

# Correcting misinformation on social media with a large language model

Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff

Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

## Abstract

Real-world misinformation, often multimodal, can be partially correct and factual but misleading by cherry-picking, conflating correlation with causation, and using other tactics. Such misinformation is severely understudied, challenging to address, and harms many social domains, including democracy, economics, justice, human rights, and public health, particularly on social media, where it can spread rapidly. High-quality and timely correction of misinformation that identifies and explains its inaccuracies and accuracies has been shown to effectively reduce false beliefs. Despite the wide acceptance of manual correction, it is difficult to be timely and scalable, a concern as technologies like large language models (LLMs) make misinformation easier to produce. LLMs also have versatile capabilities that could accelerate misinformation correction—however, they struggle due to a lack of recent information, a tendency to produce false content, and limitations in addressing multimodal information. We propose MUSE, an LLM augmented with access to and credibility evaluation of up-to-date information. By retrieving evidence as refutation or supporting context, MUSE identifies and explains (in)accuracies in a piece of content—not presupposed to be misinformation—with references. It conducts multimodal retrieval and interprets visual content to verify and correct multimodal content. Given the absence of a systematic and comprehensive evaluation approach, we propose and define 13 dimensions of misinformation correction quality, ranging from the accuracy of identifications and factuality of explanations to the relevance and credibility of references. Then, fact-checking experts correspondingly evaluate responses to social media content that are not presupposed to be (non-)misinformation but broadly include incorrect, partially correct, and correct posts that may or may not be misleading. The results demonstrate MUSE’s ability to write high-quality responses to potential misinformation—across modalities, tactics, domains, political leanings, and for information that has not previously been fact-checked online—within minutes of its appearance on social media. Overall, MUSE outperforms GPT-4 by 37% and even high-quality responses from laypeople by 29%. Our work provides a general methodological and evaluative framework to correct misinformation at scale. It reveals LLMs’ potential to help combat real-world misinformation effectively and efficiently.

## Introduction

Misinformation has far-reaching and detrimental effects on individuals and society<sup>1–5</sup>. It erodes public trust in government, decreases civil engagement in elections, and has been viewed as a threat to democracy<sup>3,6,7</sup>. Evidence has demonstrated that election misinformation helped fuel the January 6th attack on the U.S. Capitol, where five people died and more than 100 police officers were injured<sup>8</sup>. Misinformation also drastically increases during outbreaks and disasters, as seen with the “infodemic” during the COVID-19 pandemic, which significantly increased vaccine hesitancy<sup>9,10</sup>. The result is tragic, where COVID-19 vaccines could have prevented at least 232,000 COVID-19-associated deaths between May 2021 and September 2022 in the U.S. alone<sup>11</sup>. Concerns over misinformation on social media have been particularly

significant<sup>12–14</sup>, as the social media context interferes with truth discernment, where users post content without professional moderation and often consume news in a hasty and distracted way<sup>15, 16</sup>.

Fortunately, high-quality and timely correction of misinformation, which identifies what part(s) of its content is or is not accurate and explains why that part of the content is (in)accurate with references, has been shown to effectively reduce the spread of misinformation and false beliefs<sup>4, 17–19</sup>. While domain experts (e.g., on FactCheck.org) and groups of laypeople (e.g., on X Community Notes, formerly Twitter Birdwatch) have played pivotal roles in correcting misinformation<sup>17, 19, 20</sup>, keeping pace with massive social media posts is impossible. As Brandolini’s law indicates<sup>21</sup>, correcting misinformation is laborious—often requiring domain knowledge, information and media literacy, and explanatory skills—whereas creating and spreading misinformation is easy. As a consequence, 88% of suspicious content on X did not receive any response, and 93% did not receive a high-quality response within the first hour (according to X/Twitter Community Notes as of February 2023; Supp. Fig. S1). Even high-quality responses suffer from limited effectiveness when they are created after rather than before initial bursts of attention to misinformation content<sup>4, 14</sup>. The absence of sufficient coverage also markedly diminishes the impact of correction and, as the implied truth effect suggests, may even increase the perceived accuracy of misinformation that escapes correction<sup>22</sup>.

While generative AI models like LLMs raise concerns that they facilitate creating misinformation, they also potentially make scaling up and accelerating misinformation’s correction possible. Recent LLMs have exhibited proficiency in generating fluent and coherent text, laying a foundation for producing explanations that the public can understand. Indeed, LLMs have revolutionized the field of AI and presented remarkable capabilities across domains and tasks<sup>23–25</sup>. However, accurate and trustworthy misinformation correction is inseparable from accessing up-to-date and reliable information, providing accurate references to back up claims, and addressing textual and visual information, all areas where existing LLMs fall short<sup>26</sup>. Specifically, GPT-4<sup>27</sup> (as of March 2023), Llama-3 (as of July 2024)<sup>28</sup>, MisinfoCorrect<sup>29</sup>, MADR<sup>30</sup>, JustiLM<sup>31</sup>, and many other LLMs lack access to constantly growing and changing knowledge. These models are thereby ill-equipped to combat misinformation on emerging topics. They either do not provide or “hallucinate” references, which can be fabricated or irrelevant<sup>26, 32</sup>. A growing body of literature has focused on retrieval-augmented LLMs, which can retrieve up-to-date information from Wikipedia or the entire Internet<sup>33–39</sup>. However, their retrieval does not explicitly consider the factuality and bias of retrieved sources, posing risks of generating falsehood and backfiring (i.e., reinforcing rather than reducing false beliefs)<sup>40–42</sup>. Every above LLM also struggles with counteracting multimodal misinformation due to their nonacceptance of visual inputs.

