

Evaluating Counter-Argument Strategies for Logical Fallacies: An Agent-Based Analysis of Persuasiveness and Polarization

Keisuke Toyoda¹, Tomoki Fukuma¹, Koki Noda¹ Yoshiharu Ichikawa² Kyosuke Kambe² Yu Masubuchi² Hiroshi Someda² Fujio Toriumi³

¹TDAI Lab Co.,Ltd.

²NHK (Japan Broadcasting Corporation)

³The University of Tokyo

keisuke.toyoda@tdailab.com, tomoki.fukuma@tdailab.com, koki.noda@tdailab.com, ichikawa.y-gq@nhk.or.jp, kambe.k-je@nhk.or.jp, masubuchi.y-lq@nhk.or.jp, someda.h-do@nhk.or.jp, tori@sys.t.u-tokyo.ac.jp

Abstract

Misinformation on social media often involves logical fallacies, challenging traditional fact-checking methods and increasing demand for collective correction approaches like Community Notes. This study uses large language model (LLM)-based agent simulations to analyze the effectiveness of various counter-argument strategies against logical fallacies and the polarization induced by evaluator’s stance. We evaluated 10 CALSA-based rebuttal patterns against 13 common logical fallacies using independent and misinformation-aligned simulated agents. Results indicate that the “No Evidence” strategy was widely effective across various fallacies, functioning as a well-balanced rebuttal that helps curb polarization. Meanwhile, in many cases we observed a “Persuasiveness–Polarization Dilemma,” wherein strategies with higher persuasiveness can also increase polarization. Furthermore, we found that objective strategies that are less likely to trigger psychological resistance among misinformation supporters achieved both high persuasiveness and lower polarization risk. Our findings offer practical guidelines for designing effective misinformation corrections with reduced polarization risks.

Introduction

The rapid dissemination of information via social media has a detrimental aspect: the spread of “information disorder,” encompassing misinformation, disinformation, and malinformation, which undermines individuals’ ability to make rational decisions based on accurate information (Wardle and Derakhshan 2017). Among these challenges, information containing logical fallacies is often presented as personal opinions or subjective interpretations, making objective fact-checking difficult and complicating judgments about their validity. This complexity reflects the intricacies of information disorder. Therefore, it is necessary to develop strategies capable of addressing and countering the argumentative flaws inherent in such information.

To address these argumentative challenges, which traditional fact-checking struggles to handle, collective intelligence systems such as X’s Community Notes aim to identify logical fallacies and provide contextual clarifications from

diverse viewpoints. However, the effectiveness of these systems is often weakened by recipients’ pre-existing stances. Corrections touching personal beliefs can trigger motivated reasoning, such as backfire effects and confirmation bias, leading to resistance (Nyhan and Reifler 2010). Moreover, evaluations of corrections themselves tend to polarize along these stances (Taber and Lodge 2006). This resistance makes consensus on effective interventions difficult, potentially causing corrections to fail or even unintentionally deepening societal polarization. Thus, identifying rebuttal strategies that effectively address logical fallacies while minimizing polarization is an urgent research gap.

To address this research gap, we systematically analyze the effectiveness of counter-argument strategies against logical fallacies and how their evaluations vary with evaluator stance, utilizing an LLM-based agent simulation. This simulation employs diverse agent groups, constructed from actual evaluation histories of Community Notes evaluators on platform X, to realistically model the effectiveness of rebuttal strategies. Specifically, we sampled 50 Community Notes evaluators with varied evaluation patterns and modeled their cognitive biases using an LLM. Through prompt engineering, we created two agent groups: those with a bias toward agreeing with the original misinformation claims (“Aligned”) and those without such bias (“Independent”). We evaluated combinations of 13 logical fallacies (Jin et al. 2022) and 10 CALSA-based counter-arguments (Naito et al. 2024) using these agents. Our analysis addresses the following research questions: RQ1: What counter-argument strategies are most effective against specific logical fallacies? RQ2: Which logical fallacies are particularly difficult to counter? RQ3: Can persuasive counter-arguments be effective while minimizing polarization?

