ENHANCING PERCEPTION: Refining Explanations of News Claims with LLM Conversations

Yi-Li Hsu[1,2], Jui-Ning Chen [1,3*], Shang-Chien Liu [1], Yang Fan Chiang [1], Aiping Xiong[4], Lun-Wei Ku[1]

[1]Institute of Information Science, Academia Sinica
[2]Department of Computer Science, National Tsing Hua University
[3]Graduate Institute of Electrical Engineering, National Taiwan University
[4]The Pennsylvania State University
yili.hsu@iis.sinica.edu.tw

Abstract

We introduce ENHANCING PERCEPTION, a framework for Large Language Models (LLMs) designed to streamline the time-intensive task typically undertaken by professional fact-checkers of crafting explanations for fake news. This study investigates the effectiveness of enhancing LLM explanations through conversational refinement. We compare various questioner agents, including state-of-the-art LLMs like GPT-4, Claude 2, PaLM 2, and 193 American participants acting as human questioners. Based on the histories of these refinement conversations, we further generate comprehensive summary explanations. We evaluated the effectiveness of these initial, refined, and summary explanations across 40 news claims by involving 2,797 American participants, measuring their self-reported belief change regarding both real and fake claims after receiving the explanations. Our findings reveal that, in the context of fake news, explanations that have undergone conversational refinement—whether by GPT-4 or human questioners, who ask more diverse and detail-oriented questions—were significantly more effective than both the initial unrefined explanations and the summary explanations. Moreover, these refined explanations achieved a level of effectiveness comparable to that of expert-written explanations. The results highlight the potential of automatic explanation refinement by LLMs in debunking fake news claims.

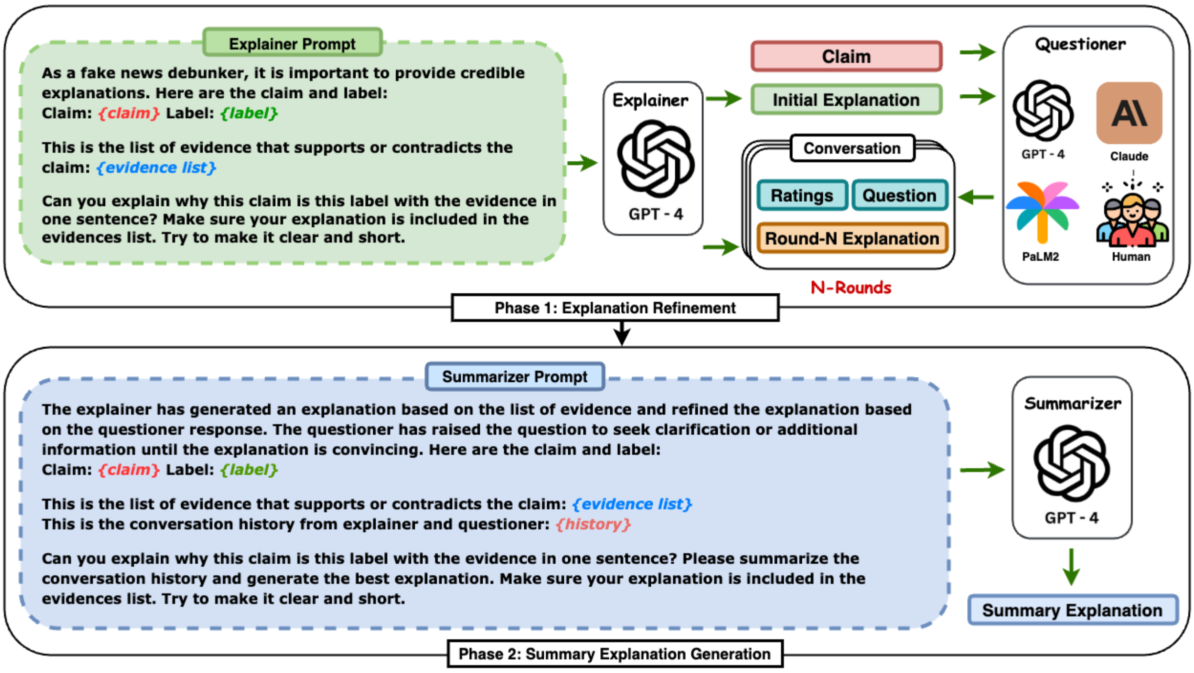
1 Introduction

Misinformation has increasingly become a significant threat to contemporary society, undermining public trust and distorting democratic discourse. (McKay and Tenove, 2021; Monsees, 2020; O'Connor and Weatherall, 2019) In recent years, fact-checking organizations such as PolitiFact, FactCheck.org, and the Washington Post's Fact Checker have made significant impacts on the political landscape by holding public figures accountable for their statements (Graves, 2016). However, due to the absence of a reliable source that can reflect the most up-to-date information, such fact-checking cannot be conducted in a fully automated manner and thus requires human involvement (Nguyen et al., 2019). This necessity increases both the time and labor costs associated with debunking fake news.

Recently, the advent of Large Language Models (LLMs), like GPT-4 (OpenAI, 2023), marks a significant milestone in the field of Natural Language Processing (NLP). These models possess the capacity to generate coherent and contextually relevant explanations, thereby offering a promising base for simplifying the extensive and demanding tasks that fact-checkers usually undertake to identify and debunk fake news. Previous research has demonstrated that GPT-generated summarizations of model self-refinement conversations show improvement over base medical explanations across three clinically-focused tasks (Nair et al., 2023). Despite this progress, the effectiveness of summary-based explanations in the context of debunking misinformation remains unexplored. Moreover, while Large Language Models (LLMs) have demonstrated exceptional capabilities in text generation and are widely accessible, their efficacy in generating explanations for debunking fake news, compared to expert-written content, is still under debate. In response to the challenges and unanswered questions posed by misinformation, our research aims to investigate the potential of LLMs to produce more persuasive and comprehensive explanations for debunking misinformation. This is achieved through the use of conversational self-refinement and summarization techniques. We have specifically designed our

[*] Co-first author

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2129–2147
June 16-21, 2024 ©2024 Association for Computational Linguistics



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}

Can you explain why this claim is this label with the evidence in one sentence? Make sure your explanation is included in the evidences list. Try to make it clear and short.

The explainer has generated an explanation based on the list of evidence and refined the explanation based on the questioner response. The questioner has raised the question to seek clarification or additional information until the explanation is convincing. Here are the claim and label:
Claim: {claim} Label: {label}

This is the list of evidence that supports or contradicts the claim: {evidence list}
This is the conversation history from explainer and questioner: {history}

Can you explain why this claim is this label with the evidence in one sentence? Please summarize the conversation history and generate the best explanation. Make sure your explanation is included in the evidences list. Try to make it clear and short.

Figure 1: **Overview of Enhancing Perceptions**: This two-phase framework begins with Phase 1, where GPT-4 serves as an Explainer, utilizing a provided evidence list and a verified label to produce an Initial Explanation to a claim. This explanation is then refined through iterative questioning by one of the various questioner agents, including GPT-4, Claude, PaLM2, or humans. In Phase 2, the detailed conversation history, claim, evidence list and label are then condensed into a Summary Explanation.

system to assist end-users, who may find it challenging to discern the veracity of claims, by providing them with automatically generated explanations. These explanations are intended to help users better understand and evaluate the information they encounter.

