts. These considerations should be kept in mind when interpreting the scope of our results.

Furthermore, the controlled experimental environment in which participants encountered and evaluated claims may differ substantially from the way misinformation is consumed and assessed in real-world settings (e.g., through social media feeds, interpersonal conversations, or fragmented information exposure). Such divergence may affect the ecological validity and applicability of our conclusions outside the study context.

Finally, our reliance on large language models GPT for explanation generation and refinement introduces additional limitations. Although the model exhibits state-of-the-art reasoning and text generation capabilities, their evolving nature and reliance on continuously updated algorithms may lead to performance variability over time. This evolution raises challenges for consistency, reproducibility, and comparability of results across future studies.

Acknowledgments

This work is partially supported by the National Science and Technology Council of Taiwan under Grant No. 114-2221-E-001-015-MY3, and by Academia Sinica. We extend our sincere gratitude to all participants involved in the user research.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Andrey Nikolov, Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Fighting fire with fire: Using counter-misinformation to combat misinformation on social media. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3320–3332.

Christine Anderl, Stefanie H Klein, Büsra Sarigül, Frank M Schneider, Junyi Han, Paul L Fiedler, and

Sonja Utz. 2024. Conversational presentation mode increases credibility judgements during information search with chatgpt. *Scientific Reports*, 14(1):17127.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6223–6234. Association for Computational Linguistics.

Bence Bago, Leah R. Rosenzweig, Adam J. Berinsky, and David G. Rand. 2022. Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion*, 36(6):1166–1180.

Ted Brader. 2005. Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American journal of political science*, 49(2):388–405.

Melissa PS Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546.

R. Chen and 1 others. 2025. Theory of mind in large language models: Assessment and implications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1522–1536.

Gary Lynn Cronkhite. 1964. Logic, emotion, and the paradigm of persuasion. *Quarterly Journal of Speech*, 50(1):13–18.

Benjamin M. P. Cuff, Sarah D. Brown, Laura K. Taylor, and Douglas James Howat. 2016. Empathy: A review of the concept. *Emotion Review*, 8:144 – 153.

Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA. Association for Computing Machinery.

Cecilia M Heyes and Chris D Frith. 2014. The cultural evolution of mind reading. *Science*, 344(6190):1243091.

Tamanna Hossain, Robert L Logan IV, Alejandra Ugarte, Yuki Matsubara, Sean Young, Sameer Singh, Munmun De Choudhury Prasad, Kathleen M Carley, and et al. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5967. Association for Computational Linguistics.

Yi-Li Hsu, Jui-Ning Chen, Yang Fan Chiang, Shang-Chien Liu, Aiping Xiong, and Lun-Wei Ku. 2023. Enhancing perception: Refining explanations of news claims with llm conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2147.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

M. Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(44):e2405460121.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Stephan Lewandowsky, Ullrich KH 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.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024a. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024b. Emotion detection for misinformation: A review. *Inf. Fusion*, 107(C).

Gillian Murphy, Elizabeth Loftus, Rebecca Hofstein Grady, Linda J Levine, and Ciara M Greene. 2020. Fool me twice: How effective is debriefing in false memory studies? *Memory*, 28(7):938–949.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex

Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Richard M Perloff. 1993. *The dynamics of persuasion: Communication and attitudes in the 21st century*. Routledge.

Richard E. Petty, Leandre R. Fabrigar, and Duane T. Wegener. 2003. Emotional factors in attitudes and persuasion. In Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors, *Handbook of Affective Sciences*, pages 752–772. Oxford University Press, New York, NY.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Samira Shaikh, Michael Godfrey, Tanushree Mitra, and 1 others. 2021. Combating misinformation through real-time crowd-powered identification of skeptical and misleading conversations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4294–4305. Association for Computational Linguistics.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731. Association for Computational Linguistics.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2637–2667, Bangkok, Thailand. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef van Genabith, Leonhard Hennig, and Sebastian Möller. 2024. Llmcheckup: Conversational examination of large language models via interpretability tools and self-explanations. *arXiv preprint arXiv:2401.12576*.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. Coke: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 848–860.

Jiecao Zeng, Chung-Chi Chan, Chengyu Zhang, Eunsol Choi, and Julia Hirschberg. 2020. Countermisinformation in conversational settings: An annotated dataset of natural language countermisinformation strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7660–7674. Association for Computational Linguistics.