Our contributions are threefold: First, we identified effective counter-argument strategies for specific logical fallacy types. Second, we found that the difficulty of providing logical rebuttals and the risk of polarization often coexist, giving rise to what we term the “Persuasiveness–Polarization Dilemma.” Finally, adopting an objective approach that minimizes psychological resistance among those inclined to believe misinformation can achieve high persuasiveness with low polarization, suggesting that fact-checking methods—verifying the basis of information

and objective facts—may be especially effective. Our insights provide practical guidance for the future development of community-driven correction systems like Community Notes and highlight the importance of tools supporting balanced and less polarizing rebuttal creation.

Methodology and Experimental Design

This section outlines our methodology and experimental design for quantitatively evaluating the effectiveness of various counter-arguments (CA) against misinformation containing logical fallacies. Our methodology and experimental design are overviewed in Figure 1.

Problem Formulation

This study aims to quantitatively analyze the effectiveness of counter-argument strategies against social media misinformation containing logical fallacies, taking into account evaluators’ stances. We define logical fallacies as a set $L = \{l_1, \dots, l_{13}\}$ based on (Jin et al. 2022), and counter-argument (CA) patterns as $C = \{c_1, \dots, c_{10}\}$ following the CALSA framework (Naito et al. 2024).

Evaluation agents are categorized into two stance groups: $S = \{\text{Independent}, \text{Aligned}\}$. "Independent" agents evaluate claims without strong prior biases, while "Aligned" agents tend to evaluate misinformation favorably. Each stance group has 50 agents, totaling 100 agents ($A = A_{\text{Independent}} \cup A_{\text{Aligned}}$). Each agent $a \in A_s$ assigns an evaluation score $E(l, c, s, a) \in \{0, 1, 2\}$ for a counter-argument c against fallacy l , where scores represent: 0 = "Not helpful," 1 = "Somewhat helpful," and 2 = "Helpful," consistent with evaluation criteria used in Community Notes.

We define the following metrics to quantify effectiveness and polarization: First, the maximum potential effectiveness for fallacy type l is defined as follows, given stance s :

$$E_1(l, s) = \max_{c \in C} \mathbb{E}_{a \in A_s} [E(l, c, s, a)].$$

Second, the average effectiveness per counter-argument pattern c for stance s is:

$$E_c(s, c) = \mathbb{E}_{l \in L, a \in A_s} [E(l, c, s, a)].$$

Finally, we measure polarization between stances for fallacies and counter-arguments, defined as:

$$\begin{aligned} E_{\text{polar}}(l) &= E_1(l, \text{Independent}) - E_1(l, \text{Aligned}), \\ E_{\text{polar}}(c) &= E_c(\text{Independent}, c) - E_c(\text{Aligned}, c). \end{aligned}$$

Selection of Logical Fallacies

We use the LOGIC dataset introduced by (Jin et al. 2022), which contains 13 common types of logical fallacies, as listed below:

Faulty Generalization, False Causality, Circular Claim, Ad Populum, Ad Hominem, Deductive Fallacy, Appeal to Emotion, False Dilemma, Equivocation, Fallacy of Extension, Fallacy of Relevance, Fallacy of Credibility, Intentional Fallacy

We denote each sampled misinformation instance from fallacy type $l \in L$ as $d_{i,l}$.

Generation of Counter-Arguments (CA)

To generate counter-arguments for each misinformation instance $d_{i,l}$, we applied 10 counter-argument patterns from the CALSA framework proposed by (Naito et al. 2024), defined as:

Mitigation (Mig), Alternative (Alt), No Evidence (No Evi), Another True Cause (ATC), Missing Mechanism #1 (MM #1), Missing Mechanism #2 (MM #2), No Need to Address (NNA), Negative Effect Due to y (Neg eff), Positive Effects from Different Perspective #1 (Dif Per #1), Positive Effects from Different Perspective #2 (Dif Per #2)

Given each misinformation instance $d_{i,l}$ and CA pattern $c \in C$, we generated counter-argument text $r_{i,l,c}$ using a large language model (LLM):

$$r_{i,l,c} = \text{LLM}(d_{i,l}, c).$$

The LLM (gpt-4o) was prompted using few-shot examples from (Naito et al. 2024) to provide clear definitions and illustrative examples of each CA pattern.