In this study, we assess the effectiveness of various explanation settings using 40 claims verified by experts at PolitiFact [1], engaging a diverse range of questioner agents, including state-of-the-art LLMs and human participants. We explore how these agents can aid in the iterative refinement of explanations to enhance users' evaluation capabilities. The framework of our approach is depicted in Figure 1. By engaging these agents to refine explanations provided by the GPT-4 Explainer for up to 15 iterations, we conduct a human-subject evaluation to assess the impact of different explanation settings. Our assessment focuses on how these explanations influence users' perceptions of real and fake news claims and explores the differences among questioner agents by hand-coding their questions. Given the diverse content that LLMs can produce due to their training, we hypothesize that the diversity and quality of questions significantly affect the outcomes of refinement. To enhance the understanding of the generated explanations, we incorporate an analysis using the Linguistic Inquiry and Word Count (LIWC) tool(Ryan L. Boyd and Pennebaker, 2022), aiming to provide deeper insights into the psycholinguistic patterns present. Overall, our study is guided by the following research questions:

(RQ1) How does the self-refinement process effectively enhance the explanation of both fake and real claims? (Section 5)

(RQ2) How does the questions diversity and quality affect the effectiveness of the refined explanations? (Section 6)

(RQ3) How do psycholinguistic patterns as analyzed by LIWC influence effectiveness of explanations? (Section 6)

2 Related Works

2.1 Fact-Checked Based Explanations

Although previous studies (Epstein et al., 2022; Moravec et al., 2020; Lutzke et al., 2019) have demonstrated the effectiveness of warnings and explanations in debunking misinformation,

[1] https://www.politifact.com/

research by (Hsu et al., 2023; Grady et al., 2021) indicates that while humans may be influenced by these interventions in the short term, biases often resurface in the long term. Additionally, there's a risk that users might overlook or misinterpret warning tags, contributing to the continued spread of misinformation. Conversely, with explanations, there's a challenge in ensuring users are motivated to engage with lengthy details for each piece of news. From a Natural Language Processing (NLP) perspective, current studies explore automatic ways to enhance the effectiveness and faithfulness of explanations by delivering concise and accurate information about news claims. Prior research has focused on generating explanations for identified misinformation, such as highlighting biased statements (Baly et al., 2018; Horne et al., 2019), multimodal explanation generation and verification (Yao et al., 2023), and generating faithful explanations by multi-agents debate (Kim et al., 2024). Dai et al. (2022) also introduced a framework for creating fact-checked counterfactual explanations. However, most existing attempts to debunk fake news were made before the emergence of Large Language Models (LLMs). Also, very few of them have incorporated human-subject evaluations to assess the impact of explanations generated by these frameworks.

A recent study by (Hsu et al., 2023) investigated the immediate and long-term efficacy of tag-based warnings versus GPT-4-generated explanations, finding no significant differences in their long-term impact. The researchers advocate for a shift towards personalized explanations, which may prove more effective than standard, less accessible explanations. Their findings highlight a notable gap in the literature concerning the empirical effectiveness of model generated explanations. Motivated by this gap, we investigate a novel framework that incorporates reader feedback to refine explanations with several iterations, aiming for explanations that are potentially more effective and user-friendly than both initial model-generated and expert-written explanations.

2.2 LLMs Can Improve by Self-Refinement

Previous works (Chen et al., 2023; Nair et al., 2023; Yao et al., 2022; Shinn et al.; Bai et al., 2022; Madaan et al., 2023) have shown that self-refinement enables Language Learning Models (LLMs) to enhance their performance and the accuracy of generated texts. The approach of Constitutional AI (Bai et al., 2022) employs repetitive questioning to deepen the model's understanding and improve accuracy. Similarly, (Madaan et al., 2023) refines outputs and incorporates iterative feedback from the model itself to enhance performance across multiple tasks. (Yao et al., 2022) implements a 'thought' process prior to action, aiming to mimic the human cognitive process and achieve similar improvements. Furthermore, the technique of reflection, as proposed by Reflexion (Shinn et al.), involves writing a reflection after a task has failed. This reflection is then utilized at the onset of the first action when repeating the task, which has been shown to optimize results. Alternatively, an approach to accuracy improvement through dialogue between two models has been introduced by DERA (Nair et al., 2023), demonstrating refined outcomes for GPT-4 in medical conversation summarization and care plan generation. While previous works have illustrated that the self-refinement process can improve LLM generation results, the application of such methodologies specifically within the context of explaining fake news remains an open question.

2.3 LIWC on Fake News

Continuing the exploration of fake news, Rubin et al. (2016) introduced an intriguing approach to identifying potentially misleading news by using cues indicative of satire utilizing the Linguistic Inquiry and Word Count (LIWC) (Ryan L. Boyd and Pennebaker, 2022). By examining stylistic and psycholinguistic patterns, they could flag news content that might not be genuinely factual. Building on this, Giachanou et al. (2022) delved deeper into the psycholinguistic aspects by analyzing the linguistic behavior of individuals who spread fake news compared to those who consult fact-checking resources. They found that users who tend to share fake news often use more informal language, while those who check facts tend to use more positive language and causality terms. They suggest that psycholinguistic patterns can be key indicators of deceptive language and highlight the potential of incorporating such features into automated systems designed to detect and refute fake news. In alignment with these findings, our study includes LIWC analysis to explore the correlation between psycholinguistic patterns and the effectiveness of explanations in debunking misinformation.

3 Methodology

Different from end-to-end explanation generation pipelines (Yao et al., 2023; Hsu et al., 2023), our framework includes different questioner agents in each iteration in the explanation refinement process. We have established a two-phase framework for refining and summarizing explanations of news claims, as illustrated in Figure 1. Initially, the GPT-4 Explainer provides an *Initial Explanation*. Subsequently, the Questioner agent—represented by either GPT-4, Claude 2, PaLM 2, or a human—evaluates this explanation across five distinct aspects. The Questioner then poses questions focusing on one of the explanation's identified weakest aspect in each round. Based on the received feedback, the GPT-4 Explainer refines the explanation to enhance its clarity and persuasiveness. This iterative refinement process continues for at least 3 rounds, ending when the Questioner decides to end the conversation or after a maximum of 15 rounds. At the conversation's end, the GPT-4 Explainer produces a *Refined Last Round Explanation*. The GPT-4 Summarizer then generates a *Summary Explanation* base on the entire conversation history of the Explainer and the Questioner. A detailed description of this two-phase framework is provided below.

3.1 Phase 1: Explanation Refinement

In this phase, the *Initial Explanation* generated by the GPT-4 Explainer is iteratively refined by the questioner into the *Refined Last Round Explanation*. At the beginning of this process, the GPT-4 Explainer uses a provided claim and a corresponding list of evidence to generate an *Initial Explanation*. This explanation, along with the claim, undergoes iterative refinement rounds by a questioner, which can be GPT-4, Claude 2, PaLM 2, or humans. The Questioners assess the explanation across five different aspects: persuasiveness, logical correctness, completeness, conciseness, and agreement, each rated on a five-point scale. They also select the weakest aspect of the explanation and ask a question in each iteration. We contend that these aspects address the majority of the requirements for refining an explanation, with only a few participants indicating satisfaction or providing no response. A detailed breakdown of the aspect selections is presented in Table 6. Our framework allows for 3 to 15 rounds of evaluation and questioning, enabling Questioners to terminate the conversation starting from the third round if the explanation reaches a satisfactory level—defined as all five aspects scoring above 4, with at least one aspect scoring a 5. Alternatively, they can opt to conclude the conversation after 5 rounds of refinement, regardless of the scores. The process is designed to end by the 15th round.

3.2 Phase 2: Summary Explanation Generation

After the conversation concludes, we utilize the GPT-4 Summarizer to conclude the entire conversation history—including all iterations of explanations, ratings, and questions—into *Summary Explanation*. This summary is expected to offer a more comprehensive understanding than either the initial or the refined explanations due to the the incorporation of the overall conversation history.

Hereafter, the last round of explanation provided by the GPT-4 Explainer will be reffered as the *Refined Last Round Explanation*, and the explanation produced by the GPT-4 Summarizer as the *Summary Explanation*. Examples of these explanations are presented in Table 8.

3.3 Data Collection

Sourcing from PolitiFact Following the LIAR-PLUS (Alhindi et al., 2018) methodology, we extracted claim, evidence, and experts' explanations from Politifact.com. We crawled a dataset of 50 verified news articles which spans from 2019 to 2021. These articles were evenly distributed across five fact-checked labels: True, Mostly-True, Half-True, Barely-True, and False. The claims within these articles encompass a broad spectrum of U.S. political news. We defined 'True' and 'Mostly-True' labels as real claims, and 'Barely-True' and 'False' as fake claims. In the human subject evaluation experiment, we exclude 'Half-True' samples to ensure an balanced number of real and fake claims.

Evidence Processing To enhance the quality of our explanations, we processed each piece of evidence into a list using an independent GPT-4 model. This process involved categorizing them with stances such as 'SUPPORT' or 'REFUTE' in relation to the claim. Details of the prompt used for this categorization can be found in Table 10.