Real-world misinformation commonly encompasses content that is partially correct and even factual but misleading through various tactics and across domains<sup>2, 3</sup>. Such misinformation can be more difficult to recognize, prevalent, and harmful<sup>2, 3</sup>. Existing LLM and other AI models to tackle misinformation focus narrowly on it that is factually incorrect (e.g., MADR<sup>30</sup>, FOLK<sup>38</sup>, and ProgramFC<sup>43</sup>), within a single domain (e.g., politics<sup>30</sup> or COVID-19<sup>29, 39</sup>), or employing one specific tactic (e.g., repurposing authentic images with fabricated stories<sup>44</sup>). A general AI framework to practically identify and correct misinformation is demanded but absent. Evaluating the quality of a correction presented in natural language is another core challenge. Related studies heavily rely on automatic evaluation, which calculates the similarity between model-generated and “ground-truth” text using metrics such as ROUGE<sup>31, 44</sup>. However, what qualifies a correction as the ground truth is unclear<sup>29</sup>, and existing metrics struggle to capture content factuality, especially when compared to fact-checking experts<sup>26</sup>. The need for an accurate and comprehensive understanding of a model’s performance in identifying and explaining (in)accuracies, generated text, and references is recognized, but a solution is still lacking.

In this article, we propose MUSE, a scalable approach for multimodal misinformation correction.

MUSE makes use of an LLM and augments it with the ability to handle images, access timely and credible knowledge on the web, retrieve evidence that refutes or contextualizes the given content that may or may not be misinformation, and generate clear explanations with accurate and trustworthy references. MUSE is a nonparametric approach and therefore easily and cheaply updated, especially compared to parametric language models. Fact-checking experts comprehensively evaluate MUSE-generated responses to real social media posts that potentially are misinformation, and compare them to baselines including GPT-4 and high-quality responses based on the collective efforts of laypeople. Our assessment measures the overall quality of a response, specifically defined as the explicitness, accuracy, comprehensiveness, and informativeness when it identifies and explains (in)accuracies; the relevance, factuality, fluency, coherence, and toxicity of generated text; and the reachability, relevance, and credibility of references. We find that MUSE outperforms GPT-4 by 37% and even high-quality responses from laypeople by 29% in accurately and promptly responding to potential misinformation. Results demonstrate MUSE's advance no matter whether the potential misinformation content is textual with or without images; fabricates a story, misinterprets or misrepresents facts, lacks context, implies false or oversimplified causation, uses loaded language, presents false or biased data, has improper analogy or equivalence, and adopts other tactics; across a broad range of domains, including politics and international affairs, economy and business, crime and law, social issues and human rights, and health and medicine; leans liberal or conservative; or whether the content has previously fact-checked online or not.

## Approach

MUSE is designed to automatically respond to content that potentially is misinformation. In other words, the content might be (partially) inaccurate or factually accurate but misleading, all of which are misinformation, or fully accurate as non-misinformation. The content can contain text with or without visuals. The response should identify what part(s) of the content is or is not accurate, explain why that part of the content is or is not accurate, and provide links as references. Overall, MUSE features three modules to reliably and efficiently address generalized misinformation: (1) a highly parallelizable relevance- and credibility-aware evidence retriever, (2) an LLM-based evidence-assisted response generator, and (3) an image describer that enables text-only LLMs to jointly interpret multimodal signals from the input.

Fig. 1 illustrates MUSE's pipeline. We start by introducing the details of MUSE with a piece of text-only potential misinformation as the input. First, MUSE generates queries based on an LLM from the potential misinformation (Fig. 1b; Methods). Each query acts as the input of a web search engine to access timely updated web content and obtain a list of web links directly relevant to the query (Fig. 1b; Methods). After scraping the content from these web links, MUSE calculates their direct relevance to the potential misinformation and removes irrelevant web pages (Fig. 1b; Methods). Then, MUSE determines the credibility of web pages by looking up their publishers' factuality and bias ratings and selects pages with high factuality and minimal bias (Fig. 1c; Methods). Next, MUSE leverages an LLM to extract text from each of the web pages as evidence. Such evidence can refute the potential misinformation, typically happening when it is misinformation with false claims, or provide additional context, which can demonstrate that the potential misinformation is accurate or part(s) of its claims are accurate (Fig. 1d; Methods). Finally, MUSE generates a response to the potential misinformation by providing an LLM with the extracted pieces of evidence and their source web links (Fig. 1d; Methods).

Note that content, especially content posted on social media, often contains extraneous information that does not need verification and is irrelevant to correction, including unverifiable opinions or emojis, such as the textual content of the false post in Fig. 1. Therefore, generating queries instead of simply using the post content improves web searches as a way of denoising the post content; see example queries in

Fig. 1b. Meanwhile, generating multiple queries helps decompose a post, which may have multiple claims that each needs verification or correction, whereas generating one query may overlook some of the claims and hence lead to not comprehensive identifications and explanations of (in)accuracies (Supp. Fig. S2). Another concern may arise from filtering retrieved web pages based on how relevant the page content is to the potential misinformation: theoretically, including all retrieved web pages increases the amount of extracted evidence, which may not hurt and perhaps even benefit correction. However, the increase in selected web pages drastically elevates the expense of MUSE (Methods). We also observe that retrieved web pages with relatively low relevance can increase the prevalence of hallucinations when generating responses (Supp. Fig. S3). Moreover, as illustrated in Fig. 1c, MUSE filters and ranks the selected web pages by their publishers' factuality and bias. It starts extracting evidence from pages with the highest factuality and least bias and then continues down the ranking, stopping when it has obtained sufficient refutations (i.e., at least two web pages were found to refute the misinformation) or gone through all the credible pages.

When the input is a piece of multimodal (textual and visual) content, MUSE first generates textual description of each image (Fig. 1a) so that the content can be handled by any LLM on downstream tasks, including query generation (Fig. 1b), evidence extraction (Fig. 1d), and response generation (Fig. 1d). Specifically, MUSE augments image captioning models developed to describe an image in natural language with recognizing celebrities and optical characters based on an LLM (Fig. 1a; Methods). Compared to existing image captioning models that capture global features of images, MUSE produces more informative descriptions with features crucial for making accurate verification and corrections<sup>45</sup> (Methods). For example, even a state-of-the-art image captioning model<sup>46</sup> may describe the visual misinformation in Fig. 1 as simply “list of banned books in Florida.” The description overlooks the listed titles of books that are essential for the visual content’s verification and correction (see more examples in Supp. Fig. S4). In addition, MUSE conducts multimodal search on the web and computes multimodal relevance to filter out irrelevant web pages (Fig. 1b; Methods).

