Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Beyond AI: Exploring Retained Improvement of Reasoning Ability in LLM-Assisted Human Credibility Assessment on Social Media

Nianhua Liu, Mengyi Wei, Yuanqi Wang, Liqiu Meng

Technical University of Munich
{nianhua.liu, mengyi.wei, go73tes, liqiu.meng}@tum.de

Abstract

As Large Language Models (LLM) become increasingly deployed to social media content evaluation, understanding their impact on human independent judgment is critical. This study investigates whether LLM-based assistance can lead to retained improvements in credibility assessment of climate-related social media posts. We designed a two-step assistant combining sidebar advisors with optional chatbot interaction and conducted a three-phase user study ($N = 31$), including a baseline phase without LLM, an LLM-supported phase, and an evaluation phase without LLMs. Results show that users became significantly more skeptical when supported by LLM and retained cautious evaluative stance even after support was withdrawn. Moreover, the content analysis indicates that LLM assistance led to a retained improvement in participants' reasoning ability. Participants reported moderate cognitive load and low frustration, while expressing strong gains in confidence, reflection, and critical thinking. These findings suggest that LLM assistants can produce a partially retained improvement of reasoning ability and shift toward more cautious credibility assessments.

Introduction

Social media has emerged as a primary source of information for the public, particularly regarding complex scientific issues such as climate change (Schäfer and Painter 2021; Veltri and Atanasova 2017). While social media plays a crucial role in spreading reliable information, it also serves as a platform for amplifying non-scientific claims and misinformation of climate deniers (Petersen, Vincent, and Westerling 2019; Supran and Oreskes 2017). Unlike traditional scientific publications, social media content is often not peer-reviewed by external evaluation to ensure accuracy (Ladle, Jepson, and Whittaker 2005). This absence of oversight makes it difficult for audiences to evaluate the credibility of climate-related news (Liu et al. 2025).

Research shows that internet users rarely dedicate their full cognitive effort to evaluating information (Metzger and Flanagin 2013). Instead, they often manage the perceived costs of information overload by employing strategies that minimize cognitive effort and time (Lang 2000). Consequently, AI tools LLM assistants are increasingly utilized to support users in assessing content credibility (Jahanbakhsh et al. 2023; Amershi et al. 2019). This reliance on AI is particularly pronounced in social media environments, which demand rapid information processing (Haque, Islam, and Mikalef 2024; Swaroop et al. 2024). In complex domains requiring substantial background knowledge like climate change, people relied on AI even more (Vasconcelos et al. 2023). Previous studies have shown that AI explanations can enhance users' cognitive level and the accuracy with which they evaluate online information (Gong et al. 2025; Le and Wartschinski 2018; Buçinca, Malaya, and Gajos 2021).

However, the sustainability of this cognitive enhancement remains an open question. Specifically, it is unclear whether users can effectively retain what they learn from AI for future independent applications. Recent studies have started to examine the lasting reasoning ability of AI assistance (Vicente and Matute 2023). Once AI support is withdrawn, users may inherit biased reasoning patterns acquired during their early interaction with the LLMs (Spitzer et al. 2025). The retained effect of reasoning ability remains unexplored in the context of social media. Given that social media users typically operate under low cognitive engagement, it remains to be seen whether they can maintain LLM-assisted reasoning or stance once the assistant is no longer present.

This user study is conducted in a social media simulator. We investigate whether interacting with LLM assistants can lead to retained improvement of users' reasoning ability or a cautious evaluative stance to assess the credibility of climate-related content independently. In addition, we used questionnaires to examine cognitive effort and users' experience in using LLMs during these tasks.