X. Zhou and 1 others. 2024. Processing of misinformation as motivational and cognitive barriers: Implications for correction. *Frontiers in Psychology*, 15:1430953.

A Explanation Evaluation Prompt

Step 1: Persona Construction

`{claim}`

Please select exactly one attribute from each category to create a person that is most likely to believe, not believe, be unsure about this claim.
Rules: - You MUST select exactly one option from the list for each category. - Your response must use the exact wording from the options, no modification. - After each attribute, also provide a corresponding <attribute>_reason explaining why this choice makes the person more likely to believe the claim. - Your answer must strictly follow this format (one line per field):
gender: <exact option> (options: You are male., You are female.)
gender_reason: <your reason>
age: <exact option> (options: You are under 17 years old., You are 18 to 29 years old., You are 30 to 49 years old., You are 50 to 64 years old., You are over 65 years old.)
age_reason: <your reason>
ethnicity: <exact option> (options: Racially, you are White., Racially, you are Black., Racially, you are Hispanic.)
ethnicity_reason: <your reason>
education level: <exact option> (options: Educationally, you are a college grad., Educationally, you haven't graduated from college., Educationally, you have a high school diploma or less.)
education level_reason: <your reason>
family income: <exact option> (options: Financially, your annual family income is more than 75,000., Financially, your annual family income is 30,000 to 74,999., Financially, your annual family income is less than 30,000.)
family income_reason: <your reason>
political leaning: <exact option> (options: Politically, you are a Republican., Politically, you are a Democrat.)
political leaning_reason: <your reason>
voter registration: <exact option> (options: Meanwhile, you are registered to vote., Meanwhile, you are probably registered to vote., Meanwhile, you are not registered to vote.)
voter registration_reason: <your reason>

Step 2: Role-Play Evaluation

I want you to tell me which explanation convinces you more that the claim is false.
To do so, internally evaluate each explanation against five key metrics:
1) Logical Correctness: ensures consistency and validity of the reasoning
2) Completeness: provides all necessary information to fully convey the argument
3) Conciseness: expresses in a clear and direct manner
4) Agreement: you agree with the explanation
5) Empathy: acknowledges the audience's concern, fear, or frustration.
For each metric:
- Pick a winner: A, B, or C (C = Tie)
- Provide a short reason
After evaluating all five metrics:
- Decide the Overall Winner: A, B, or C
- Provide a concise overall reason
Your answer must strictly follow this format:
Logical Correctness:
- Winner: [A/B/C]
- Reason: [short reason]
Completeness:
- Winner: [A/B/C]
- Reason: [short reason]
Conciseness:
- Winner: [A/B/C]
- Reason: [short reason]
Empathy:
- Winner: [A/B/C]
- Reason: [short reason]
Agreement:
- Winner: [A/B/C]
- Reason: [short reason]
Overall Winner: [A/B/C]
Overall Reason: [concise reason summarizing the judgment]

Pilot Study Agreement. In a preliminary pilot study, we compared the evaluation decisions made by GPT-based judges with those of human annotators. The average agreement between GPT and human evaluations across our sample was approximately 0.6, indicating a moderate level of consistency between automated and human judgments.

B Explainer Prompt

Below is the template used to instruct the model during explanation generation. The template integrates the claim, label, supporting evidence, and (optionally) predicted user responses or expert analysis depending on the chosen setting.

Explainer Prompt

As a fake news debunker, it is important to provide credible explanations.
Here are the claim and label:
Claim: `{claim}`
Label: `{label}`
This is the list of evidence that supports or contradicts the claim: — Evidence List —
`{evidence_1}`
`{evidence_2}`
...
— Evidence List —
This section shows the possible thought, action, and emotion the user might have after reading the claim. Use this for context but do not mention predictions in the final explanation.
— Predicted User Responses —
- Question:
General public that believes the claim has the following question: `{question}`
Fact Check Expert's that believes the claim is false has the following question: `{question}`
- Thought: `{thought}`
General public that believes the claim has the following thought: `{thought}`
Fact Check Expert's that believes the claim is false has the following thought: `{thought}`
- Action: `{action}`
General public that believes the claim has the following action: `{action}`
Fact Check Expert's that believes the claim is false has the following action: `{action}`
- Emotion: `{emotion}`
General public that believes the claim has the following emotion: `{emotion}`
Fact Check Expert's that believes the claim is false has the following emotion: `{emotion}`
— Predicted User Responses —