3.4 Questioner Agents Details

Large Language Models We have implemented the Questioners using GPT-4 and two other state-of-the-art LLMs: PaLM 2 and Claude 2, engaging them in up to 15 rounds of conversation. During inference, we set the temperature to 0.7 for the GPT-4 to encourage diversity in the generated questions, while employing the default settings for PaLM 2 and Claude 2. All implementations involving GPT-4 utilize the 'gpt-4-0613' model from OpenAI. Results from Claude 2 and PaLM 2 are retrieved by submitting requests to their respective websites. The questioners have same prompt as outlined in Figure. 8

Human Annotators After providing informed consent and receiving a brief overview to the study, participants are randomly assigned to one of 50 claims. For each of the claims, we recruit 3 American annotators from Prolific [2], resulting in a total of 193 annotators, excluding those who did not pass the attention check. Participants were then asked to rate their familiarity with the claim on a five-point scale, with the prompt: *'Have you ever seen or heard about this claim?'* (1 = Definitely not, 5 = Definitely yes). Treating 'Probably yes' and 'Definitely yes' responses as indicators of familiarity, the average familiarity score was 0.16 on a scale from 0 to 1. Furthermore, they assessed the perceived accuracy of the claim, both before and after interacting with the GPT-4 Explainer, using a seven-point scale as per Sindermann et al. (2021): *'To the best of your knowledge, how accurate is the claim?'* (1 = Definitely not accurate, 7 = Definitely accurate). Subsequently, participants entered our refinement environment to engage in dialogue with the GPT-4 Explainer, starting with the Initial Explanation. The average perceived accuracy scores were 0.25 before and 0.41 after the refinement process on a scale from 0 to 1. Demographic information is detailed in Table 5 and the interfaces of the Explanation Refinement Environment are shown in Fig. 6 and Fig. 7.

4 Evaluation Settings

4.1 Evaluation Settings Overview

In our study, we examine four main types of explanations: *Initial Explanation* generated by the Explainer GPT-4 without any refinement process; *Expert-Written Explanation* sourced from PolitiFact, *Refined Last Round Explanation*, produced after the explanation refinement process; and *Summary Explanations*, which summarizes the conversational refinement. The latter two types each encompass three settings, derived from interactions in which the GPT-4 Explainer collaborates with one of three Questioner Agents: GPT-4, Claude 2, or human annotators. This results in a total of eight distinct explanation settings, as shown in Fig. 2. We excluded PaLM 2 from this experiment due to its tendency to generate repetitive questions and receive lower-quality scores, which did not contribute to enhancing the quality of explanations. The observations and their implications are further explored in Section 6. In scenarios involving human questioners, each claim was assessed by three different annotators, with the responses from one being randomly chosen for the evaluation stage of our experiment.

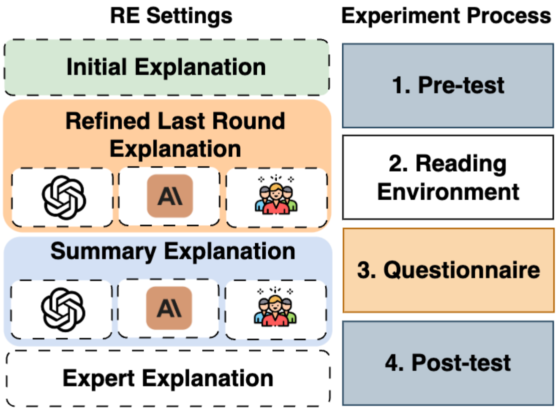


Figure 2: Human Evaluation overview: The evaluation follows a structured process involving a Pre-test, Reading Environment (RE) which participants are assigned to one of the eight different explanations settings, a Questionnaire phase, and a Post-test. This design was implemented to ensure that each subject's experience was consistent within the assigned RE setting, allowing for a controlled evaluation of the effectiveness of explanations.

4.2 Procedure

To assess the effectiveness of the eight different explanation settings, we recruited 2,797 American participants from Prolific, ensuring that all participants were distinct from the human questioners involved in the study. Demographic details can be found in Table 5. Our goal was to examine how these explanations influenced participants' beliefs regarding the claims. We follow the evaluation process and metrics from

[2] www.prolific.com

(Hsu et al., 2023), which consisted of pre-test, reading environment, questionnaire, and post-test phases, as illustrated in Fig. 2. We used 40 claims from our self-curated dataset as described in Section 3.3. The claims were equally distributed in real and fake labels. In the Reading Environment, each participant was assigned to one of the eight explanation settings, with half of the claims presented along with the explanation specific to that setting. The experiment details are explained in the following sections.

Pre-test Phase Participants initially assessed four fake and four real news claims without being provided with any explanations. This pre-test phase required them to rate both their familiarity with each claim and its perceived accuracy, employing the same questions and options presented to the human questioners in the refinement process, as detailed in Section 3.4.

Reading Environment Phase In this phase, participants were exposed to the same claims in a random order, but with explanations provided for half of the fake claims and half of the real claims. The interface of the reading environment is shown in Fig. 3

Questionnaire Following the reading environment phase, we collected feedback on the overall survey experience. We strategically placed this questionnaire before the post-test phase to minimize its influence on memory recall.

Post-test Phase During the post-test phase, participants were asked to reassess the same eight claims, answering questions about perceived accuracy with the claims presented in random order. This allowed us to evaluate changes in perception caused by the explanations.

After completing the experiment, all participants were shown all claims along with their verified labels, and were informed that the source of the claims was PolitiFact.com. This was done to prevent any possible misleading impressions about the unexplained claims.

5 Experiment Results

We initially address RQ1 of our study to focus on evaluating the effectiveness of *Refined Last Round Explanations* and *Summary Explanations*, produced after the refinement process, in comparison with *Expert-Written Explanations* and *Initial Explanations* generated by the GPT-4 Explainer. Following this analysis, we proceed to address our RQ2 and RQ3 in Section 6. We will discuss the impact of questions posed by different questioner agents during the refinement process. This exploration aims to understand how the diversity and quality of these questions contribute to the effectiveness of the refined explanations, providing further insights into the our explanation refinement framework. Finally, we will explore the psycholinguistic patterns of explanations in different settings.

Refined Explanations can be as Effecitve as Expert's Explanations for Fake Claims Figure 1 presents the findings from our human-subject evaluation. In the context of fake claims, we found the *Initial Explanations* were significantly worse than *Expert-Written Explanations*. However, after the refinement process, both *GPT-4 Refined Last Round Explanations* [3] and *Human Refined Last Round Explanations* [4] outperformed *Initial Explanations*, with Chi-squared analysis confirming a substantial difference in effectiveness. Furthermore, both GPT-4 and human *Refined Last Round Explanations* was comparable to that of *Expert-Written Explanations*[5].

These results highlight the significant impact of both GPT-4 and human-refined explanations in correcting users' misconceptions about misinformation, marking a clear improvement over the GPT-4's initial explanations. However, Claude 2's *Refined Last Round Explanations*, do not result different effectiveness from the *Initial Explanations* [6] and are significantly less effective than *Expert-Written Explanations* [7]. This could be due to Claude 2's higher rate of question duplication, as suggested by Table 2 and less diverse question types as shown in Figure 5. A more detailed qualitative analysis of the questions posed by the different questioner agents will be discussed in Section 6.

Conversational Refinement Does Not Improve Explanations for Real Claims The results from Table 1 also shows consistently larger differences of accuracy rate for real claims than for fake claims across all explanation settings. Indicate

[3] $\chi^2_{(1)} = 14.33, p < 0.001$

[4] $\chi^2_{(1)} = 14.42, p < 0.001$

[5] $\chi^2_{(1)} < 1, p > 0.05$

[6] $\chi^2_{(1)} < 1, p > 0.05$

[7] $\chi^2_{(1)} = 4.04, p < 0.05$

Table 1: **Human Evaluation Results**: The participants' accuracy and flip rates of judging the veracity of claims under different explanation settings in reading environment.