Construction of Evaluation Agents

Recent studies demonstrate that LLMs can simulate individual evaluators based on prompt engineering (Park et al. 2024) or evaluation histories (Yan et al. 2025). Leveraging this capability, we construct diverse agent groups using historical Community Notes data. Specifically, we focus on evaluators who have rated at least 20 notes. Using the official Birdwatch algorithm (Allen, Martel, and Rand 2022)—a matrix factorization-based approach for helpfulness evaluation—we derive two-dimensional latent vectors representing evaluators’ rating behaviors. We then apply K-Means clustering ($k = 50$) to group similar evaluators. From each cluster, we select the evaluator whose latent vector is closest to the cluster centroid as representative, using their historical evaluations as prompt input for constructing evaluation agents. The most important reason for using LLM-based agents to simulate evaluators is that it enables us to realistically replicate diverse evaluation tendencies and biases derived from actual historical evaluator data, providing a controlled yet authentic assessment environment. Stance is controlled by using stance-specific prompt designs based on (Piao et al. 2025): the "Aligned" prompt induces an agreement bias toward misinformation, while the "Independent" prompt remains independent. Consequently, we simulate two stance-based agent groups (A_{Aligned} and $A_{\text{Independent}}$, each with $|A_s| = 50$), enabling agents to produce evaluation scores $E(l, c, s, a)$.

Results and Discussion

This section analyzes the results of the LLM-agent simulation and addresses the three research questions (RQs) presented. First, it clarifies what counter-argument strategies are most effective against specific logical fallacies (RQ1). Next, it identifies which logical fallacies are particularly difficult to counter (RQ2). Finally, it examines whether it is possible to reconcile argument persuasiveness with the suppression of evaluative polarization (RQ3).

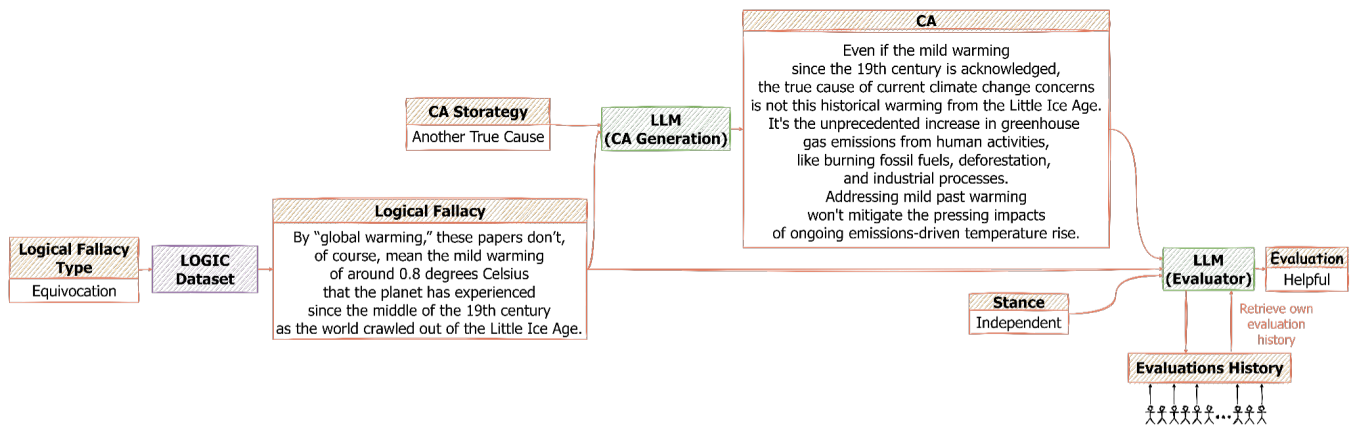


Figure 1: Overview of our Our Methodology and Experimental Design

Fallacy Type	Best Counter Argument (Independent)	Best Counter Argument (Aligned)
ad hominem	Another True Cause	Another True Cause
ad populum	No Evidence	No Evidence
appeal to emotion	Another True Cause	Another True Cause
circular reasoning	Another True Cause	No Evidence
equivocation	Missing Mechanism 1	Missing Mechanism 1
fallacy of credibility	Another True Cause	No Evidence
fallacy of extension	No Evidence	No Evidence
fallacy of logic	No Evidence	No Evidence
fallacy of relevance	No Evidence	No Evidence
false causality	Another True Cause	No Evidence
false dilemma	Missing Mechanism 1	No Evidence
faulty generalization	Mitigation	Another True Cause
intentional	Negative Effect due to y	Negative Effect due to y

Table 1: Most Effective Counter-Arguments for Each Fallacy Type (by Evaluator Type)

RQ1: Effective Counter-Argument Strategies for Specific Logical Fallacies

To address RQ1, we analyzed the effectiveness of counter-argument strategies against 13 logical fallacies, considering both Independent and Aligned evaluator stances. Three results are summarized in Table 1.