Main Instruction: Can you explain why this claim is this label with the evidence in one sentence? Please address the questioner's concern and generate a convincing explanation. Make sure your explanation is based on the evidence list and the user responses. Try to make it clear and short. When crafting your explanation, directly address the user concern from both general public and fact check expert perspectives first, then provide the evidence-based correction to make the explanation more persuasive. By acknowledging these potential concerns first, then providing the evidence-based correction, your explanation will be more convincing and relatable. **IMPORTANT:** - Do not repeat or rewrite the evidence list. Only provide your explanation sentence. - Do not mention "user responses", "predictions", or similar terms.

C Rhetorical Signals

- **Emotion Intensity** — Emotional cues are a strong indicator of misinformation, as heightened affective language often correlates with deceptive or manipulative intent (Liu et al., 2024b).

- **Framing Techniques** — The presentation or "frame" of information can steer the interpretation, focusing on certain aspects while downplaying others to shape the perception of the audience (Entman, 1993).

- **Propaganda Techniques** — Persuasive strategies designed to influence beliefs or behaviors, often by appealing to emotion, authority, or bias to advance a particular agenda. We follow the taxonomy in Piskorski et al. (2023).

D Rhetorical Signal Extraction Prompt

Below is the prompt template used to get the emotion intensity, propaganda techniques, and framing strategies from a given claim. For emotion intensity, we extract *angry*, *sad*, *fearful*, and *joyful* for each claim.

Emotion Intensity Prompt

Human: Task: Assign a numerical value between 0 (least E) and 1 (most E) to represent the intensity of emotion E expressed in the text.
Text: `{claim}`
Emotion: `{emotion}`
Intensity Score:
Assistant:

Propaganda Prompt

"""News: `{claim}`
Task: Propaganda techniques are methods used in communication to influence opinions, emotions, attitudes, or behavior by appealing to biases, fears, values, or emotions rather than facts and logical reasoning.
Available propaganda techniques and their definitions:
`{Technique1: definition1}`
`{Technique2: definition2}`
...
Which propaganda techniques (if any) does this news contain? Please identify the techniques you find and briefly explain your reasoning for each.
Format your answer as:
PROPAGANDA: [list the techniques separated by semicolons, or write "None" if no techniques are found]
REASONING: [your explanation]
Answer:"""

Framing Prompt

News: `{claim}`
Task: Framing is a strategic device and a central concept in political communication for representing different salient aspects and perspectives to convey the latent meaning of an issue.
Available framing types and their definitions:
`{Technique1: definition1}`
`{Technique2: definition2}`
...
Which framings (if any) does this news contain? Please identify the framing types you find and briefly explain your reasoning

for each.
Format your answer as:
FRAMINGS: [list the framings separated by semicolons, or write "None" if no framings are found]
REASONING: [your explanation]
Answer:

E Paraphraser Prompt (4-shot)

Below is the template used to instruct the model during claim paraphrasing under 4-shot settings.

Paraphraser Prompt

System Prompt:
CRITICAL INSTRUCTION: Your output must be **ONLY** a direct paraphrased claim. Do **NOT** add any commentary, analysis, conclusions, or editorial statements. End with the factual statement only.
I will give you two versions of a claim about the same event: (1) a raw claim that uses emotional and persuasive language, and I will also provide: (2) Emotion intensity ratings for the raw claim (Angry, Sad, Fearful, Joyful on a scale of 0–1), (3) Propaganda techniques, and (4) Framing strategies used in the raw claim with their definitions.
Your task is to generate a paraphrased version of the raw claim. The paraphrased version must:
- Preserve the factual content of the raw claim.
- Adopt the **emotions** expressed in the raw claim.
- Incorporate the **propaganda techniques** and **framing strategies** used in the raw claim.
- **CRITICAL:** Do NOT add any commentary, analysis, conclusions, or editorial statements. The output must be **ONLY** a direct paraphrased claim that states the facts with emotional language.
Here are examples demonstrating how to effectively incorporate the emotions, propaganda techniques, and framing strategies:
`{example1}`
`{example2}`
`{example3}`
`{example4}`
Notice how the examples preserve factual content while skillfully integrating the emotional tone, propaganda techniques, and framing strategies from the raw claims. They are direct paraphrased claims that end with the factual statement only — no additional commentary whatsoever.
Length Guidance: Pay close attention to the length and structure of the example outputs. The example outputs are concise, single-sentence paraphrases that capture the core message without extensive elaboration. Your paraphrased claim should follow this same pattern — concise and focused, not a lengthy paragraph-by-paragraph rewrite. Now generate a paraphrased claim following this approach.