## Evaluation

MUSE’s evaluation was based on X Community Notes data. Community Notes empowers people on X, often laypeople, to collaboratively fact-check tweets, which has been shown to reduce the spread of misinformation<sup>17</sup>. Every laypeople’s free-response fact-check is associated with a helpfulness score by aggregating the assessments of people with diverse backgrounds, e.g., different political ideologies. A response with a sufficiently high helpfulness score is then displayed on the corresponding tweet and publicly visible<sup>17</sup> (Supp. Fig. S6). We included the tweets from Community Notes with at least two responses ( $n = 247$ , posted from February 2021 to February 2023); one has a high helpfulness score and the other has an average helpfulness score (as of February 2023; Methods). Though we do not presuppose the accuracy of the tweets in MUSE’s design and evaluation, we found that more than half of the tweets are not fully (in)accurate or misleading but frequently presented in a way that combines accurate claims and inaccurate or misleading claims (Methods). We further generated responses to these tweets based on MUSE and GPT-4 (as of June 2023) (Discussion; Methods). Experts in fact-checking and journalism evaluated the quality of responses by various approaches to the same tweet (Methods); they were blinded to which approach had generated each response. The evaluation contains 13 specific criteria, covering how well a response identifies and explains (in)accuracies, the quality of generated text, and the quality of references (Methods). It also contains the overall quality of a response by taking all 13 evaluation criteria into account (Methods).

Our primary finding is that *the overall quality of MUSE-generated responses is higher than responses*

by GPT-4 and even high-helpfulness responses by laypeople (Fig. 2a). The overall quality of MUSE-generated responses has an average score of 8.1 out of 10, 29% higher than laypeople's high-helpfulness responses (mean: 6.3;  $p = 3 \times 10^{-48}$ , by Mann-Whitney U test unless otherwise specified;  $N = 464$ ), 37% higher than GPT-4-generated responses (mean: 5.9;  $p = 4 \times 10^{-42}$ ;  $N = 464$ ), and 56% significantly higher than laypeople's average-helpfulness responses (mean: 5.2;  $p = 5 \times 10^{-81}$ ;  $N = 462$ ). Despite statistical insignificance between the overall quality of laypeople's high-helpfulness responses and GPT-4-generated responses ( $p = 0.4$ ;  $N = 464$ ), the overall quality of GPT-4-generated responses has the highest variability, and GPT-4 generates more responses with extremely low quality. The standard deviation of the overall quality of GPT-4-generated responses is 2.7, vs only 2.0 for MUSE and laypeople. 10% of GPT-4's generated responses have a quality score of 0 (lowest) or 1 out of 10, whereas this proportion is 5% for laypeople's average-helpfulness responses, 3% for laypeople's high-helpfulness responses, and 2% for MUSE-generated responses. Note that laypeople's responses were created on average 14 hours after the tweet was posted on social media. Here, MUSE only retrieved web pages published *before* the tweet was posted (Methods).

Examining specific components of response quality, results show that MUSE *outperforms GPT-4 and laypeople who produce even high-helpfulness responses in identifying and explaining (in)accuracies* (Fig. 2b-f). Experts assessed that MUSE-generated responses more explicitly identify and explain where and why a tweet is (in)accurate than GPT-4's and laypeople's high-helpfulness responses (Fig. 2b). 89% of MUSE's generated responses explicitly identify and explain (in)accuracies, 16% more than GPT-4-generated responses, 29% more than laypeople's high-helpfulness responses, and 43% more than laypeople's average-helpfulness responses (Fig. 2b). As for identifying where a tweet is (in)accurate, we found that MUSE more comprehensively identifies a tweet's (in)accuracies with fewer mistakes—here, mistakes indicate falsely claiming where a tweet should be inaccurate as accurate or where a tweet should be accurate as inaccurate—than GPT-4 and laypeople who produce even high-helpfulness responses (Fig. 2c-d). 91% of MUSE's generated responses have at least one correct identification without *any* mistake, 11% more than laypeople's high-helpfulness responses, 19% more than GPT-4-generated responses, and 26% more than laypeople's average-helpfulness responses (Fig. 2c). MUSE has 61% of generated responses accurately identifying *all* the (in)accuracies in a tweet, vs GPT-4 has 38%, laypeople who produce high-helpfulness responses have 26%, and laypeople who produce average-helpfulness responses have 17% only (Fig. 2d). Furthermore, MUSE explains (in)accuracies more precisely and informatively than GPT-4 and laypeople who produce even high-helpfulness responses (Fig. 2e-f). 70% of responses by MUSE have *fully* accurate explanations, vs 55% for laypeople's high-helpfulness responses, 47% by GPT-4, and 37% for laypeople's average-helpfulness responses only (Fig. 2e). Meanwhile, the average informativeness score of MUSE-generated responses is 7.9, 32% higher than laypeople's high-helpfulness responses, 36% higher than GPT-4-generated responses, and 65% higher than laypeople's average-helpfulness responses (Fig. 2f).

Results also demonstrate that MUSE *outperforms GPT-4 and laypeople who produce even high-helpfulness responses in the quality of generated text* (Fig. 2g-k). MUSE, when it augments GPT-4 with the capabilities of accessing timely updated knowledge and addressing visuals (Methods), exhibits enhanced relevance ( $p = 10^{-15}$ ;  $N = 460$ ) and factuality ( $p = 2 \times 10^{-20}$ ;  $N = 459$ ) of text compared to GPT-4 without sacrificing fluency ( $p = 0.6$ ;  $N = 464$ ), coherence ( $p = 0.1$ ;  $N = 459$ ), and toxicity ( $p = 0.8$ ;  $N = 464$ ). Meanwhile, MUSE-generated text is more relevant to the responded tweet ( $p = 2 \times 10^{-30}$ ;  $N = 464$ ), factual ( $p = 4 \times 10^{-6}$ ;  $N = 463$ ), fluent ( $p = 2 \times 10^{-10}$ ;  $N = 464$ ), and coherent ( $p = 10^{-5}$ ;  $N = 451$ ) than the text of high-helpfulness responses by laypeople and additionally less toxic than the text of average-helpfulness responses by laypeople ( $p = 4 \times 10^{-12}$ ;  $N = 462$ ). MUSE-generated text has an average relevance score of 8.7, 18% higher than GPT-4-generated text, 21% higher than the text of

high-helpfulness responses by laypeople, and 43% higher than the text of average-helpfulness responses by laypeople (Fig. 2g). 74% of MUSE-generated text is *completely* factual, vs 59% for the text of even high-helpfulness responses by laypeople and 45% for GPT-4-generated text (Fig. 2h). Almost all of MUSE-generated text does not have any mistake in the use of English (Fig. 2i) and is not biased, impolite, and provoking (Fig. 2k), and 91% is highly coherent and logical (vs 76% and 61% for laypeople, Fig. 2j).