Certain counter-argument strategies demonstrated broad effectiveness across different fallacies and evaluator stances. "Another True Cause" (ATC) and "No Evidence" (No Evi) were notably effective. ATC performed well against fallacies like 'ad hominem' and 'appeal to emotion' for both stances, suggesting it effectively undermines the claim's logical foundation by offering alternative explanations. Independent evaluators also favored ATC for 'circular reasoning', 'fallacy of credibility', and 'false causality'. No Evi was consistently effective regardless of stance against fallacies lacking substantial evidence, such as 'ad populum', 'fallacy of extension', 'fallacy of logic', and 'fallacy of relevance', highlighting the power of pointing out the absence of objective evidence. Additionally, "Missing Mechanism 1"

(MM #1) was identified as the best strategy for 'equivocation', and "Negative Effect due to y" (Neg eff) for 'intentional' fallacies, for both evaluator groups.

Conversely, the optimal strategy for several fallacies depended significantly on the evaluator's stance. For 'circular reasoning', 'fallacy of credibility', 'false causality', and 'false dilemma', Independent evaluators favored strategies offering alternative explanations or mechanisms (ATC or MM #1), while Aligned evaluators rated No Evi highest. This suggests Aligned evaluators might resist alternative explanations conflicting with their beliefs, potentially preferring the identification of formal flaws, possibly influenced by confirmation bias or motivated reasoning. Furthermore, for 'faulty generalization', Independent evaluators preferred "Mitigation," whereas Aligned evaluators selected ATC. This indicates that for Aligned individuals, presenting alternative causes might be more persuasive than merely pointing out exceptions for this type of fallacy.

In conclusion, RQ2 strongly suggests a "Persuasiveness-Polarization Dilemma" where, for many fallacies, particu-

larly those in Quadrant 1 like ad hominem, ease of logical refutation (high $E_l(l, \text{Independent})$) coexists with a high risk of polarization (high $E_{\text{polar}}(l)$). This differs from the challenge posed by Quadrant 2 fallacies like appeal to emotion, which are difficult to refute logically and also trigger polarization. This finding underscores that combating fallacies requires addressing both logical and socio-psychological dimensions, necessitating strategies tailored to the nature of the difficulty. For systems like Community Notes, advanced support for crafting rebuttals that mitigate polarization while maintaining persuasiveness, especially for Quadrant 1 and 2 fallacies, will be key. Understanding this multi-faceted difficulty is crucial for developing effective interventions against information disorder.

RQ2: Identifying Difficult-to-Counter Logical Fallacies

To address RQ2: "Which logical fallacies are particularly difficult to counter?", we analyze the scatter plot in Figure 2a. The horizontal axis represents the potential persuasiveness of counter-arguments as perceived by Independent evaluators ($E_l(l, \text{Independent})$), while the vertical axis represents the polarization between stances ($E_{\text{polar}}(l)$). These axes illustrate two facets of the "difficulty" in countering fallacies: difficulty in achieving persuasion (low $E_l(l, \text{Independent})$) and the risk of inducing backlash or polarization (high $E_{\text{polar}}(l)$). We examine these dimensions below.

First, examining the horizontal axis ($E_l(l, \text{Independent})$) reveals the difficulty associated with achieving persuasive counter-arguments among independent evaluators. Fallacies positioned on the left side of the plot, such as appeal to emotion and fallacy of credibility, exhibit low $E_l(l, \text{Independent})$ scores. These fallacies are thought to rely heavily on non-logical or heuristic cues like emotion or source authority, making it relatively difficult to construct effective, logically-based counter-arguments that achieve high persuasiveness. In contrast, many fallacies on the right side (e.g., ad hominem, false dilemma) tend to show higher $E_l(l, \text{Independent})$ scores. For these fallacies, the flaws in logical structure or reasoning are considered relatively clear (from a neutral perspective), suggesting that constructing a counter-argument via objective refutation or logical pointing is comparatively easier.