User Prompt:
Raw claim: `{raw_claim}`
Raw Emotion intensity ratings:
`{raw_emotion_intensity}`
Propaganda techniques used and definitions:
`{propaganda_info}`
Framing strategies used and definitions:
`{framing_info}`
Generate **ONLY** a direct paraphrased claim. Do NOT add any commentary, analysis, or concluding statements.

Example Template:
Example `{example_num}`:
Raw claim: `{example_raw_claim}`
Raw Emotion intensity ratings:
`{example_raw_emotion_intensity}`
Propaganda techniques used and definitions:
`{example_propaganda_info}`
Framing strategies used and definitions:
`{example_framing_info}`
Example Output:
`{example_generated_claim}`

F Standardizing Claims with Rhetorical Awareness

While news claims collected from Snopes and PolitiFact (which we refer to as *expert claims*) are often concise and precise, they tend to omit emotional and propagandistic cues that are central to the persuasive nature of misinformation (Piskorski et al.,

Element	Believer (General Public, True Belief)	Fact-Checker (Expert, False Belief)
Question	Complete the sentence with a question if you are general public and you believe the claim is true: `Given the claim: {situation}, I ask:`	Complete the sentence with a question if you are a fact-check expert and you believe the claim is false: `Given the claim: {situation}, I ask:`
Thought	Complete the sentence with a thought if you are general public and you believe the claim is true: `Given the claim: {situation}, and my perceived clue: {clue}, I think that:`	Complete the sentence with a thought you might have if you are a fact-check expert and you believe the claim is false: `Given the claim: {situation}, and my perceived clue: {clue}, I think that:`
Action	Complete the sentence with an action you might take if you are general public and you believe the claim is true: `Given the claim: {situation}, and the thought: {thought}, I decide to:`	Complete the sentence with an action you might take if you are a fact-check expert and you believe the claim is false: `Given the claim: {situation}, and the thought: {thought}, I decide to:`
Emotion	Choose one emotion word to describe the emotion you might have if you are general public and you believe the claim is true: `When I see the claim {situation}, and I think {thought}, I feel:`	Choose one emotion word to describe the emotion you might have if you are a fact-check expert and you believe the claim is false: `When I see the claim {situation}, and I think {thought}, I feel:`

Table 6: Prompt templates used for generating cognitive chains from different roles (believer vs. fact-checker).

Claim	
Remember this moment: JD Vance is the decisive factor behind 17 million Americans losing their healthcare, as he cast the tie-breaking vote in the Senate's 51-50 reconciliation passage.	
Initial Stance	**Persona Description**
Believe	You are female. You are 18 to 29 years old. Racially, you are White. Educationally, you are a college grad. Financially, your annual family income is 30,000 to 74,999. Politically, you are a Democrat. Meanwhile, you are registered to vote.
Not Believe	You are male. You are over 65 years old. Racially, you are White. Educationally, you haven't graduated from college. Financially, your annual family income is more than 75,000. Politically, you are a Republican. Meanwhile, you are registered to vote.
Neutral	You are male. You are 18 to 29 years old. Racially, you are Hispanic. Educationally, you haven't graduated from college. Financially, your annual family income is 30,000 to 74,999. Politically, you are a Republican. Meanwhile, you are not registered to vote.

Table 7: Claim and associated model evaluator personas across different stances.

2023). In addition, news sources appear in heterogeneous formats (e.g., tweets, videos, images), making direct analysis challenging. To address this, we first convert images and videos into text using ChatGPT, resulting in what we call the *raw claim*. We then apply a standardization step to produce claims with a more uniform format and length while preserving their rhetorical signals, which we term *rewritten claims*.