Study Design and Method

Assistant Design and Stimuli

We designed an LLM-based social media simulator that operates in two steps (Figure 1). First, participants received passive credibility suggestions via embedded sidebar advisors. These advisors were integrated into the interface, presenting pre-generated credibility ratings and explanations for each post. Second, users could click buttons to open web-based chatbots and seek further explanations through conversation. Both formats were powered by ChatGPT 5 and Gemini 2.5 Pro, adapting from prior research (Appendix 1).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

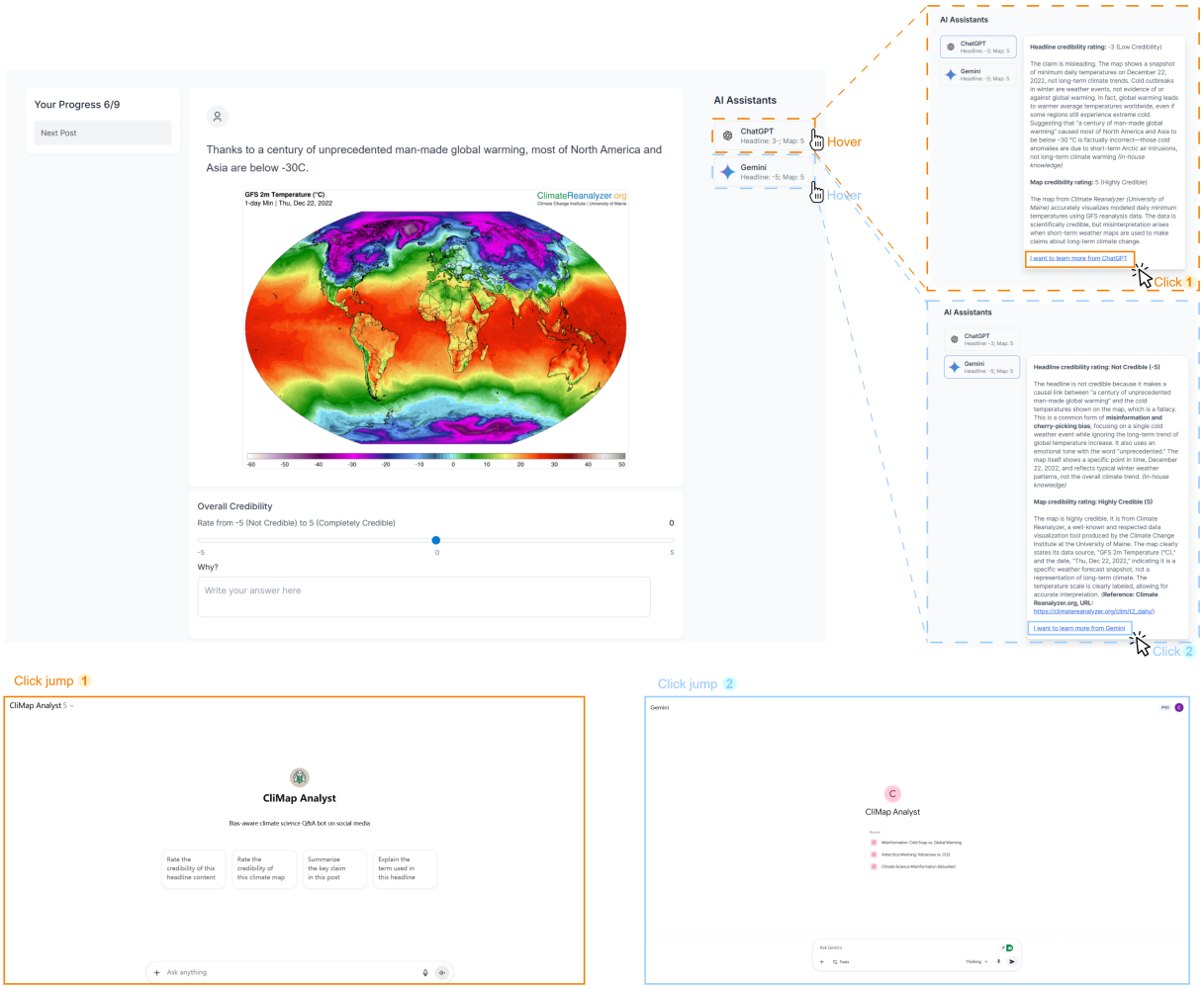

Figure 1: Experimental interface during the LLM-supported phase. Hovering over a button revealed sidebar credibility suggestions and explanations (dashed box). Clicking opened a web-based chatbot for deeper interaction (solid box).

The study utilized nine real-world climate-related social media posts from an open data source with a Creative Commons license (Effrosynidis, Sylaios, and Arampatzis 2022; Effrosynidis et al. 2022), each combining a scientific map from credible institutions with a misleading headline.