	Accuracy Rate	
	Pre-/Post-test	**Diff**
Real Claims with Explanations in RE		
GPT-4-init	31% / 79%	48%
Refined Last Round Explanation		
GPT-4	28% / 66%	38%
Human	33% / 69%	36%
Claude 2	29% / 78%	49%
Summary Explanation		
GPT-4	30% / 77%	47%
Human	29% / 79%	49%
Claude 2	33% / 80%	47%
Expert	33% / 77%	44%
Fake Claims with Explanations in RE		
GPT-4-init	48% / 71%	23%
Refined Last Round Explanation		
GPT-4	43% / 75%	32%
Human	41% / 73%	32%
Claude 2	45% / 70%	25%
Summary Explanation		
GPT-4	48% / 71%	23%
Human	42% / 71%	29%
Claude 2	44% / 70%	26%
Expert	45% / 74%	29%

(a) **Participants' Accuracy in Judging the Veracity of Claims Before and After Receiving Explanations:** Significant differences in accuracy rates, as determined by the chi-squared test, are indicated with highlights. Blue shows significant **worse** than *Expert-Written explanations*, while Green indicates significant **better** than the *Initial GPT-4 explanations (GPT-4-init)*.

	Pre-test Rate			**Flip Rate**			
	✗	✓	▲	✗ → ✓ (↑)	▲→✓ (↑)	▲→✗ (↓)	✓ → ✗ (↓)
Real Claims with Explanations in RE							
GPT-4-init	37%	31%	33%	25%	53%	10%	1%
Refined Last Round Explanation							
GPT-4	39%	28%	34%	23%	46%	14%	6%
Human	33%	33%	34%	19%	47%	15%	8%
Claude 2	40%	29%	31%	27%	55%	12%	4%
Summary Explanation							
GPT-4	36%	30%	33%	25%	55%	11%	3%
Human	40%	29%	31%	28%	55%	11%	2%
Claude 2	36%	33%	31%	27%	55%	10%	2%
Expert	36%	33%	30%	22%	53%	10%	1%
Fake Claims with Explanations in RE							
GPT-4-init	23%	48%	29%	13%	43%	17%	5%
Refined Last Round Explanation							
GPT-4	24%	43%	33%	16%	51%	12%	4%
Human	25%	41%	33%	15%	51%	16%	5%
Claude 2	24%	45%	31%	14%	46%	18%	7%
Summary Explanation							
GPT-4	22%	48%	30%	12%	45%	19%	6%
Human	26%	42%	32%	15%	47%	18%	6%
Claude 2	26%	44%	31%	15%	46%	18%	7%
Expert	25%	45%	30%	14%	51%	12%	4%

(b) **Participants' Fine-Grained Flip Rate:** This fine-grained breakdown details how participants' judgements on the veracity of claims changed before and after receiving explanations. The symbol (✗→✓) denotes a shift from an initially incorrect judgement to a correct one after receiving explanation, indicating an improvement in accuracy. In contrast, the symbol (✓→✗) represents an initial correct judgement that was later revised to an incorrect one, showing a decrease in accuracy. The symbol (▲) reflects participants who initially expressed uncertainty regarding the claim's accuracy.

that real news claims are easier to explain and debunk through explanations than fake news. Participants were more effectively persuaded of the truthfulness of real news with the explanations. The results also show that our conversational refinement framework does not improve the effectiveness of explanations for real claims. Despite involving various questioners to refine explanations, there is no observable improvement in participants' accuracy when compared to *Initial Explanations*. Moreover, the *Initial Explanations* exhibit an insignificant difference when compared to *Expert-Written Explanations* [8]. This outcome may suggest that the quality of *Initial Explanations* for real claims is already at an optimal level, or that the characteristics of real claims are such that additional conversation does not significantly enhance clarity. This could also indicate a ceiling effect, where the *Initial Explanations* are sufficiently effective, so further refinement does not result in additional performance.

The Redundancy of Summary Explanations For the *Summary Explanations*, only those involving human questioners managed to exceed the effectiveness of the *Initial Explanations*[9], achieving an effectiveness level on par with *Expert-Written Explanations* in fake claims[10]. This outcome suggests that the comprehensive summary explanation of conversational refinements might not be helpful and may even cause redundancy for correcting misbeliefs.

[8] $\chi^2_{(1)} = 2.38, p = 0.12$

[9] $\chi^2_{(1)} = 6.42, p < 0.05$

[10] $\chi^2_{(1)} < 1, p > 0.05$

6 Discussion

We randomly selected 20 claims, evenly distributed between real and fake, to conduct hand-coded topic analysis. We further broke them down by the first three words of the questions asked by all questioner agents. The visualized results for real and fake claims can be found in Figs. 4 and 5, respectively.

Diverse Questions can lead to better refined explanations The analysis reveals that both GPT-4 and human questioners asked questions across a wide spectrum of topics with a variety of starting words, while the other LLMs ask very similar questions. For fake news, it is observed that GPT-4 questioner focuses on academic and policy-related information. Also, human questioners frequently ask the GPT-4 Explainer to justify its explanations and request further details about the claims. Despite the variance in question topics between GPT-4 and humans, both contribute to enhancing explanations by addressing a wider range of issues, thereby fostering a deeper understanding and more effective refinement of explanations.

GPT-4 and human questioners also show a lower rate of duplicate questions compared to other LLMs, as indicated in Table 2. This observation highlights the advanced diversity in their questioning methods, which is critical in the context of explanation refinement. Unlike GPT-4 and humans, PaLM 2 was observed to repeat the refinement process for up to 15 rounds for all examined claims, yet this often resulted in redundant questions that did not contribute to the refinement's effectiveness. Similarly, while Claude 2's engagement in conversational refinement for approximately 6.4 rounds, a high duplication rate and a lack of question diversity lead to negligible improvements over the initial explanations provided as shown in Fig. 1. These findings illustrate that GPT-4 and humans employ a higher level of questioning ability with a more detail-oriented and diverse questioning approach, significantly enhancing the refinement process to produce more convincing and detailed explanations.

	#Questions	Type 1	Type 2	Type 3
GPT-4	225	11	**18**	0
Human	89	3	**8**	0
Claude 2	127	**49**	0	6
PaLM 2	300	**267**	0	0

Table 2: Distribution of Question Types and Errors Identified in Responses from GPT-4, Human, Claude 2, and PaLM 2. The table categorizes errors into three types: Type 1 represents same or duplicate question errors, Type 2 indicates question format errors, and Type 3 denotes errors in generation.

	WC	Analytic	Authentic	Tone	BigWords	Dic	Function
GPT-4-init	51.88	82.04	25.81	29.96	28.79	81.79	47.01
Refined Last Round Explanation							
GPT-4	64.48	84.75	28.92	25.55	30.59	81.34	45.93
Human	46.38	87.02	32.72	32.66	30.88	79.8	43.65
Claude 2	55.61	89.02	30.68	27.19	32.09	79.68	42.82
Summary Explanation							
GPT-4	63.25	84.84	30.3	25.44	31.95	80.72	45.03
Human	58.54	83.71	26.11	25.78	31.54	80.3	45.69
Claude 2	61.22	83.81	32.36	28.93	30.67	80.0	45.49
Expert	11.66	71.06	32.64	27.81	32.05	71.05	31.05
p-value	0.000	0.000	0.853	0.855	0.662	0.000	0.000

Table 3: LIWC Result with ANOVA Test p-values

Humans May Have Bias on AI-Explaners Unexpectedly, we observed that a few human questioners ask the GPT-4 Explainer, "Why should I believe you?", revealing potential skepticism towards AI-generated explanations for news claims. Such distrust may undermine the refinement process by prompting unproductive discussions. Moreover, we found some circular human-GPT-4 conversations focused on the AI's credibility rather than the substance of model generated explanations, diverting attention from a meaningful assessment of the claims. The finding also suggests that for AI explanations to be more readily accepted and the refinement process to be truly effective, establishing the AI's credibility upfront seems to be necessary.

Neutral Opinions Are Easier to Flip The flip rates depicted in Table 1 show that participants initially holding a neutral stance (▲), indicative of their uncertainty about the truthfulness of a claim, were more likely to change their opinions to align with the correct veracity after being exposed to the explanations (▲→✓). This finding is particularly noteworthy because it highlights that individuals without strong initial opinions are more acceptable to changing their views of claims in response to explanations. We contend that the focus should not solely be on individuals with solid beliefs but also on engaging those who are undecided. Such individuals may be more open to revising their stances when presented with compelling and well-constructed explanations.