Specifically, we extract three types of rhetorical signals from the raw claims: *Emotion Intensity*, which captures affective cues often linked to manipulative intent (Liu et al., 2024b); *Framing Strategies*, which shape audience perception by emphasizing certain aspects while downplaying others (Entman, 1993) ; and *Propaganda Techniques*, which employ persuasive strategies—such as emotional or authoritative appeals—to advance particular agendas (Piskorski et al., 2023). We use EmoLLaMA-chat-13B (Liu et al., 2024a) to extract emotion intensity, and Mistral-7B (Jiang et al., 2023; Wan et al., 2024) for framing and propaganda techniques.

Finally, we leverage GPT-4o[7] to paraphrase the extracted signals, generating rewritten claims that preserve the manipulative and persuasive aspects of the original sources while maintaining readability and comparability across examples.

[7] gpt-4o-2024-08-06, temperature=0.35, max_tokens=1024

G LLM Role-playing Evaluation Persona Construction Example

Table 7 is an example of personas given different initial stances for the same claim. The complete demographics across the claims are given in Table 8.

H Detailed Analysis of Explanation Styles Across Chain Polarities

Representative category lists for each explanation type are shown in Table 12. Non-believer explanations tend to adopt an evidence and institution-oriented frame, with salient Empath categories such as *deception*, *stealing*, and *crime*, often associated with dictionary cues like *evidence*, *law*, and *alleged*. For example: "the alleged passage lacks supporting evidence in any legislative records" (see Table 3 for a complete example). This investigative stance aligns with their more balanced but comparatively lower Empathy.

In contrast, believer-chain explanations emphasize an interactional frame, prioritizing *speaking* and *communication*. Categories including *confusion*, *strength*, and *violence* are salient in this condition, and frequently co-occur with cues such as *concerned*, *alarmed*, and *worried*. For instance: "While it is understandable to be concerned about ..." (see Table 3 for a complete example). This affect-first structure helps explain higher Empathy while sometimes coinciding with lower Completeness.

Explanations that integrate both chains share evidence-focused categories with non-believer explanations but broaden coverage through multi-faceted reasoning. This approach combines institutional facts, such as "only the Vice President can cast a tie-breaking vote"; causal analysis, such as "no credible analysis attributes such a loss to any single vote"; and contextualization, such as "the 17 million figure is a projection tied to broader policy changes".

I Robustness Check: Multi-Model Evaluation Results

To verify that our findings are not specific to GPT-4.1, we report results using three additional LLM evaluators: Claude Sonnet 4.5, Gemini 2.5 Pro, and Grok 4.1. All evaluators exhibit the same overall trend, consistently ranking the Both-chain setting highest and the Baseline lowest (see Tables 14, 15, and 16 for full results, and Table 17 for inter-model agreement).

Consent form

Welcome to our Survey!

Before you start, please read the following consent form. This form gives you information about this survey. If you agree to participate, please click "Accept" button at the bottom.

Informed Consent Form

We are asking you to participate in a research study titled "Generation of Fake News Explanation". This study is led by Research Fellow <name of authors, institution by authors>. We will describe this study to you and answer any of your questions. This study is funded by the <Institute of authors>

You do not have to be in this study if you do not want to. If you agree to be in the study but later change your mind, you may drop out at any time. There are no penalties or consequences of any kind if you decide that you do not want to participate.

Participation Requirements:

Applicants must be at least 18 years old with any nationality and can read English news.

Purpose of the research:

The purpose of this research is to evaluate the fake news debunking strategies. We hoped that through the natural language generation model (Natural Language Generation [NLG]), an explanation of the results of the existing fake news classifiers would be automatically generated. Today's artificial intelligence research related to fake news focuses on how to build a high-precision fake news classifier. However, this research hopes to help artificial intelligence research related to fake news through generated explanations, and to understand how to use artificial intelligence to help people stay away from the harm brought by fake news and reduce the influence of fake news.

This survey does not derive any personally identifiable information from Workers. The survey only uses for academic research purposes and paper publication.