Additionally, results reveal that MUSE *outperforms GPT-4 and laypeople who produce even high-helpfulness responses in the quality of references* (Fig. 2l-n). First, GPT-4 hallucinates references frequently. 49% of its links result in “page-not-found” errors (Fig. 2l), and only 76% of reachable links are relevant to the generated text (Fig. 2m). MUSE significantly reduces such hallucinations with nearly 100% links being reachable (Fig. 2l) and 96% reachable links being relevant to the generated text (Fig. 2m). Meanwhile, MUSE’s references are more credible than the references offered in even high-helpfulness responses by laypeople ( $p = 4 \times 10^{-11}$ ;  $N = 744$ ; Fig. 2n).

We further analyzed the robustness of MUSE in addressing diverse forms of real-world misinformation, which may vary in modality, difficulty to correct, political leaning, domain, and adopted tactics. We first observed that the quality of MUSE-generated responses to textual and multimodal (textual and visual) content is consistently higher than responses by GPT-4 and even high-helpfulness responses by laypeople, with a margin of 21% or higher (Fig. 3a). GPT-4 can be comparable to laypeople who write high-helpfulness responses when responding to text-only tweets. However, its performance significantly declines when responding to multimodal tweets, often resulting in a quality lower than laypeople who produce average-helpfulness responses.

Second, we removed tweets where MUSE’s response references at least one related fact-checking article, accounting for 16% of all tweets. Notably, in only 6% of all tweets, MUSE’s response exclusively references related fact-checking article(s). The quality of MUSE-generated responses to these tweets can be more associated with MUSE’s ability to summarize retrieved professional fact-checking articles, rather than fact-checking such content from scratch. We found that the quality of MUSE-generated responses to content that has *not* been fact-checked online is at least 28% higher than responses by GPT-4 and even high-helpfulness responses by laypeople (Fig. 3b). The average quality score of MUSE-generated responses for these potentially more challenging tweets is 8.0, close to its 8.1 average score for all tweets.

Third, LLMs have exhibited political bias<sup>47</sup>, which may diminish their ability to debunk misinformation that aligns with their political stance. We annotated the political leaning of tweets and found that GPT-4, which has an observed liberal bias<sup>47</sup>, indeed performs 5% worse in correcting misinformation that aligns with liberal ideology compared to conservative ideology (Fig. 3c). Nevertheless, the quality of MUSE-generated responses to both liberal and conservative content has the same average score of 8.3. This score is consistently higher than GPT-4-generated responses and even high-helpfulness responses from laypeople by at least 26%.

Finally, we annotated the domain of tweets and tactics used to make them or part of them false or misleading. Results in Fig. 3d-e indicate that the quality of MUSE-generated responses to potential misinformation across different domains and tactics maintains an average score around 8, consistently higher than responses by GPT-4 and even high-helpfulness responses from laypeople, with a margin of at least 19%. Here, domains broadly include politics and international affairs, economy and business, crime and law, social issues and human rights, and health and medicine. Tactics primarily include fabricating, e.g., a story with or without an authentic image, or using a digitally fabricated or altered screenshot or photo; misinterpreting or misrepresenting someone’s claim, a symbol, a policy, and others; lacking context; implying false or oversimplified causation, such as conflating correlation with causation; using loaded language; presenting false, partial, or biased data, such as cherry-picking; and having improper analogy or equivalence (see examples in Supp. Fig. S7).

## Discussion

While concerns have arisen about LLMs in facilitating the creation of misinformation<sup>48,49</sup>, our work demonstrates LLMs' potential to improve the online information ecosystem by correcting misinformation<sup>2,3</sup>. Correcting misinformation requires identifying what part(s) of the content is (in)accurate and explaining why that part of the content (in)accurate with trustworthy references. Such identifications with clear explanations have been shown to reduce misinformation's spread and people's false beliefs, but its effectiveness is affected by quality, timeliness, and scalability<sup>4,22</sup>. However, research on scaling up and accelerating misinformation correction is still at the early stage. Existing AI models are ill-equipped to identify and correct misinformation on social media. Such misinformation often spreads rapidly and is not restricted to narrow domains. It can combine (in)accurate, factually accurate but misleading, and unverifiable claims. It can employ various tactics, deliberately or accidentally, to create falsehood or misleadingness. Addressing it requires comprehending content that can be multimodal and the context beyond it that often involves emerging events. Existing methods to evaluate the quality of corrections presented in natural language also struggle to be accurate and comprehensive. Such evaluation relies on professional knowledge and skills. It should thoroughly assess the identification and explanation of (in)accuracies, as well as the use of references from diverse perspectives.

We propose MUSE, which augments existing powerful LLMs (here, it is GPT-4) with the capabilities of addressing images, accessing up-to-date knowledge, and finding accurate references, and 13 dimensions of misinformation correction quality. Evaluation by fact-checking experts demonstrates the high quality of MUSE's automatically generated responses to social media posts that potentially are misinformation. Results further validate that GPT-4 struggles to effectively respond to visual content but MUSE excels (Fig. 3a). MUSE also exhibits significantly fewer "hallucinations" by having fewer errors in identifying and explaining (in)accuracies (Fig. 2c,e), generating text that is more factual and relevant to the responded content (Fig. 2g-h), and providing more references that are real and relevant to the generated text (Fig. 2l-m) than GPT-4. Note that addressing LLM 'hallucination' has been a long-standing challenge stressed in many works<sup>26,50</sup>.