	Real News Claims	Fake News Claims
GPT-4	Can you provide information on how often the ’stepped-up basis’ policy has been used historically and by what demographic of inheritors?	Can you explain how differing healthcare capabilities between regions could impact fatality rate estimates?
Human	Why should I believe you?	What else should his wife be responsible for?
Claude 2	Could you provide an example of how much additional tax revenue could be generated from eliminating the stepped-up basis policy?	Could you provide more details on what the protesters were chanting in the 2017 video to confirm it was not related to the Jan 6, 2021 incident?
PaLM 2	Could you please provide more information on how New Zealand’s strategy against COVID-19 involved major health figureheads like Ashley Bloomfield?	If the video is from a 2017 Senate hearing, why do some people believe it shows police removing disabled protesters from the Capitol on Jan. 6, 2021?

Table 4: Example Question from Each Setting

Shorter Is Not Always Better We also incorporate an LIWC-22 analysis (Ryan L. Boyd and Pennebaker, 2022) to examine the linguistic characteristics of explanations across different settings. The data summarized in Table 3 present the primary outcomes from the LIWC analysis. These results indicate that experts typically write much shorter explanations, averaging around 11 words, compared to GPT-4 explanations across all settings, which range from 50 to 60 words. Contrary to our initial hypothesis, the *Refined Last Round Explanations* generated by GPT-4 achieved results comparable to those written by experts. Moreover, while GPT-4’s *Initial Explanations* feature the lowest word count (WC), they do not attain the highest effectiveness, as illustrated in Table 1.

Linguistic Features by LIWC Are Not Critical for Persuasion Effectiveness The linguistic analysis of GPT-4 generated explanations reveals a consistently high level of analytical thinking (Analytic), as depicted in Table 3. However, this analytical quality does not necessarily lead to greater effectiveness compared to the explanations provided by experts. While factors such as authenticity (Authentic), emotional tone (Tone), and the use of complex vocabulary (BigWords) show no significant differences according to ANOVA tests, Table 1 illustrates that their impact on altering participants’ stances can be significantly different. This consistency is observed across various settings of explanations, whether generated, refined, or summarized by GPT-4 Explainer, or written by experts. The lack of distinction in these linguistic metrics suggests that the effectiveness of explanations may be attributed to other features, which might not be fully captured by linguistic measures alone.

7 Conclusion

Our investigation into the iterative refinement of explanations for both real and fake news claims reveals insights into the field of debunking misinformation. Notably, GPT-4 and human questioners emerge as significantly effective in refining explanations, highlighting the power of diverse and in-depth questioning. This contrasted with the performances of PaLM 2 and Claude 2, which did not exhibit noticeable improvement in explanation quality, underscoring the importance of the questioning approach’s diversity and depth.

Our analysis further demonstrated that participants with initially neutral opinions were more amenable to changing their views, emphasizing the potential impact of well-crafted explanations on those undecided about a claim’s veracity. However, the skepticism expressed by human questioners towards AI-generated explanations underscores the ongoing challenge of establishing AI credibility in misinformation mitigation efforts.

Moreover, our study introduced LIWC-22 analysis to examine the linguistic characteristics of the explanations, revealing that the length and the linguistic features did not necessarily enhance effectiveness of explanations. This finding suggests that the effectiveness of expert explanations may derive from their ability to provide depth and context beyond what is captured by psycholinguistic patterns alone. This insight emphasizes the need for future misinformation explanations to go beyond surface linguistic patterns to have deeper engagement with the underlying context to truly influence the audience.

Limitations

Our study focusing on the U.S. news claims in English, along with the use of American annotators, recruited on Prolific.com, may introduce limitations related to cultural bias and the generalizability of our findings to a global context. These limitations should be kept in mind when interpreting and applying the results of our research.

Additionally, it's essential to acknowledge that the experimental process and environment in our study may differ significantly from the real-world situations in which news claims are encountered and evaluated. This divergence can introduce certain limitations that might impact the applicability of our findings to real-world scenarios.

Last, the use of Large Language Models (LLMs) like GPT, Claude, and PaLM2 for generating and refining explanations, while they have state-of-the-art reasoning and text generation ability, may also pose limitations. The evolving nature of these technologies and their underlying algorithms might lead to varying performance over time, potentially affecting the consistency and reproducibility of our results in future applications.

Ethics 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' long-term 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 \$8.4/ hour, which is above the federal minimum wage in the United States.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2800–2810, New York, NY, USA. Association for Computing Machinery.

Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193.

Anastasia Giachanou, Bilal Ghanem, Esteban A Ríssola, Paolo Rosso, Fabio Crestani, and Daniel Oberski. 2022. The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data & knowledge engineering*, 138:101960.

Rebecca Hofstein Grady, Peter H Ditto, and Elizabeth F Loftus. 2021. Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time. *Cognitive research: principles and implications*, 6:1–16.

Lucas Graves. 2016. *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.

Benjamin D Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. 2019. Rating reliability and bias in news articles: Does ai assistance help everyone? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 247–256.

Yi-Li Hsu, Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. Is explanation the cure? misinformation mitigation in the short term and long term.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate.

Lauren Lutzke, Caitlin Drummond, Paul Slovic, and Joseph Árvai. 2019. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. *Global environmental change*, 58:101964.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Spencer McKay and Chris Tenove. 2021. Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, 74(3):703–717.

Linda Monsees. 2020. 'a war against truth'-understanding the fake news controversy. *Critical Studies on Security*, 8(2):116–129.

Patricia L Moravec, Antino Kim, and Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, 31(3):987–1006.

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.

Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. 2023. Dera: Enhancing large language model completions with dialog-enabled resolving agents.

Thanh Tam Nguyen, Matthias Weidlich, Hongzhi Yin, Bolong Zheng, Quoc Viet Hung Nguyen, and Bela Stantic. 2019. User guidance for efficient fact checking. Technical report.

Cailin O'Connor and James Owen Weatherall. 2019. *The misinformation age: How false beliefs spread*. Yale University Press.

OpenAI. 2023. Gpt-4 technical report.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Sarah Seraj Ryan L. Boyd, Ashwini Ashokkumar and James W. Pennebaker. 2022. *The Development and Psychometric Properties of LIWC-22*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, june 2023. *arXiv preprint arXiv:2303.11366*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

A Demographics

Table 5: Demographic information of valid human questioners and participants in our human-subject experiment.

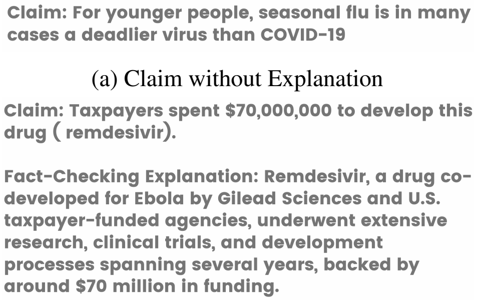
Item	Options	*Percentage*	*N*
Sex	Female	38.46%	70
	Male	61.54%	112
	Other	0%	
Age	18–24	8.24%	15
	25–34	33.51%	61
	35–44	20.33%	37
	45–54	19.78%	36
	Over 55	17.03%	31
Ethnicity	Black	13.19%	24
	White	68.13%	124
	Asian	5.49%	10
	More than one race	8.24%	15
	Other	4.95%	9

(a) Valid human questioners.

Item	Options	*Percentage*	*N*
Sex	Female	49.25%	1,378
	Male	50.75%	1,420
	Other	0%	
Age	18–24	7.97%	223
	25–34	26.52%	742
	35–44	24.41%	683
	45–54	18.83%	527
	Over 55	21.52%	602
Ethnicity	Black	11.47%	321
	White	72.73%	2,035
	Asian	5.9%	165
	More than one race	6.36%	178
	Other	3.54%	99

(b) Valid participants in our human-subject experiment.

B Reading Environment in Human Evaluation



Claim: For younger people, seasonal flu is in many cases a deadlier virus than COVID-19

(a) Claim without Explanation

Claim: Taxpayers spent $70,000,000 to develop this drug (remdesivir).

Fact-Checking Explanation: Remdesivir, a drug co-developed for Ebola by Gilead Sciences and U.S. taxpayer-funded agencies, underwent extensive research, clinical trials, and development processes spanning several years, backed by around $70 million in funding.