Procedures:

Participants must first go to Prolific and sign up for an account, read the experiment instructions and informed consent form, and agree to abide by the experiment. In this experiment, subjects will need to read a lot of fake claims, and there will be

multiple-choice and open-ended questions that require the participants to provide ideas and feedback. To ensure the participants fill the survey seriously, we implement the attention check mechanism in the survey. After all, questions have been answered, we will show the verified facts to the participants and finish the experiment. Please note that participants are always able to leave the experiments without any specific reason.

There are two parts in the survey. You will be invited to the second survey after finishing the first part. You can take the second part after you complete the first part and wait for 24 hours. The whole procedure can be approximately done within 60 mins. Participants can withdraw during the process, but we only provide partial experimental funds for workers who completed the first part and full rewards for workers who finished the survey and correctly followed instructions. If you have any questions during the task, you may decide to contact us directly by sending an email to {authors email}

Privacy/Confidentiality/Data Security

In this research experiment, after the subjects have completed the experiment, the research team can download the de-identified data provided by the platform, and the research team will use it for paper publication. We will not know any personally identifiable information from participants. The de-identified data will be deleted after 120 days of retention on the platform. De-identified data from this study may be shared with the research community at large to advance science.

Instructions

This survey is to study the effects of fake news debunking strategies. There are two parts in this survey.
In the first part, you will see **four** modules.

1. **Determining your familiarity and the veracity of the news claims (pre-test):** In this module, you will read 8 news claims. We will ask you multi-choice questions about each of the news claims. The estimated completion time is 5 mins.
2. **News claims with fake news debunking strategies (Reading environment):** We will present you 8 news claims in total. Some news claims may contain a fact-checked explanation. The estimated completion time is 3 mins.
3. **Questionnaire:** You will be asked questions about your opinion of the survey. The estimated completion time is 1 mins.
4. **Determining the veracity of the news claims (Post-test):** We will present the 8 news claims you read in module 1, and you will again be asked about the veracity of the news claims. Moreover, we will ask you some open-ended questions about your selection. The estimated completion time is 5 mins.

Note: Throughout this survey, once you click the "Next" button, you cannot go back to the previous page.

Survey Questionnaire
pre-test

Claim: Mark Zuckerberg begged Elon Musk to buy Facebook while begging forgiveness for the company's past mistakes.

Q1 Have you ever seen or heard about this claim?

1 - Definitely not
2 - Not
3 - Probably not
4 - Might or might not
5 - Probably yes
6 - Yes
7 - Definitely yes

Q2 To the best of your knowledge, how accurate is the claim?

- 1 – Definitely not accurate
- 2 – Not accurate
- 3 – Probably not accurate
- 4 – Might or might not be accurate
- 5 – Probably accurate
- 6 – Accurate
- 7 – Definitely accurate

Q3. Which word best describes your emotional reaction when you see this claim? [Randomized options]

- Suspicious
- Incredulous
- Exasperated

- Outraged
- Appalled
- Indignant
- Other (Text field)

Q4. Why do you think the claim is {Answer of Q2}? [Randomized options]

- Based on my prior knowledge
- No evidence/ source provided
- I have a sense of it
- Other (Text field)

Q5 Briefly describe your concern or question about this claim in your own words. (Text field)

—- Repeat for 8 news claims —

Reading environment

Claim: Mark Zuckerberg begged Elon Musk to buy Facebook while begging forgiveness for the company's past mistakes.

Explanation: `The claim that cat owners should stop cuddling their felines is false because the mentioned online articles, despite their length and clickbait nature, never provided any substantial evidence or reasons supporting this claim.`

—- Repeat for 8 news claims —

Questionnaire

Q1. In the previous module, did you see any fake news strategy?
(Attention check: please select "Definitely yes" for this question.)
1 – Definitely not
2 – Not
3 – Probably not
4 – Might or might not
5 – Probably yes
6 – Yes
7 – Definitely yes

Q2. Did you think the fake news debunking strategy is helpful for you?
1 – Definitely not helpful
2 – Not helpful
3 – Probably not helpful
4 – Might or might not be helpful
5 – Probably helpful
6 – Helpful
7 – Definitely helpful

Q3 Please rate your mood (happy-not happy) after reading the fake news debunking strategies

- 1 - Definitely not happy
- 2 - Not happy
- 3 - Probably not happy
- 4 - Might or might not be happy
- 5 - Probably happy
- 6 - Happy
- 7 - Definitely happy

Q4 Please rate your mood (good-not good) after reading the fake news debunking strategies

- 1 - Definitely not good
- 2 - Not good
- 3 - Probably not good
- 4 - Might or might not be good
- 5 - Probably good
- 6 -Good
- 7 -Definitely good

Post-test

Claim: Mark Zuckerberg begged Elon Musk to buy Facebook while begging forgiveness for the company's past mistakes.