We provide MUSE as a solution to assist social media users and platforms in reliably, scalably, promptly, and transparently responding to suspicious content. MUSE is end-to-end and thereby simple to use. It is nonparametric and thereby easily and cheaply updated<sup>51</sup>. Meanwhile, our results reveal that MUSE-generated responses have high quality in identifying and explaining inaccuracies, generated text, and provided references, significantly surpassing GPT-4 and laypeople who produce even high-helpfulness responses in correcting misinformation across modalities, domains, political leanings, and tactics and that has not even been fact-checked online (Fig. 2-3). Besides the highest accuracy and factuality (Fig. 2c,e,g-h,l-m), MUSE's generated responses show the highest readability by being the most explicit, fluent, and coherent (Fig. 2b,i-j). MUSE can reduce the risk of a correction backfiring (i.e., reinforcing rather than reducing people's false beliefs) by generating responses with the least toxic text and most credible references<sup>29,41,42</sup> (Fig. 2k,n). It can also reduce the implied truth effect (i.e., increasing people's perceived accuracy of overlooked inaccuracies)<sup>22</sup> by comprehensively identifying all the inaccuracies in a social media post (Fig. 2d) and being more capable of correcting misinformation at scale. These high-quality responses by MUSE can generally be obtained within minutes of suspicious content appearing on social media (Methods). By transparently providing references that refer to the retrieved web pages where evidence was collected, users can become more informed and also verify responses themselves (Fig 2,3f). MUSE can also inform users of the current lack of evidence, provide an accuracy nudge, and when applicable, express uncertainty (Supp. Fig. S8).

MUSE's responses cost about 0.5 USD per social media post at the time of our evaluation, though this

cost has now been reduced to 0.2 USD, as GPT-4's price has lowered (as of February 2024; Methods). Our focus in designing MUSE was in maximizing the quality of corrections. Considering the task's complexity and our significant improvement in quality, the cost is relatively inexpensive compared to alternatives. For example, a crowd of laypeople can already cost about 0.9 USD<sup>19</sup> to identify whether a new article's headline and lede contains misinformation *without* writing down the explanation. Notably, we also changed MUSE's foundation LLM from GPT-4 to open-sourced Llama-3 (70B) and observed two MUSES are comparable without modifying *any* other implementation details (Supp. Fig. S9; Methods), suggesting MUSE's generalizability. Note that the knowledge cutoff for Llama-3 is December 2023, whereas that of GPT-4 is September 2021. Llama-3's training data might have included some or even all tweets used in evaluation and related events, knowledge, and corrections (e.g., their Community Notes data). As a comparison, GPT-4's training data might only include up to 10% (Methods, where we show that MUSE's performance remains consistent after removing these tweets). Hence, it is unfair to compare between Llama-3-based and GPT-4-based MUSE, and we did not include Llama-3 and its based MUSE in expert evaluation. Nevertheless, our qualitative analysis illustrated in Supp. Fig. S9 reveals the potential to significantly reduce MUSE's cost, as GPT-4 was MUSE's most costly component. Furthermore, as one of the most capable open-sourced LLMs, Llama-3 still cannot accept visual inputs (as of August 2024), emphasizing the importance of MUSE's image describer module.

This work also faces the following limitations. First, although MUSE is capable of responding to multimodal misinformation with text and images, it cannot accept video inputs. Second, we only focus on English, one of the most spoken languages in the world. Third, we evaluated MUSE using real social media content on a single platform, X, as its Community Notes system has been shown to reduce the spread of misinformation<sup>17</sup> and transparent. X is also a popular social media platform, and one where more than half of users consume news regularly<sup>52</sup> and where misinformation has been shown to diffuse faster than the truth<sup>13</sup>. Fourth, experts assessed and compared MUSE against one foundation LLM, GPT-4, which can also be seen as an ablation study, since MUSE augments GPT-4 (see Methods and Supp. Fig. S24 for additional results of our ablation study). GPT-4 was chosen as a comparison as it is one of the best performing LLMs across a wide range of tasks available today<sup>53,54</sup>.