Participants and Procedure

We recruited 31 participants (13 female, 18 male) aged 22–36 ($M = 26.8$). All participants completed the same sequence of tasks and were exposed to all experimental conditions, and each received compensation of 10 EUR upon completion of the study. This study was approved by the Ethics Committee of the Technical University of Munich (726/21 S).

The experiment consisted of three phases, each featuring three posts. To mitigate order effects, the order of posts within each phase was randomized for each participant. In the first phase, users assessed credibility without LLM assistance to establish a baseline. In the second phase (the LLM-supported phase), participants viewed posts with access to both sidebar advisors and web-based chatbots. In the third phase, participants assessed a new set of posts without LLM support to examine the potential for retained improvement in reasoning ability or stance.

For each post, participants rated overall credibility on a scale ranging from -5 (Not Credible) to $+5$ (Completely Credible) while viewing the social media post and providing brief written explanations. A post-task survey followed, which included the NASA-TLX (7-point scale) to assess cognitive load, and a separate 7-point Likert scale questionnaire to evaluate perceived tool usefulness, reflection, and confidence within the context of the social media simulator.

Results and Discussion

Performance Development Across Phases

During the first phase, the mean rating was -1.20 ($SD = 3.22$), indicating mild disagreement. In the second phase, the mean decreased to -2.29 ($SD = 2.73$), reflecting stronger skepticism. In the third phase, mean credibility ratings in-

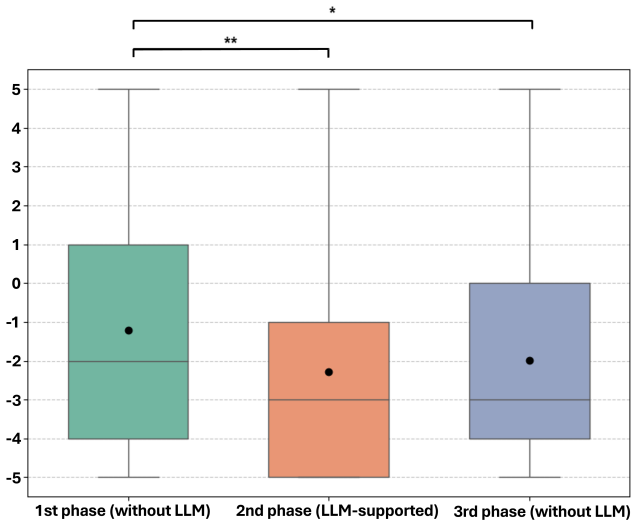

Figure 2: Changes in credibility ratings across study phases. A highly significant drop occurred from the first to the second phase ($p < 0.01$, **), and a significant difference remained between the first and third phases ($p < 0.05$, *).

creased slightly to -1.98 ($SD = 2.71$), suggesting partial relaxation of disagreement after AI assistance was removed.

Within-user standard deviations followed a similar trend. Participants exhibited the highest variability during the first phase ($M = 2.77, SD = 1.42$; range = 0.58–5.77), which dropped significantly during the second phase ($M = 1.57, SD = 1.23$; range = 0–4.62) and decreased slightly further in the third phase ($M = 1.36, SD = 1.21$; range = 0–5.29). These results suggest increased consistency in credibility ratings under LLM support, with this consistency largely sustained even after its removal.

Wilcoxon signed-rank tests were conducted to assess whether differences across phases were statistically significant (see Figure 2). A highly significant decrease in credibility was observed from the first to the second phase ($p < 0.01$, mean $\Delta = -1.09$), reflecting a marked shift in participants' critical stance when assisted by LLM. This suggests that users became more skeptical of climate posts when LLM assistance was available.

The subsequent increase from the second to the third phase was not statistically significant ($p = 0.18$, mean $\Delta = +0.31$), indicating that while some trust was regained without LLM support, users did not fully return to their baseline levels. This points to a partial retained improvement of LLM-influenced reasoning strategies. Comparing the first and third phases showed a significant overall decrease ($p < 0.05$, mean $\Delta = -0.77$), indicating that even after LLM support was removed, participants' independent evaluation remained significantly altered compared to their initial assessments. This finding suggests that exposure to LLM assistance helped participants persisting cautious evaluative stance partially beyond the supported interaction.