Next, the vertical axis ($E_{\text{polar}}(l)$) indicates the risk of backlash and polarization induced by counter-arguments. Many fallacies located in the upper half of the plot, including appeal to emotion, fallacy of credibility, ad hominem, and false dilemma, exhibit high $E_{\text{polar}}(l)$ scores. It is considered that these fallacies, or counter-arguments against them, strongly engage recipients' emotions, values, or social identities. Consequently, even logically sound counter-arguments are likely to be perceived as attacks on personal beliefs or group affiliations, leading to strong resistance and significant divergence in evaluation between stances (polarization).

Viewed through these two axes, the difficulty of countering fallacies is particularly evident for those in the upper half of the plot. Fallacies in Quadrant 2 (top-left), such

as appeal to emotion, are characterized by high difficulty on both axes: constructing persuasive counter-arguments is challenging (low $E_l(l, \text{Independent})$), and they tend to provoke strong backlash (high $E_{\text{polar}}(l)$). This is likely because they rely on non-logical elements and engage identity. Fallacies concentrated in Quadrant 1 (top-right), like ad hominem, while having relatively clear logical flaws (high $E_l(l, \text{Independent})$), pose a significant challenge due to the extremely high risk of polarization ($E_{\text{polar}}(l)$), as they often stimulate emotions and identity. Therefore, the answer to RQ2 depends on the dimension of difficulty considered, but fallacies in Quadrants 1 and 2 warrant particular attention. Among these, Quadrant 2 fallacies, exhibiting high difficulty on both dimensions, are potentially the most challenging to address effectively. Developing effective counter-strategies requires not only focusing on logical correctness but also carefully considering the socio-psychological impact ($E_{\text{polar}}(l)$) that counter-arguments can trigger.

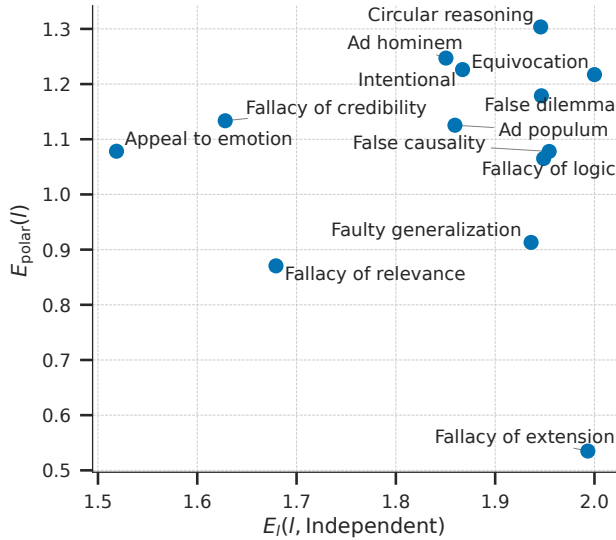
RQ3: Can Persuasive Counter-Arguments Be Effective While Minimizing Polarization?

To answer RQ3, we analyzed the relationship between the effectiveness and polarization associated with each counter-argument strategy. Figure 2b plots the average effectiveness for the Independent stance, $E_c(\text{Independent}, c)$, on the horizontal axis, and the polarization index, $E_{\text{polar}}(c) = E_c(\text{Independent}, c) - E_c(\text{Aligned}, c)$, on the vertical axis for each counter-argument strategy c .