## Conclusion

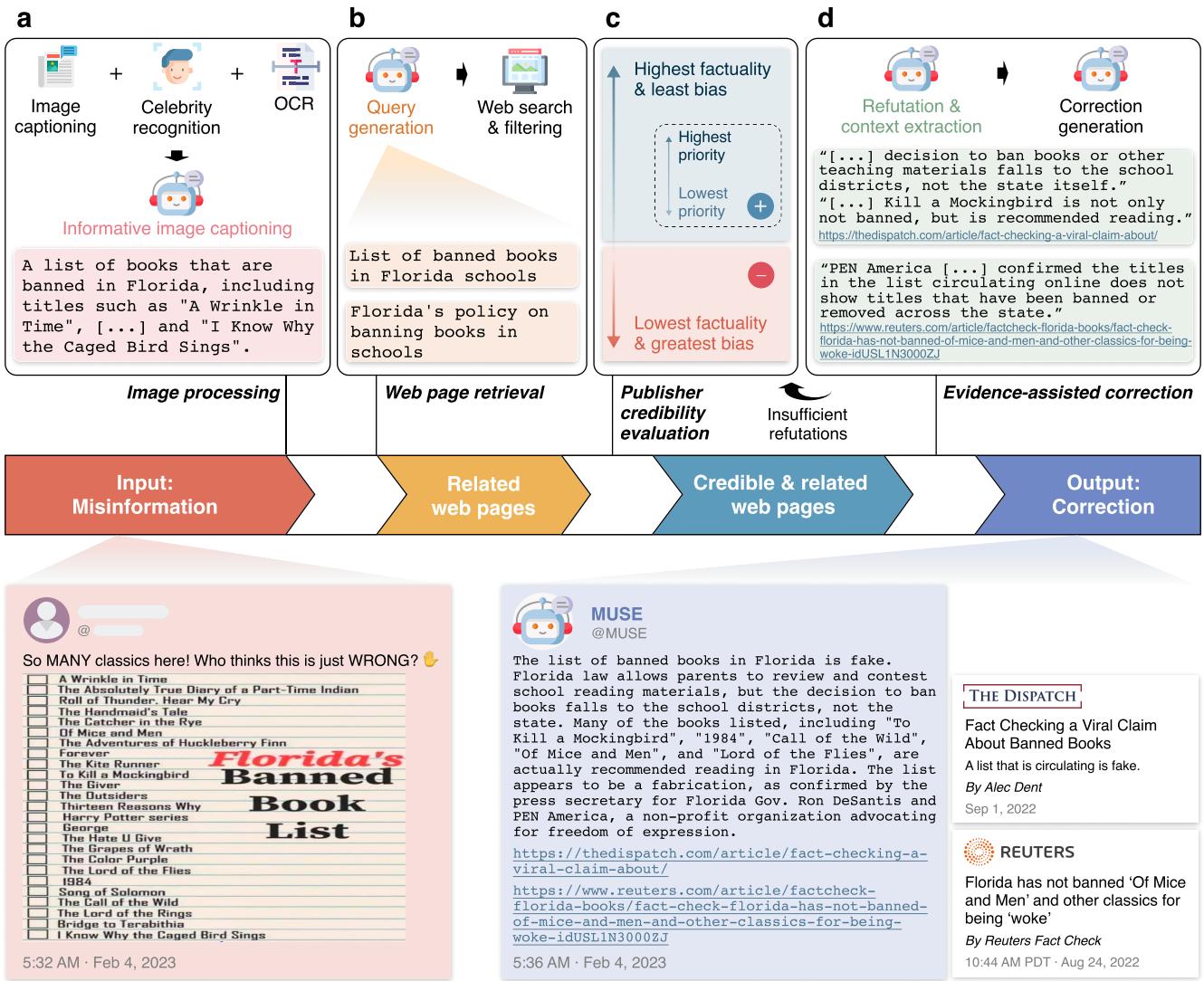
We proposed MUSE, a nonparametric LLM to identify and explain (in)accuracies in social media content. MUSE can integrate textual and visual information, access timely knowledge, and generate responses that use natural languages and have accurate and trustworthy references. Experts assessed that MUSE significantly outperforms GPT-4 and even laypeople who write high-helpfulness responses on X Community Notes in identifying and explaining (in)accuracies, generating high-quality text, providing high-quality references, and producing high-quality responses overall. MUSE excels when the responded content is textual or multimodal, related to a broad range of domains, false or misleading by applying various tactics, and even when the content has not been fact-checked online. MUSE also performs consistently well when responding to liberal and conservative content. This work demonstrates the potential of LLMs to respond to online misinformation in a scalable, prompt, reliable, and transparent manner—addressing key bottlenecks where existing studies struggle. This work also contributes to a comprehensive framework, incorporating with 13 associated standards (Methods, where we analyzed their importance), for evaluating misinformation correction models. Additionally, we publicly release rich annotations on social media posts covering aspects such as political leanings, domains, and tactics that make them false or misleading, along with expert-annotated responses to enable research.

## References

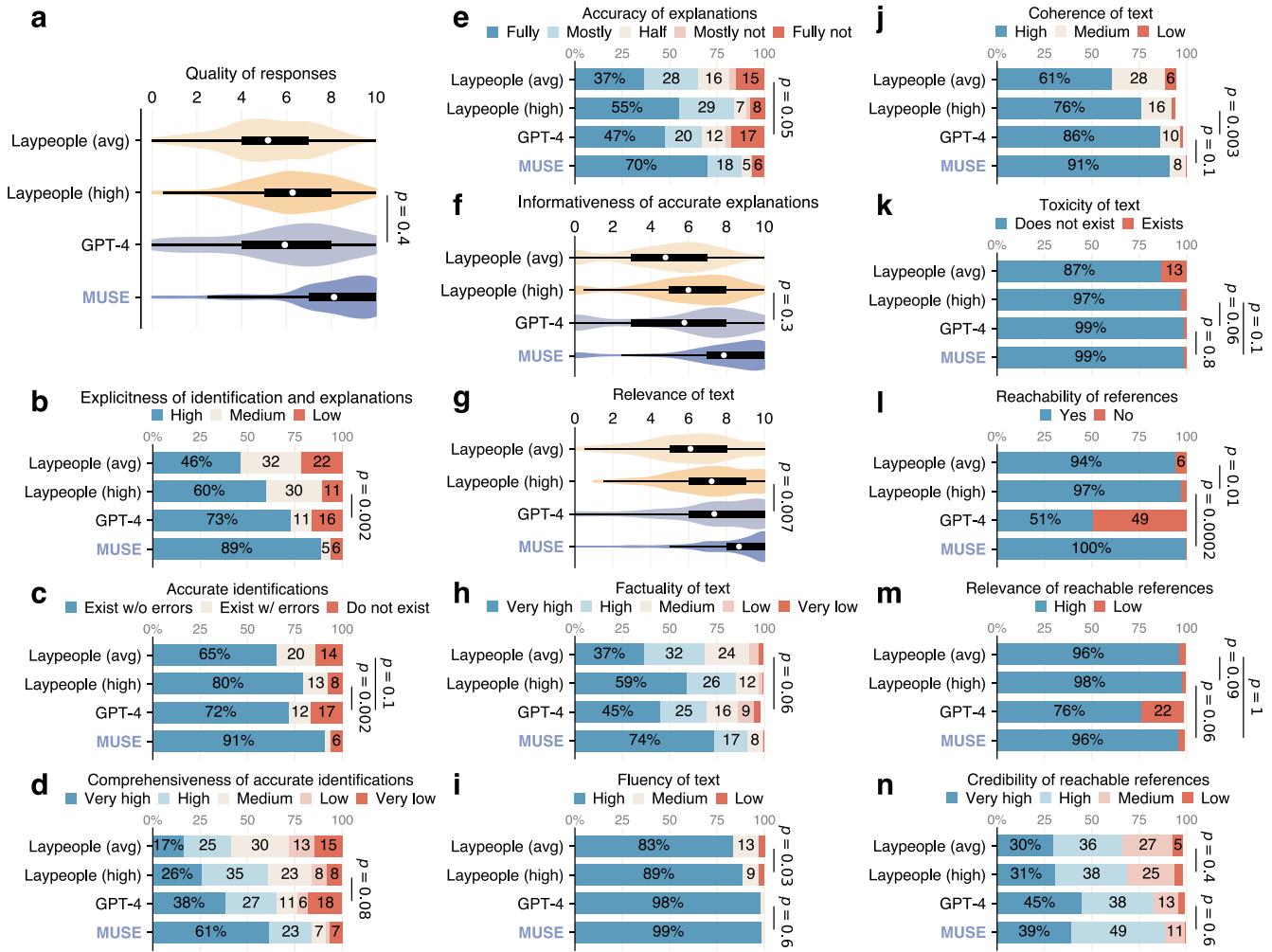
1. Lazer, D. M. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
2. Traberg, C. S. Misinformation: broaden definition to curb its societal influence. *Nature* **606**, 653–653 (2022).
3. Watts, D. J., Rothschild, D. M. & Mobius, M. Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci.* **118**, e1912443118 (2021).
4. Van Der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat. Medicine* **28**, 460–467 (2022).