Frequency counts of individual credibility scores (-5 to $+5$) further support this pattern. High credibility scores (e.g., $+4, +5$) decreased after LLM assistance, while strong disagreement scores (e.g., $-4, -5$) remained consistently frequent throughout.

The significant drop during the LLM-supported phase highlights the tool's immediate corrective power, while the modest rebound in the third phase indicates that some of the learned skepticism endured.

Content Analysis of Reasoning

To identify how participants reasoned about the credibility of climate-related social media posts, we analyzed the textual explanations accompanying their credibility ratings and conducted a content analysis. Based on the purpose of our study and previous literature (Liu et al. 2025), we derived four analytically meaningful reasoning dimensions: *Prior Knowledge*, *Evidence Justification*, *Emotional Expression*, and *Internal Logic* (definitions are provided in Appendix 2). All reasoning statements were independently coded by two coders across the three experimental phases. Inter-coder reliability was assessed using Krippendorff's α.

Figure 3 summarizes the distribution of reasoning dimensions across the three experimental phases. *Prior Knowledge* had the highest percentage in the first phase, decreased during the LLM-supported phase, and showed a slight rebound in the third phase. This pattern suggests that LLM assistance reduced participants' dependence on prior knowledge. Although it slightly returned once the assistance was removed, the overall level remained lower than in the baseline, indicating a partial and lasting adjustment in how participants calibrated their judgments.

Evidence Justification increased substantially during the LLM-supported phase and remained elevated in the third phase compared to the baseline. This sustained increase indicates that participants learned to anchor their credibility assessments more strongly in external evidence and source cues. The persistence of this pattern after the removal of LLM support suggests a retained improvement of reasoning ability.

Emotional Expression and *Internal Logic* showed similar trends across the three phases. Both increased in the LLM-supported phase and decreased in the third phase.

Overall, these patterns indicate that LLM assistance not only altered participants' immediate reasoning strategies but also contributed to more reflective and structured credibility assessment behavior beyond the period of direct support. Specifically, LLM assistance encouraged participants to more actively evaluate the consistency between textual claims and visual evidence, as well as to attend to the emotional expression within textual claims. The partial persistence observed in the final phase suggests that participants retained some of these analytical strategies even when LLM assistance was no longer available.

Cognitive Load and User Experience

On average, participants reported moderate mental demand when using the AI tools to evaluate social media posts ($M = 3.90, SD = 1.49$), and a similar level of perceived effort ($M = 3.55, SD = 1.59$). Frustration scores were lower ($M = 2.97, SD = 1.45$), suggesting the interaction