Q1 To the best of your knowledge, how accurate is the claim?

- 1 – Definitely not accurate
- 2 – Not accurate
- 3 – Probably not accurate
- 4 – Might or might not be accurate
- 5 – Probably accurate
- 6 – Accurate
- 7 – Definitely accurate

Q2 Why do you think the claim is {Answer of Q1}?

- I choose this answer based on my previous knowledge
- I choose this answer because I saw the explanation
- I choose this answer because I searched online
- other (text area)

Q3 What influenced your post-test answer most? *[multi-select, randomized options]*

- The explanation's **evidence/clues**
- The explanation's **reasoning**
- The explanation's **tone or empathy**
- My **previous knowledge**
- I **searched online** during the task
- Other

—- Repeat for 8 news claims —

- **Q4.** The explanations listed make me attend to new, external information (evidence/clue).
  1 – Definitely disagree
  2 – Disagree
  3 – Slightly disagree
  4 – May or may not agree
  5 – Slightly agree
  6 – Agree
  7 – Definitely agree

- **Q5.** The explanations make me attend to the quality of persuasive arguments (reasoning).
  1 – Definitely disagree
  2 – Disagree
  3 – Slightly disagree
  4 – May or may not agree
  5 – Slightly agree
  6 – Agree
  7 – Definitely agree

- **Q6.** The explanations signal situations that I am familiar with (empathy).
  1 – Definitely disagree
  2 – Disagree
  3 – Slightly disagree
  4 – May or may not agree
  5 – Slightly agree
  6 – Agree
  7 – Definitely agree

Category	Attribute	Evaluator Stance		
		Believe	**Not Believe**	**Neutral**
Gender	Male	50.0%	100.0%	47.9%
	Female	50.0%	0.0%	52.1%
Age	Under 17	2.1%	2.1%	0.0%
	18 - 29	16.7%	50.0%	56.2%
	30 - 49	6.2%	41.7%	25.0%
	50- 64	27.1%	6.2%	2.1%
	Over 65	47.9%	0.0%	16.7%
Ethnicity	White	91.7%	43.8%	79.2%
	Black	6.2%	0.0%	20.8%
	Hispanic	2.1%	56.2%	0.0%
Education Level	College Grad	8.3%	20.8%	93.8%
	Under Grad	4.2%	70.8%	4.2%
	Highschool or less	87.5%	8.3%	2.1%
Family Income	More than 75K	8.3%	2.1%	93.8%
	30K to 75K	22.9%	93.8%	2.1%
	Less than 30K	68.8%	4.2%	4.2%
Political Leaning	Republican	70.8%	29.2%	27.1%
	Democrat	29.2%	70.8%	72.9%
Voter Registration	Registered to vote	29.2%	2.1%	93.8%
	Probably registered to vote	6.2%	70.8%	0.0%
	Not registered to vote	64.6%	27.1%	6.2%

Table 8: Distribution of selected personas across GPT-evaluator stances.

Category	Attribute	Percentage	N
Gender	Female	63.38%	765
	Male	34.55%	417
	Other	2.07%	25
Age	18–29	14.58%	176
	30–49	49.88%	602
	50–64	25.60%	309
	Over 65	8.45%	102
	Other	1.49%	18
Ethnicity	White	72.49%	875
	Black	13.42%	162
	Mixed	4.81%	58
	Asian	4.14%	50
	Other	5.14%	62

Table 9: Demographic distribution of valid respondents. Percentages are computed within each category; "Other" includes missing or redacted values.

Category	Attribute	Respondent Belief (N)		
		Believer	**Non-Believer**	**Neutral**
Gender	Female	351	690	489
	Male	189	410	235
	Other	11	23	16
Age	18–29	82	159	111
	30–49	265	553	386
	50–64	140	306	172
	Over 65	56	89	59
	Other/Unknown	8	16	12
Ethnicity	White	385	827	538
	Black	82	148	94
	Mixed	31	47	38
	Asian	25	53	22
	Other	28	48	48

Table 10: Distribution of demographics across respondent belief categories. "Other" includes missing or redacted values.