5. West, J. D. & Bergstrom, C. T. Misinformation in and about science. *Proc. Natl. Acad. Sci.* **118**, e1912444117 (2021).
6. Moore, R. C., Dahlke, R. & Hancock, J. T. Exposure to untrustworthy websites in the 2020 US election. *Nat. Hum. Behav.* 1–10 (2023).
7. Forum, W. E. Global risks report 2024 (World Economic Forum, 2024).
8. Thompson, B. G. *et al.* *Final Report of the Select Committee to Investigate the January 6th Attack on the United States Capitol* (U.S. Government Publishing Office, 2022).
9. Zarocostas, J. How to fight an infodemic. *The Lancet* **395**, 676 (2020).
10. Pertwee, E., Simas, C. & Larson, H. J. An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nat. Medicine* **28**, 456–459 (2022).
11. Jia, K. M. *et al.* Estimated preventable COVID-19-associated deaths due to non-vaccination in the United States. *Eur. J. Epidemiol.* 1–4 (2023).
12. Ledford, H. Deepfakes, trolls and cybertroopers: how social media could sway elections in 2024. *Nature* (2024).
13. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
14. Bak-Coleman, J. B. *et al.* Combining interventions to reduce the spread of viral misinformation. *Nat. Hum. Behav.* **6**, 1372–1380 (2022).
15. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Sci. Adv.* **9**, eab06169 (2023).
16. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl. Acad. Sci.* **120**, e2216614120 (2023).
17. Wojcik, S. *et al.* Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723* (2022).
18. Porter, E., Velez, Y. & Wood, T. J. Factual corrections eliminate false beliefs about COVID-19 vaccines. *Public Opin. Q.* **86**, 762–773 (2022).
19. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).
20. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Crowds can effectively identify misinformation at scale. *Perspectives on Psychol. Sci.* 17456916231190388 (2022).
21. Williamson, P. Take the time and effort to correct misinformation. *Nature* **540**, 171–171 (2016).