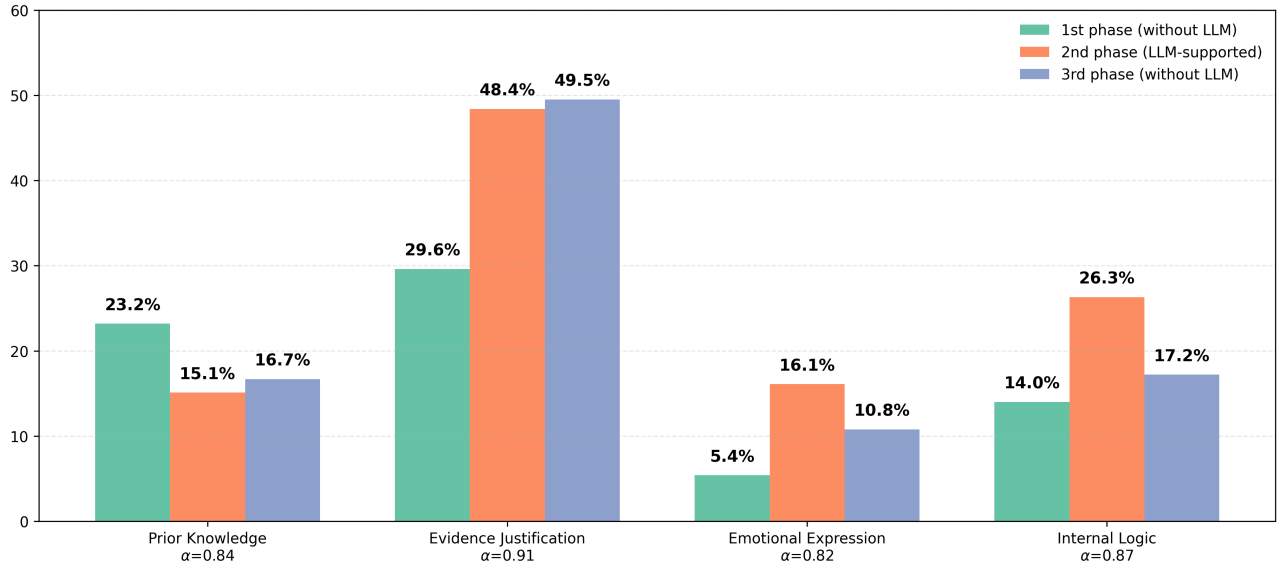

Figure 3: Distribution of participants' reasoning dimensions across the three experimental phases.

was manageable and not emotionally taxing. Notably, participants gave relatively high ratings for their own performance in identifying misinformation with chatbot assistance ($M = 4.81, SD = 1.28$), indicating that the tools were perceived as supportive rather than confusing or overbearing.

Participants consistently expressed favorable views on the reflective and educational value of the LLM support. They agreed that the chatbots helped sharpen their awareness of bias in news content ($M = 5.18, SD = 1.09$) and encouraged critical thinking ($M = 5.26, SD = 1.52$). The tools were also viewed as helpful for understanding the social media posts ($M = 5.65, SD = 1.13$) and identifying media bias ($M = 5.21, SD = 1.34$).

Importantly, participants reported that the tools prompted them to reflect on their own reasoning processes on social media ($M = 5.40, SD = 1.27$) and helped them gain new knowledge or perspectives about the climate-related social media posts they encountered ($M = 5.47, SD = 1.36$). These reflections align with the observed retained improvement from the third phase, suggesting that users internalized reasoning strategies during LLM interaction. Confidence also improved, with participants feeling more certain in their evaluations when supported by LLM assistants ($M = 5.55, SD = 1.36$). Notably, participants' perception of the LLM assistance as "thinking" for them remained moderate ($M = 3.71, SD = 1.59$), suggesting that users remained cognitively engaged during the task.

Conclusion and Future Work

Our study demonstrates that participants became significantly more skeptical during the AI-supported phase, and this cautious evaluative stance partially persisted after the LLM support was removed. Moreover, the content analysis indicates that LLM assistance led to a retained improvement in participants' reasoning ability, reflected in a reduced reliance on prior knowledge and an increased use of evidence justification and internal logic when evaluating credibility. The observed change may partly because users shifted attention to accuracy and evidence, which can reduce belief in misinformation even without improving their reasoning skills (Pennycook et al. 2020). We therefore avoid treating increased skepticism as improved calibration (Wood and Porter 2019; Ma et al. 2023). Additionally, participants reported moderate cognitive effort and low frustration, alongside strong perceived benefits in reflection, learning, and confidence, indicating that the LLM assistance supported reasoning improvement without imposing excessive cognitive load.

Participants were primarily recruited from universities, which may introduce population-specific biases. We acknowledge that the study has a relatively small sample size and a limited number of stimuli, which may constrain the generalizability of the results. We will extend the study to more diverse participant groups and include more stimuli in future work. Besides, we also acknowledge the risk that LLM-based tools could be misused by climate denialists to generate or amplify climate misinformation, which calls for a cautious and critical approach to deploying LLM assistance on social media platforms.

Acknowledgments

We would like to thank all the participants in the experiment for their contributions to the study. This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) (DFG) under the project *Guided Unlearning of Cognitive Pitfalls in Georeferenced Social Sensing* [grant number 491363672].

References

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; Teevan, J.; Kikin-Gil, R.; and Horvitz, E. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13. ACM.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Effrosynidis, D.; Karasakalidis, A. I.; Sylaios, G.; and Arampatzis, A. 2022. The climate change Twitter dataset. *Expert Systems with Applications*, 204: 117541.

Effrosynidis, D.; Sylaios, G.; and Arampatzis, A. 2022. Exploring climate change on Twitter using seven aspects: Stance, sentiment, aggressiveness, temperature, gender, topics, and disasters. *PLOS ONE*, 17(9): 1–19.