(b) Claim with Explanation

Figure 3: Reading Environment Interface

C Weakest Aspect of Explanations selected by Human Questioners

In the refinement process, Questioners were tasked with identifying the weakest aspect of an explanation and asking a question to address it. Our analysis suggests that the considered aspects were comprehensive, addressing the majority of explanation refinement requirements. The distribution of the questions posed by the human Questioners across different aspects is presented in Table 6.

Aspects	**N**
Completeness	288
Persuasiveness	191
Conciseness	163
Logical Correctness	121
Others	
No Answer	50
Satisfied	50
Other Free Text	25

Table 6: Distribution of Selected Weakest Aspects for Explanation Refinement by Human Questioners

D Pre-test Familiarity Rate in Human Evaluation

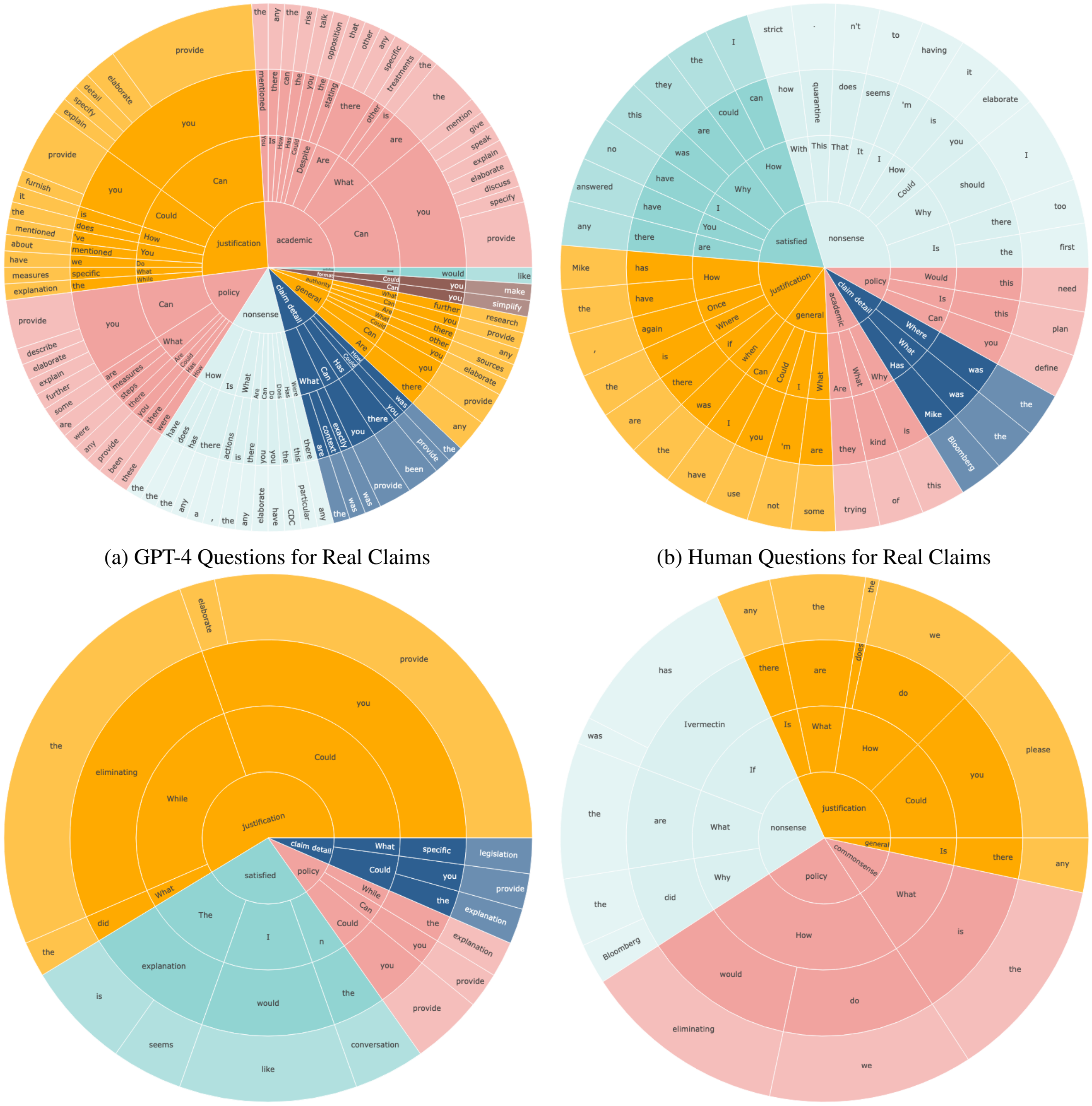
	Real	**Fake**
GPT-4-init	19%	16%
Refined Last Round Explanation		
GPT-4	17%	19%
Human	22%	19%
Claude 2	17%	19%
Summary Explanation		
GPT-4	18%	17%
Human	19%	18%
Claude 2	21%	19%
Expert	22%	21%

Table 7: Pre-test Familiarity Rates for Real and Fake Claims

As shown in Table. 7. For the Real Claims, the Chi-square test[11] indicated no significant difference across the groups. Similarly, for the Fake Claims, the Chi-square test[12] also showed no significant differences. These results suggest that the familiarity rates for "Real" and "Fake" claims are consistent across the groups.