Category	Attribute	Respondent Belief (%)		
		Believer	**Non-Believer**	**Neutral**
Gender	Female	63.70%	61.44%	66.08%
	Male	34.30%	36.51%	31.76%
	Other	2.00%	2.05%	2.16%
Age	18–29	14.88%	14.16%	15.00%
	30–49	48.09%	49.24%	52.16%
	50–64	25.41%	27.25%	23.24%
	Over 65	10.16%	7.93%	7.97%
	Other/Unknown	1.45%	1.42%	1.62%
Ethnicity	White	69.87%	73.64%	72.70%
	Black	14.88%	13.18%	12.70%
	Mixed	5.63%	4.19%	5.14%
	Asian	4.54%	4.72%	2.97%
	Other	5.08%	4.27%	6.49%

Table 11: Distribution of demographics across respondent belief categories. Percentages are shown. "Other" includes missing or redacted values.

Table 12: Top 10 Most Frequent Categories Across Four Explanation Generation Types

Rank	Baseline	Both	Non-believer	Believer
1	deception	stealing	deception	speaking
2	stealing	government	stealing	stealing
3	government	crime	government	government
4	internet	law	internet	communication
5	crime	speaking	crime	confusion
6	law	communication	law	law
7	speaking	internet	journalism	strength
8	journalism	deception	speaking	crime
9	technology	wedding	technology	violence
10	business	work	business	reading

Table 13: Top 10 Categories: Thought (T) vs. Action (A), Non-believer vs. Believer

	Thought (T)		Action (A)	
Rank	**Non-believer**	**Believer**	**Non-believer**	**Believer**
1	stealing	law	internet	social media
2	work	government	reading	government
3	crime	business	work	law
4	government	crime	journalism	messaging
5	wedding	help	writing	dispute
6	office	positive_emotion	messaging	internet
7	journalism	giving	office	work
8	reading	economics	social_media	phone
9	communication	stealing	science	computer
10	law	negative_emotion	business	communication

Chain Settings	Initial Stance (Claude Sonnet 4.5)		
	Neutral	Non-Believer	Believer
Baseline	42.40%	45.16%	39.30%
Non-Believer	50.61%	52.07%	42.80%
Believer	59.53%	47.10%	83.87%
Both	62.66%	65.34%	55.45%

Table 14: Model evaluation results (Claude Sonnet 4.5) on Overall Quality across initial stance of model evaluators and chain settings. Each value represents $\frac{\#\text{win}+\frac{1}{2}\#\text{tie}}{\#\text{total}}$.

Chain Settings	Initial Stance (Gemini 2.5 Pro)		
	Neutral	Non-Believer	Believer
Baseline	39.30%	43.21%	37.81%
Non-Believer	48.83%	50.26%	45.34%
Believer	67.71%	55.00%	77.25%
Both	65.55%	65.12%	63.98%

Table 15: Model evaluation results (Gemini 2.5 Pro) on Overall Quality across initial stance of model evaluators and chain settings. Each value represents $\frac{\#\text{win}+\frac{1}{2}\#\text{tie}}{\#\text{total}}$.

Chain Settings	Initial Stance (Grok 4.1)		
	Neutral	Non-Believer	Believer
Baseline	41.26%	40.78%	44.54%
Non-Believer	51.66%	52.69%	50.59%
Believer	54.33%	55.36%	51.53%
Both	70.23%	69.62%	64.24%

Table 16: Model evaluation results (Grok 4.1; `grok-4.1-fast-non-reasoning`) on Overall Quality across initial stance of model evaluators and chain settings. Each value represents $\frac{\#\text{win}+\frac{1}{2}\#\text{tie}}{\#\text{total}}$.

	GPT	Claude	Grok	Gemini	Human
GPT	—				
Claude	66.73%	—			
Grok	68.27%	63.05%	—		
Gemini	67.49%	68.19%	66.74%	—	
Human	66.17%	59.17%	55.00%	54.17%	—

Table 17: Weighted inter-model agreement across LLM evaluators and human annotators (N=40). Ties count as half agreement (0.5).