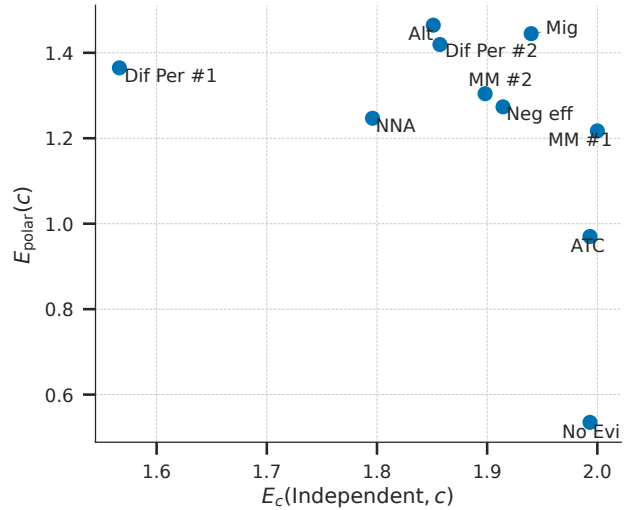
Relatively moderate counter-arguments, such as "No Evidence" and "Another True Cause," exhibited high persuasiveness from the Independent stance while maintaining low polarization scores. These strategies tend to emphasize objective aspects, like the presence or absence of evidence or alternative causes, rather than directly negating or attacking the recipient's beliefs. This approach appears less likely to excessively provoke cognitive resistance or emotional defensiveness in the recipient. Consequently, even individuals aligned with the misinformation may find such rebuttals more palatable, as they feel less personally invalidated, allowing these strategies to achieve a certain level of effectiveness without exacerbating polarization.

Conversely, among counter-arguments involving more direct critique or stronger implications of refutation, such as "Missing Mechanism #1" or "No Need To Address," we observed a tendency where some strategies were highly persuasive to the Independent stance but prone to generating significant psychological resistance among Aligned evaluators, thus leading to polarization. Such rebuttals are more likely to induce cognitive dissonance among those aligned with the misinformation, carrying a risk of increasing polarization. This phenomenon aligns with motivated reasoning and cognitive dissonance theories, which suggest that defensive biases arise when information implies skepticism or negation towards one's existing belief system (Taber and Lodge 2006; Nyhan and Reifler 2010).

In summary, this analysis reveals that the impact on persuasiveness and polarization varies significantly across counter-argument strategies. Specifically, relatively moderate rebuttals focusing on objective facts show potential for



(a)



(b)

Figure 2: (a) Scatter plot showing counter-argument persuasiveness for Independent evaluators ($E_l(l, \text{Independent})$, x-axis) versus polarization ($E_{\text{polar}}(l)$, y-axis) for each logical fallacy type. (b) Scatter plot showing average effectiveness for Independent evaluators ($E_c(\text{Independent}, c)$, x-axis) versus polarization ($E_{\text{polar}}(c)$, y-axis) for each counter-argument strategy.

maintaining persuasiveness from an independent perspective while mitigating excessive psychological resistance among those aligned with the misinformation, thereby suppressing polarization. On the other hand, more direct refutations, while potentially effective, carry the risk of fostering polarization. These findings suggest that approaches emphasizing the verification of evidence and objective facts—akin to fact-checking—rather than simply denying the opponent’s claim, may be promising for countering misinformation effectively while reducing the risk of deepening social divisions.

Conclusion

This study analyzed the relationship between the effectiveness of counter-argument strategies against logical fallacies in misinformation and the polarization of evaluations based on evaluator stance, utilizing LLM agent simulations grounded in actual Community Notes user evaluation histories. Our primary objective was to identify counter-argument strategies capable of maintaining persuasiveness while mitigating the risk of exacerbating social divisions, addressing this challenge through three related research questions (RQs).

Our analysis revealed, first, that the optimal counter-argument strategy is contingent upon the specific type of logical fallacy being addressed (RQ1). Strategies such as “Another True Cause” and “No Evidence” demonstrated broad effectiveness, particularly against fallacies rooted in faulty reasoning or lacking substantiation. Second, we identified two key facets defining the difficulty of countering fallacies (RQ2): some fallacies are inherently difficult to refute persuasively even for independent evaluators (e.g., “appeal to emotion”), while others are highly prone to in-

ducing polarization despite being relatively easy to counter logically, often because they engage emotions or identity (e.g., “ad hominem”). This highlights that merely addressing logical flaws is insufficient and suggests the existence of a “Persuasiveness-Polarization Dilemma.”