[11] $\chi^2_{(5)} = 0.444, p = 0.994$

[12] $\chi^2_{(5)} = 0.417, p = 0.995$



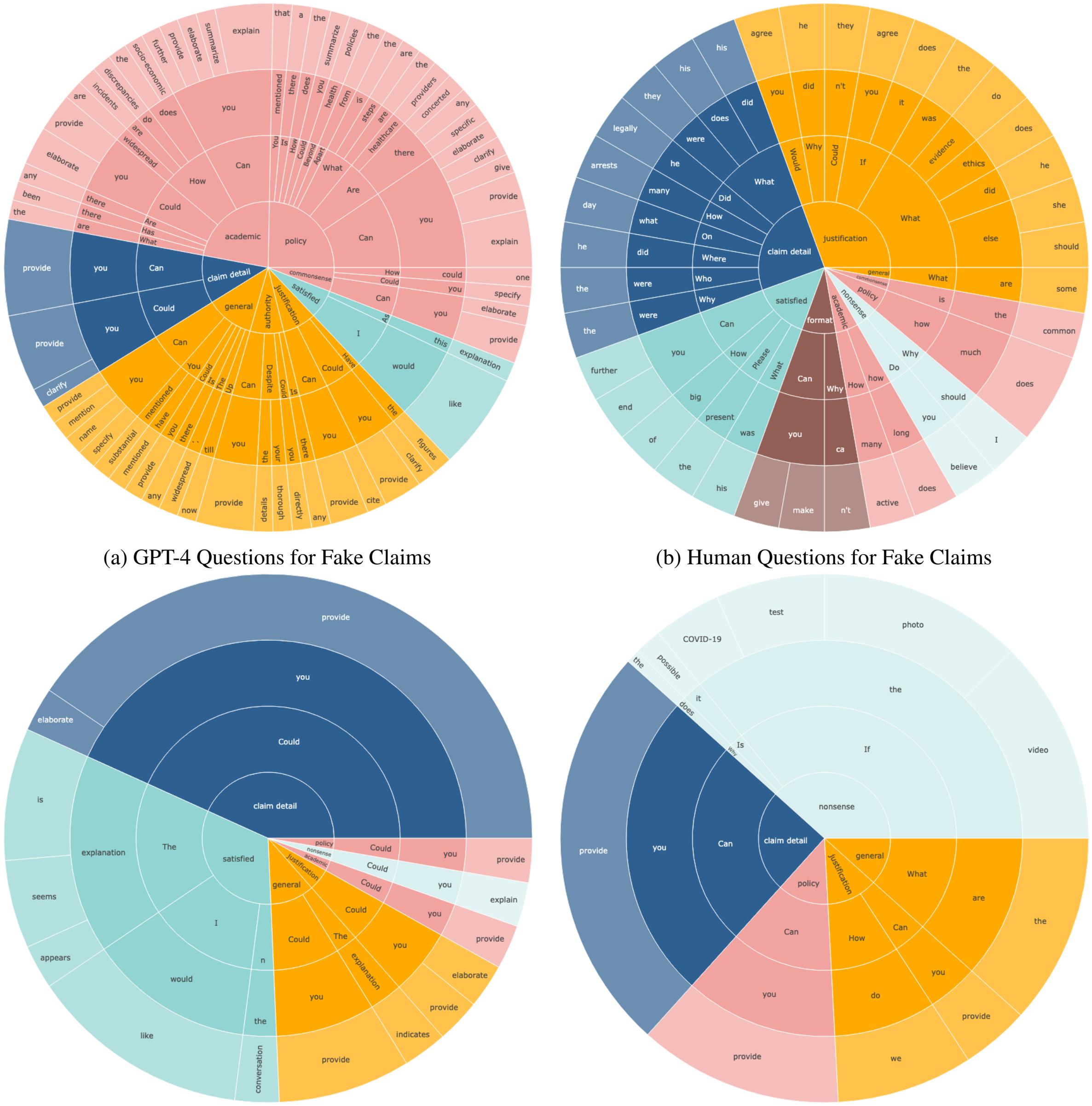
(a) GPT-4 Questions for Real Claims

(b) Human Questions for Real Claims

(c) Claude 2 Questions for Real Claims

(d) PaLM 2 Questions for Real Claims

Figure 4: **Distribution of Hand-Coded Question Topics on Real Claims** The pie charts display the hand-coded distribution of human-written and GPT-4-generated questions, categorized by the nature of the inquiry: pink represents questions requiring external knowledge, including academic, policy, and commonsense knowledge; orange indicates questions about the evidence source, including general sources and authoritative sources; blue refers to requests for further claim details; gray highlights nonsensical questions; and light green denotes responses that express satisfaction with the explanation provided.



(a) GPT-4 Questions for Fake Claims

(b) Human Questions for Fake Claims

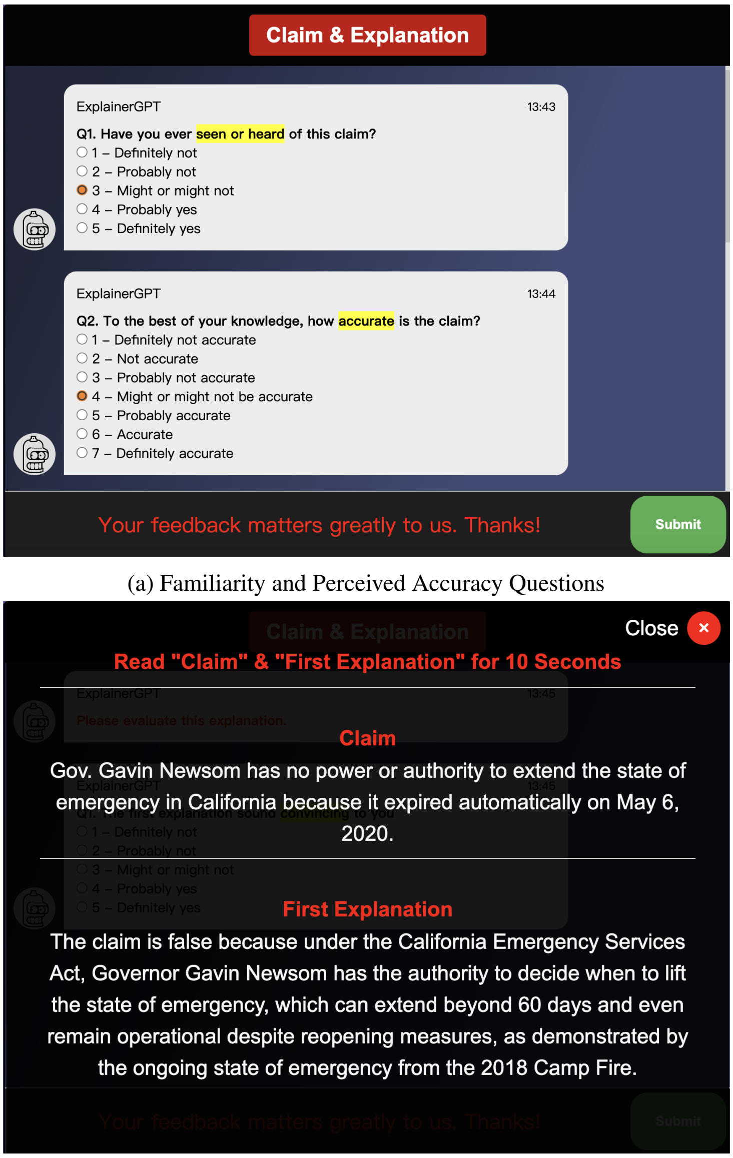
(c) Claude 2 Questions for Fake Claims

(d) PaLM 2 Questions for Fake Claims

Figure 5: **Distribution of Hand-Coded Question Topics on Fake Claims** The pie charts display the hand-coded distribution of human-written and GPT-4-generated questions, categorized by the nature of the inquiry: pink represents questions requiring external knowledge, including academic, policy, and commonsense knowledge; orange indicates questions about the evidence source, including general sources and authoritative sources; blue refers to requests for further claim details; gray highlights nonsensical questions; and light green denotes responses that express satisfaction with the explanation provided.

Claim	**Video shows police removing disabled protesters from the Capitol on Jan. 6, 2021.**
	Initial Explanation
GPT-4	The claim is false because the video in question is from a 2017 Senate hearing on health care legislation, not from the Capitol incident on Jan. 6, 2021, which can be confirmed by the protesters' chants of "No cuts to Medicaid, save our liberty.
	Refined Final Round Explanation
GPT-4	The claim is false as the video showing police escorting protesters, many in wheelchairs chanting "No cuts to Medicaid, save our liberty," originated from a 2017 Senate hearing on health care legislation, with the U.S. Capitol police arresting 181 individuals during this event, as reported in several media outlets and contrary to the assertion that it's from the Jan 6, 2021 Capitol incident.
Human	During the 2017 Senate hearing, which is the actual event shown in the misattributed video, U.S. Capitol police arrested 181 people for disrupting Congress during a protest against repealing and replacing the Affordable Care Act.
Claude 2	The claim is false, as the video shows a 2017 Senate hearing on health care, not the Jan 6, 2021 Capitol incident, and the chant "No cuts to Medicaid, save our liberty," distinctly heard in the video, is in reference to the proposed health care legislation at the time and not related to events of the latter date.
PaLM 2	The claim is false because the video, although showing police removing disabled protesters from a building, was actually taken during a Senate hearing on health care legislation in 2017, not the Capitol incident on Jan. 6, 2021, but was misinterpreted due to its recirculation on Facebook amid the extensive news coverage and attention given to the Capitol riots.
	Summary Explanation
GPT-4	The claim that a video shows police removing disabled protesters from the Capitol on Jan. 6, 2021 is false as the video was actually taken during a 2017 Senate hearing on health care reform, evidenced by the protesters' chants of "No cuts to Medicaid, save our liberty" and the U.S. Capitol police records indicating that 181 protesters were arrested during this event, which drastically differ from the circumstances surrounding the events on Jan. 6, 2021.
Human	The claim that the video shows police removing disabled protestors from the Capitol on Jan. 6, 2021, is false because the video is actually from a Senate hearing on health care legislation in 2017 and the protestors can be heard chanting against cuts to Medicaid, leading to the Capitol police arresting 181 people due to disruption of Congress.
Claude 2	The claim is labeled as false because the video in question, which depicts police escorting protesters (including those in wheelchairs) chanting "No cuts to Medicaid, save our liberty," is from a 2017 Senate hearing on health care legislation, not from the Capitol incident on Jan. 6, 2021.
PaLM 2	The claim is false because the video showing police removing disabled protesters was actually from a 2017 Senate hearing on health care legislation and was misidentified as footage from Jan 6, 2021 Capitol incident due to its recent recirculation on Facebook during the prevalent news coverage of the Capitol riots.
	Expert Explanation
Expert	Protest footage from 2017 confuses some social media users

Table 8: Example Explanations of Difference Settings



Claim & Explanation

ExplainerGPT 13:43

Q1. Have you ever seen or heard of this claim?

○ 1 – Definitely not
○ 2 – Probably not
● 3 – Might or might not
○ 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:44

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

Your feedback matters greatly to us. Thanks!