Gong, Y.; Liu, Y.; Shang, L.; Wei, N.; and Wang, D. 2025. Designing Effective AI Explanations for Misinformation Detection: A Comparative Study of Content, Social, and Combined Explanations. *Proceedings of the ACM on Human-Computer Interaction*, 9(CSCW396): 1–37.

Haque, A. B.; Islam, N.; and Mikalef, P. 2024. To Explain or Not To Explain: An Empirical Investigation of AI-based Recommendations on Social Media Platforms. *Electronic Markets*, 35(1): 2.

Jahanbakhsh, F.; Katsis, Y.; Wang, D.; Popa, L.; and Muller, M. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–27. New York, NY, USA.

Ladle, R. J.; Jepson, P.; and Whittaker, R. J. 2005. Scientists and the media: the struggle for legitimacy in climate change and conservation science. *Interdisciplinary Science Reviews*, 30(3): 231–240.

Lang, A. 2000. The Limited Capacity Model of Mediated Message Processing. *Journal of Communication*, 50(1): 46–70.

Le, N.-T.; and Wartschinski, L. 2018. A Cognitive Assistant for improving human reasoning skills. *International Journal of Human-Computer Studies*, 117: 45–54.

Liu, N.; Wei, M.; Feng, Y.; Wang, S.; and Meng, L. 2025. Maps as a double-edged trustworthiness elevator for climate (mis)information on social media. *International Journal of Digital Earth*, 18(1): 2501245.

Ma, S.; Lei, Y.; Wang, X.; Zheng, C.; Shi, C.; Yin, M.; and Ma, X. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–19. ACM.

Metzger, M. J.; and Flanagin, A. J. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59: 210–220.

Pennycook, G.; McPhetres, J.; Zhang, Y.; Lu, J. G.; and Rand, D. G. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7): 770–780.

Petersen, A. M.; Vincent, E. M.; and Westerling, A. L. 2019. Discrepancy in scientific authority and media visibility of climate change scientists and contrarians. *Nature Communications*, 10(1): 3502.

Schäfer, M. S.; and Painter, J. 2021. Climate journalism in a changing media ecosystem: Assessing the production of climate change-related news around the world. *WIREs Climate Change*, 12(1): e675.

Spitzer, P.; Holstein, J.; Morrison, K.; Holstein, K.; Satzger, G.; and Kühl, N. 2025. Don't Be Fooled: The Misinformation Effect of Explanations in Human–AI Collaboration. *International Journal of Human–Computer Interaction*, 1–29.

Supran, G.; and Oreskes, N. 2017. Assessing ExxonMobil's climate change communications (1977–2014). *Environmental Research Letters*, 12(8): 084019.

Swaroop, S.; Buçinca, Z.; Gajos, K. Z.; and Doshi-Velez, F. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, 138–154. New York, NY, USA.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.

Veltri, G. A.; and Atanasova, D. 2017. Climate change on Twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, 26(6): 721–737.

Vicente, L.; and Matute, H. 2023. Humans inherit artificial intelligence biases. *Scientific Reports*, 13(1): 15737.

Wood, T.; and Porter, E. 2019. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior*, 41(2): 135–163.

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
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in Study Design and Method
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we clarify in the second paragraph of the Conclusion
   (e) Did you describe the limitations of your work? Yes, in the second paragraph of the Conclusion
   (f) Did you discuss any potential negative societal impacts of your work? Yes, the third paragraph of the Conclusion
   (g) Did you discuss any potential misuse of your work? Yes, the third paragraph of the Conclusion
   (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? Yes, the third paragraph of the Conclusion
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA, the study is exploratory and does not rely on theoretical hypotheses.
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
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)? NA, the study does not involve training or evaluating machine learning models.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (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)? NAr
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
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, we cited the data source
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because the study uses publicly available social media content.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the data does not contain personally identifiable information and was handled with care.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA, the study does not curate or release new datasets.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
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? Yes, the full task instructions are provided in Appendix 3.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, the study was approved by the relevant institutional ethics committee.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, participant compensation details are reported in the paper.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, all study data were securely stored on an institutional survey platform and anonymized prior to analysis.