Third, and critically, our analysis of the counter-argument strategies themselves showed that high persuasiveness and low polarization are not necessarily mutually exclusive (RQ3). We found that relatively moderate strategies focusing on objective facts (No Evi) or presenting alternative causes (ATC) are less likely to excessively provoke cognitive resistance or emotional defensiveness in recipients. Consequently, they show potential to maintain high effectiveness among Independent evaluators while also garnering reasonable evaluations from those aligned with the misinformation, thereby suppressing polarization. This suggests that approaches emphasizing objective verification processes (akin to fact-checking) over direct negation of claims hold promise for addressing misinformation effectively while reducing the risk of deepening societal divides.

The main contributions of this research include: (1) providing empirical insights into effective counter-argument strategies tailored to specific logical fallacies; (2) offering a multi-faceted characterization of the difficulty in countering fallacies along the axes of persuasiveness and polarization potential; (3) identifying promising counter-argument patterns that can mitigate polarization, along with practical guidance for platforms like Community Notes; and (4) demonstrating the validity of using LLM agent simulations based on real user data as a methodological approach to reproduce nuanced evaluation tendencies.

At the same time, several limitations must be acknowl-

edged. Foremost is the reliance on LLM simulations, which may not fully capture the complexity of human psychological responses. Generalizability of the findings also requires caution, given their basis on a specific dataset (LOGIC) and evaluator pool (Community Notes users). Furthermore, the quality and naturalness of LLM-generated counter-arguments, and the simplification inherent in the binary classification of evaluator stances, represent additional limitations.

Future research should prioritize validating these simulation findings through experiments with diverse human participants. Detailed analysis of how specific wording and framing choices within counter-arguments affect effectiveness and polarization is crucial. Developing and evaluating practical platform tools designed to assist users in crafting less polarizing rebuttals based on these findings would be a valuable extension. Constructing more sophisticated, potentially multi-dimensional models of user stance, and investigating the long-term effects of repeated exposure to different counter-arguments, also constitute important avenues for future inquiry.

References

- Allen, J.; Martel, C.; and Rand, D. G. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.
- Jin, Z.; Lalwani, A.; Vaidhya, T.; Shen, X.; Ding, Y.; Lyu, Z.; Sachan, M.; Mihalcea, R.; and Schoelkopf, B. 2022. Logical Fallacy Detection. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 7180–7198. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Naito, S.; Wang, W.; Reiser, P.; Inoue, N.; Guerraoui, C.; Yamaguchi, K.; Choi, J.; Robbani, I.; Pothong, S.; and Inui, K. 2024. Designing Logic Pattern Templates for Counter-Argument Logical Structure Analysis. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11313–11331. Miami, Florida, USA: Association for Computational Linguistics.
- Nyhan, B.; and Reifler, J. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, 32(2): 303–330.
- Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109.
- Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J. Y.; Zhou, D.; Gao, C.; Xu, F.; Zhang, F.; Rong, K.; Su, J.; and Li, Y. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691.
- Taber, C. S.; and Lodge, M. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3): 755–769.
- Wardle, C.; and Derakhshan, H. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy-Making.
- Yan, Y.; Shang, Y.; Zeng, Q.; Li, Y.; Zhao, K.; Zheng, Z.; Ning, X.; Wu, T.; Yan, S.; Wang, Y.; Xu, F.; and Li, Y. 2025. AgentSociety Challenge: Designing LLM Agents for User Modeling and Recommendation on Web Platforms. arXiv:2502.18754.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, the research aims to understand and mitigate issues related to misinformation and polarization, which is beneficial.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes, the abstract and introduction clearly state the use of LLM-based agent simulations to analyze counter-argument strategies against logical fallacies, their effectiveness, and polarization, which aligns with the contributions detailed in the paper.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, the "Methodology and Experimental Design" section details the LLM-based agent simulation, selection of fallacies and counter-arguments, and construction of evaluation agents, justifying its appropriateness for addressing the research questions.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, the limitations section acknowledges that "Generalizability of the findings also requires caution, given their basis on a specific dataset (LOGIC) and evaluator pool (Community Notes users)."**
 - (e) Did you describe the limitations of your work? **Yes, the limitations are discussed in the Conclusion section, mentioning reliance on LLM simulations, generalizability, quality of LLM-generated counter-arguments, and binary classification of evaluator stances.**
 - (f) Did you discuss any potential negative societal impacts of your work? **No, the paper does not explicitly discuss potential negative societal impacts. The focus is on mitigating negative impacts of misinformation.**
 - (g) Did you discuss any potential misuse of your work? **No, the paper does not explicitly discuss potential misuse of the findings (e.g., crafting more sophisticated misinformation).**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **No, while the model used (gpt-4o) is named, and the methodology for agent construction implies some level of abstraction from raw user data (using latent vectors and cluster centroids), specific steps for data anonymization, responsible release, access control, or detailed reproducibility instructions (beyond methodology description) are not explicitly described as mitigation measures.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes. Because only public data are used, the study is IRB-exempt, yet we follow the ICWSM ethics checklist.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **N/A. The study primarily presents empirical findings from simulations rather than formal theoretical results with stated mathematical assumptions. A core methodological assumption is that LLM agents can simulate human evaluators, which is mentioned.**
 - (b) Have you provided justifications for all theoretical results? **N/A. As above, the paper focuses on empirical findings from simulations.**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **No, the paper discusses theories that complement the findings (e.g., motivated reasoning, cognitive dissonance) but does not explicitly discuss "competing" hypotheses in detail.**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes, for example, when discussing why Aligned evaluators preferred "No Evidence" for certain fallacies, the paper suggests it might be "possibly influenced by confirmation bias or motivated reasoning," implying consideration of different underlying mechanisms.**
 - (e) Did you address potential biases or limitations in your theoretical framework? **N/A. The paper does not propose a specific "theoretical framework" but discusses limitations of its methodological approach in the Conclusion.**
 - (f) Have you related your theoretical results to the existing literature in social science? **Yes, the Introduction and Discussion sections relate the findings to concepts like motivated reasoning, confirmation bias, and cognitive dissonance from social science literature.**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, the Conclusion discusses practical guidance for platforms like Community Notes and suggests future research directions, such as validation with human participants.**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **N/A. The paper does not include theoretical proofs.**
 - (b) Did you include complete proofs of all theoretical results? **N/A. The paper does not include theoretical proofs.**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **No, the paper does not mention the public availability of code, data, or detailed reproduction instructions, though there is a commented-out section for such links.**

- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **No, while the LLM model (gpt-4o) and some parameters (K-Means k=50) are specified, and the prompting strategy is described, comprehensive training details, hyperparameter choices, or data splits beyond sampling from Community Notes data are not provided.**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **No, the figures and tables present mean scores or categorizations without error bars or other measures of variance from multiple runs.**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **No, the paper does not describe the computational resources used.**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes, the paper states: "The most important reason for using LLM-based agents to simulate evaluators is that it enables us to realistically replicate diverse evaluation tendencies and biases derived from actual historical evaluator data, providing a controlled yet authentic assessment environment." The defined metrics also align with the research questions.**
 - (f) Do you discuss what is the cost of misclassification and fault (in)tolerance? **No, the paper does not explicitly frame "cost of misclassification" in these terms, although polarization is discussed as an undesirable outcome, implying a cost.**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **Yes, citations are provided for the LOGIC dataset (Jin et al. 2022), CALSA framework (Naito et al. 2024), Birdwatch algorithm (Allen, Martel, and Rand 2022), and relevant LLM agent simulation studies. The LLM model (gpt-4o) is also named.**
 - (b) Did you mention the license of the assets? **No, the licenses of the existing assets (datasets, frameworks, models) are not mentioned.**
 - (c) Did you include any new assets in the supplemental material or as a URL? **No, the paper does not mention the release of new assets.**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **No, the paper uses historical Community Notes data but does not discuss how consent was obtained from the original evaluators.**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **No, the paper does not discuss whether the Community Notes data contains PII or offensive content, or specific steps taken for de-identification beyond the methodological description of using latent vectors and cluster centroids.**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **N/A. The paper does not state an intention to release new datasets.**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **N/A. The paper does not state an intention to release new datasets.**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **N/A. The study uses LLM agents based on existing historical data, not newly recruited human subjects or crowdsourcing.**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **N/A. As above.**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **N/A. As above.**
 - (d) Did you discuss how data is stored, shared, and de-identified? **N/A. As above. While existing data is used, this question pertains to newly conducted human subject research data handling.**