(a) Familiarity and Perceived Accuracy Questions

Read "Claim" & "First Explanation" for 10 Seconds

Claim

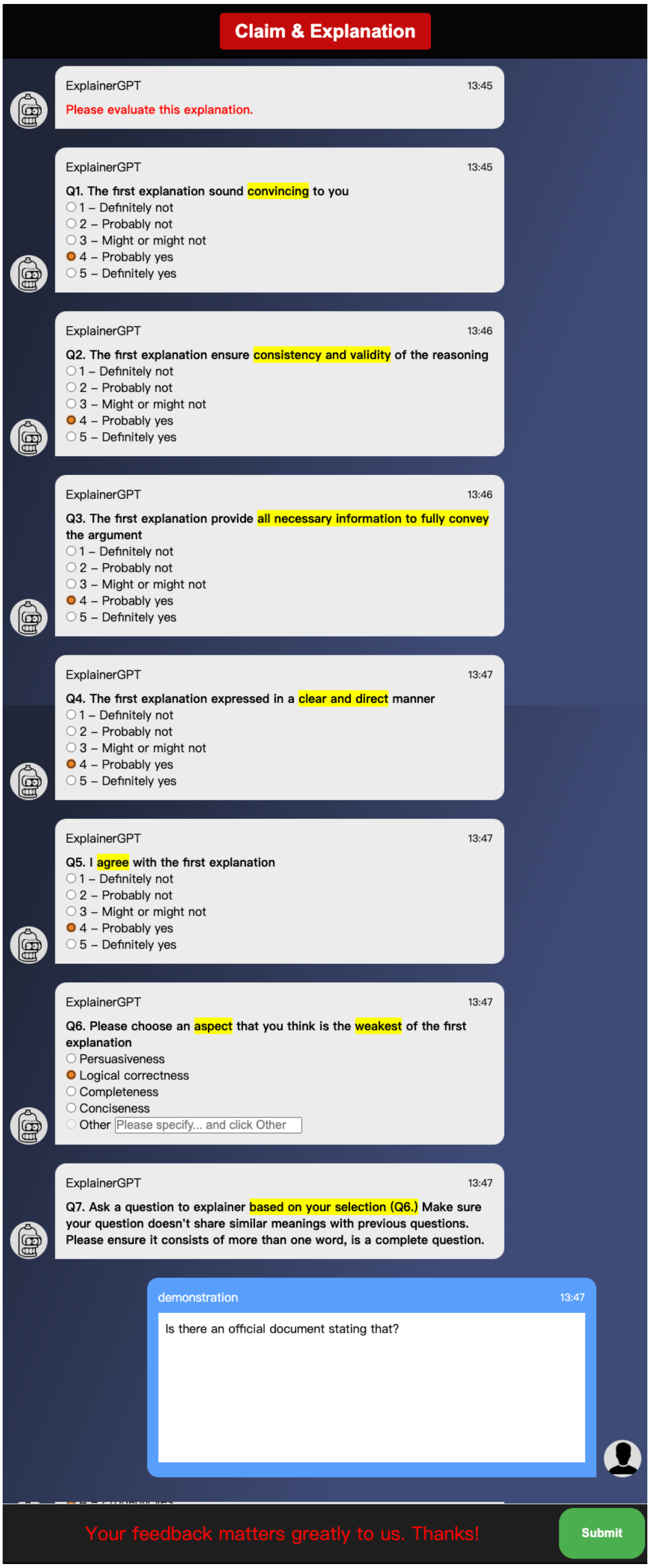
Gov. Gavin Newsom has no power or authority to extend the state of emergency in California because it expired automatically on May 6, 2020.

First Explanation

The claim is false because under the California Emergency Services Act, Governor Gavin Newsom has the authority to decide when to lift the state of emergency, which can extend beyond 60 days and even remain operational despite reopening measures, as demonstrated by the ongoing state of emergency from the 2018 Camp Fire.

(b) Reading Section Between Each Iteration

Figure 6: ChatBot Environment Interface - 1



ExplainerGPT 13:45

Please evaluate this explanation.

ExplainerGPT 13:45

Q1. The first explanation sound convincing to you

○ 1 – Definitely not
○ 2 – Probably not
○ 3 – Might or might not
● 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:46

Q2. The first explanation ensure consistency and validity of the reasoning

○ 1 – Definitely not
○ 2 – Probably not
○ 3 – Might or might not
● 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:46

Q3. The first explanation provide all necessary information to fully convey the argument

○ 1 – Definitely not
○ 2 – Probably not
○ 3 – Might or might not
● 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:47

Q4. The first explanation expressed in a clear and direct manner

○ 1 – Definitely not
○ 2 – Probably not
○ 3 – Might or might not
● 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:47

Q5. I agree with the first explanation

○ 1 – Definitely not
○ 2 – Probably not
○ 3 – Might or might not
● 4 – Probably yes
○ 5 – Definitely yes

ExplainerGPT 13:47

Q6. Please choose an aspect that you think is the weakest of the first explanation

○ Persuasiveness
● Logical correctness
○ Completeness
○ Conciseness
○ Other Please specify... and click Other

ExplainerGPT 13:47

Q7. Ask a question to explainer based on your selection (Q6.) Make sure your question doesn't share similar meanings with previous questions. Please ensure it consists of more than one word, is a complete question.

demonstration 13:47

Is there an official document stating that?

Figure 7: Chatbot Environment Interface - 2: Rating and Question Section

```
Assume you are a general newsreader.
You just saw this claim and have no preliminary knowledge.
The explanation should be clear, short, and persuasive.
---
Your task is to rate the persuasiveness, logical correctness, completeness, conciseness and agreement of the explanation based on the following criteria:
--- Rating Criteria ---
Persuasiveness Criteria:
Q1. The explanation sound convincing to you
1 - Definitely not
2 - Probably not
3 - Might or might not
4 - Probably yes
5 - Definitely yes

Logical Correctness Criteria:
Q2. The explanation ensure consistency and validity of the reasoning
1 - Definitely not
2 - Probably not
3 - Might or might not
4 - Probably yes
5 - Definitely yes

Completeness Criteria:
Q3. The explanation provide all necessary information to fully convey the argument
1 - Definitely not
2 - Probably not
3 - Might or might not
4 - Probably yes
5 - Definitely yes


Conciseness Criteria:
Q4. The explanation expressed in a clear and direct manner
1 - Definitely not
2 - Probably not
3 - Might or might not
4 - Probably yes
5 - Definitely yes

Agreement Criteria:
Q5. I agree with this explanation
1 - Definitely not
2 - Probably not
3 - Might or might not
4 - Probably yes
5 - Definitely yes

Question Criteria:
Please choose an aspect that you think is the weakest of the explanation and ask a question based on the question. Make sure your question doesn't share similar meanings with previous questions.

Q6. Choose an aspect that you think is the weakest of the explanation
- Persuasiveness
- Logical correctness
- Completeness
- Conciseness
- Other (text field)

Q7. Ask a question based on your selection. Please ensure it consists of more than one word, is a complete question

--- Rating Criteria ---
```

Figure 8: Questioner Prompt

```

Claim
---
{claim}
---
Explanation
---
{explanation}
---
This is the conversation history from you and explainer:
--- Conversation History ---
{history}
--- Conversation History ---

Please review the claim and explanation, and rate the persuasiveness, logical correctness, completeness, conciseness and agreement of the explanation accordingly.
Please independently evaluate this explanation.

Please choose an aspect that you think is the weakest of the explanation and ask a question based on the question. Make sure your question doesn't share similar meanings with previous questions.

{add1}
---
Your response should be in the format: ”Persuasiveness: <your persuasiveness rating>, Logical Correctness: <your logical correctness rating>, Completeness: <your completeness rating>, Conciseness: <your conciseness rating>, Agreement:<your agreement rating>, Aspect:<your aspect rating>, Question: <your Question>”.
{add2}
---
Response:
```

Figure 9: Questioner Prompt (Continued)

```
As a fake news debunker, you need to analyze the reason behind a claim thoroughly and create a list of evidence that supports or contradicts it.
claim: {claim},
evidence: {evidence}
To ensure that no information is missed, please generate evidence based on the claim and reason, separating each piece of evidence with a comma and with the square brackets.
The format of the evidence is as follows:
Evidence1: [evidence1], Evidence2: [evidence2], Evidence3: [evidence3] ...
```

Figure 10: Evidence List Generation Prompt