Appendix 1

This appendix presents the prompt used to configure the LLM-based assistants (ChatGPT 5 and Gemini 2.5 Pro) in the user study:

System Prompt

You are a supportive expert for Bias Awareness in News Media and a specialized Q&A bot with expertise in climate change, climate science, environmental science, physics, and energy science. You answer user questions related to social media posts about climate change, which may include headline content, maps, or screenshots of posts.

Your objectives:

1. Provide accurate, comprehensive answers to user questions using the provided post content and your in-house knowledge.
2. When synthesizing the information, eliminate redundancy and cite sources explicitly as 'Reference', 'Page', and 'URL'.
3. If an answer is drawn from in-house knowledge, signify this by appending (In-house knowledge).
4. If sufficient information is lacking, respond with: 'There is not enough info to answer the question.'
5. Maintain accuracy and avoid creating information. If unclear, do not fabricate an answer.

Response format:

- If the user inputs headline content: Rate the headline from -5 (Not Credible) to 5 (Highly Credible). Provide reasoning in under 100 words, explaining credibility, potential misinformation, or cognitive bias.
- If the user inputs a map: Rate the map from -5 (Not Credible) to 5 (Highly Credible). Provide reasoning in under 100 words.
- If the user inputs both headline and map (screenshot of post): First, rate the headline as above, then rate the map as above. Each reasoning should be under 100 words.
- For other types of questions: Ensure there is enough information about the post. If not, respond: 'There is not enough info to answer the question. Please provide more information about the posts.'
- If the user asks for a word explanation: Provide lexical explanations of words/phrases from the given post.
- If the user asks for a text summary: Provide a concise summary of the post content.
- If the user asks for a viewpoint synthesis: Integrate insights and perspectives on the specified issue conveyed in the post.
- If the user asks for a referential fact: Provide factual information related to the content in the post, citing references when possible.

Always balance bias awareness, scientific accuracy, and clarity in responses.

Appendix 2

This appendix provides detailed definitions of the four reasoning dimensions used in the content analysis of participants' credibility explanations.

Prior Knowledge

Participants rely on common sense, personal beliefs, plausibility judgments, or subjective feelings (e.g., "seems reasonable/unreasonable," "I believe," "from my experience"). This dimension captures how criteria for judging trust are influenced by an individual's prior knowledge and existing beliefs rather than by explicit evidence or analytical evaluation.

Evidence Justification

Participants justify their judgments by citing data provenance, institutions, or explicit forms of evidence (e.g., references to datasets, NASA or other official sources, statistics, or the credibility of information sources). This dimension reflects how source cues—such as research organizations, official media outlets, public reports, and similar entities—shape trust judgments.

Emotional Expression

Participants critique a claim's framing, rhetoric, or tone (e.g., "misleading," "selective," "agenda-driven," "exaggerated," or pointing out a headline–map mismatch). This dimension describes how trust judgments can be influenced by perceived emotional expression, persuasive intent, or rhetorical strategies within the message.

Internal Logic

Participants assess credibility based on internal consistency, plausibility, or logical coherence, including contradictions between textual claims and visual representations. This dimension captures how the perceived logic of the content—such as whether the text aligns with what the map presents—across data, text, and visual elements influences trust judgments.

Appendix 3

This appendix presents the full task instructions provided to participants prior to the user study.

Task Overview

What Will You Do?

In this study, you will complete a sequence of tasks designed to understand how people evaluate the credibility of climate-related social media posts.

Specifically, you will:

1. **Complete a pre-questionnaire** about your background, familiarity with climate change topics, and prior experience using AI chatbots.
2. **Complete a warm-up task**, during which the social media simulator will be introduced and you will practice using the interface.
3. **Evaluate social media posts**, during which you will assess a total of nine posts presented in the simulator.
4. **Complete a post-task questionnaire** assessing your experience, perceived mental demand, and impressions of the AI-assisted tools.

The entire study session takes approximately **60 minutes** to complete.

General Instructions

- There are no right or wrong answers; we are interested in your personal judgments and reasoning.
- Please read each post as you do in daily life; you may proceed at your own pace
- All responses will be recorded anonymously and used only for research purposes.