22. Pennycook, G., Bear, A., Collins, E. T. & Rand, D. G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manag. Sci.* **66**, 4944–4957 (2020).
23. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
24. Editorials. AI will transform science – now researchers must tame it. *Nature* **621** (2023).
25. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
26. Augenstein, I. *et al.* Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189* (2023).
27. OpenAI. GPT-4 technical report. *ArXiv abs/2303.08774* (2023).
28. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
29. He, B., Ahamad, M. & Kumar, S. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, 2698–2709 (2023).
30. Kim, K. *et al.* Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401* (2024).
31. Zeng, F. & Gao, W. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions Assoc. for Comput. Linguist.* **12**, 334–354 (2024).
32. Peskoff, D. & Stewart, B. M. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 427–438 (2023).
33. Shuster, K., Poff, S., Chen, M., Kiela, D. & Weston, J. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803 (2021).
34. Peng, B. *et al.* Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813* (2023).
35. Mallen, A. *et al.* When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822 (2023).
36. Vu, T. *et al.* FreshLLMs: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214* (2023).
37. Gao, L. *et al.* RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508 (2023).
38. Wang, H. & Shu, K. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6288–6304 (Association for Computational Linguistics, 2023).

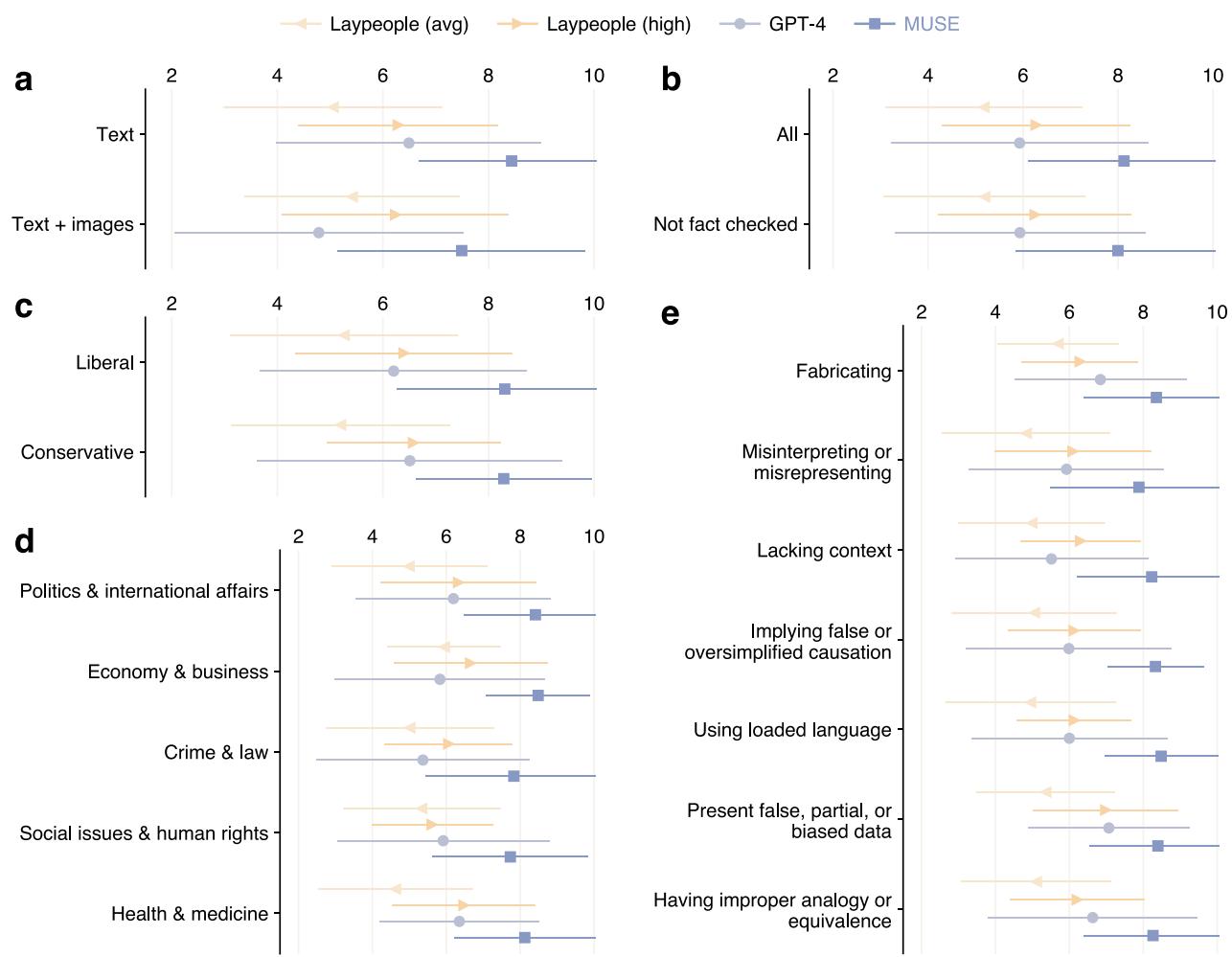
- 39.** Yue, Z. *et al.* Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5628–5643 (2024).
- 40.** Menick, J. *et al.* Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022).
- 41.** Ecker, U. K. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).
- 42.** Swire-Thompson, B., DeGutis, J. & Lazer, D. Searching for the backfire effect: Measurement and design considerations. *J. Appl. Res. Mem. Cogn.* **9**, 286–299 (2020).
- 43.** Pan, L. *et al.* Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6981–7004 (2023).
- 44.** Qi, P., Yan, Z., Hsu, W. & Lee, M. L. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062 (2024).
- 45.** Stefanini, M. *et al.* From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 539–559 (2022).
- 46.** Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- 47.** Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11737–11762 (2023).
- 48.** Menczer, F., Crandall, D., Ahn, Y.-Y. & Kapadia, A. Addressing the harms of AI-generated inauthentic content. *Nat. Mach. Intell.* **5**, 679–680 (2023).
- 49.** Pan, Y. *et al.* On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1389–1403 (Association for Computational Linguistics, 2023).
- 50.** Hu, X. *et al.* Do large language models know about facts? *arXiv preprint arXiv:2310.05177* (2023).
- 51.** Asai, A., Min, S., Zhong, Z. & Chen, D. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, 41–46 (2023).
- 52.** Walker, M. & Matsa, K. E. News consumption across social media in 2021. *Pew Res. Cent.* (2021).
- 53.** Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. GPT-4 passes the bar exam. *Philos. Transactions Royal Soc. A* **382**, 20230254 (2024).
- 54.** Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).



**Figure 1.** Overview of MUSE, an LLM augmented by addressing images and accessing timely knowledge from credible publishers to enable identifying and explaining (in)accuracies in a piece of multimodal content with accurate and trustworthy references. Given a piece of content that may or may not be misinformation, MUSE searches for related and credible web pages, from which extracts evidence as refutations or contexts, with which generates a response identifying and explaining the (in)accuracies within it. **a:** Image processing. MUSE augments image captioning models with celebrity and optical character recognition (OCR) to generate informative descriptions of images. **b:** Retrieval of related web pages. MUSE retrieves web pages using LLM-generated queries and a web search engine and filters them based on their multimodal relevance to the given content. **c-d:** Credibility evaluation of the publishers of web pages (**c**) and evidence-assisted response generation (**d**). MUSE filters and ranks publishers based on their professionally rated factuality and bias. It starts from the web pages with the highest factuality and least bias and leverages an LLM to extract evidence refuting or contextualizing the given content. It continues down the ranking, stopping when it has obtained sufficient refutations (i.e., at least two pages were found to refute the misinformation) or gone through all the credible pages. Finally, it generates a response by providing an LLM with the extracted evidence. Besides identifying and correcting a false post shown here, MUSE can also identify and respond to accurate, partially accurate, and factually accurate but misleading (see examples in Supplementary Fig. S5).



**Figure 2.** Results of expert evaluation ( $p < 2 \times 10^{-5}$  for each approach pair respectively in **a-n** by Mann-Whitney U test; experiments=84). **a:** The overall quality of MUSE-generated responses (mean  $\pm$  SD:  $8.1 \pm 2.0$ ;  $n = 232$ ) is 29% higher than laypeople’s high-helpfulness responses ( $6.3 \pm 2.0$ ; 232), 37% higher than GPT-4-generated responses ( $5.9 \pm 2.7$ ; 232), and 56% higher than laypeople’s average-helpfulness responses ( $5.2 \pm 2.1$ ; 230). **b-f:** The quality of identifying and explaining inaccuracies. MUSE-generated responses more explicitly identify and explain inaccuracies (**b**), more comprehensively identify inaccuracies with fewer mistakes that falsely state an accurate claim as inaccurate or an inaccurate claim as accurate (**c-d**), and more accurately and informatively explain inaccuracies (**e-f**) than GPT-4-generated and laypeople’s high- and average-helpfulness responses. **g-k:** The quality of generated text. MUSE’s generated text is more relevant to the responded misinformation and factual than GPT-4’s generated text and the text of high- and average-helpfulness responses by laypeople (**g-h**). MUSE-generated text is more fluent and coherent than the text of high-helpfulness responses by laypeople and additionally less toxic than the text of average-helpfulness responses by laypeople (**i-k**). **l-n:** The quality of links as references. MUSE rarely while GPT-4 frequently hallucinates references; MUSE provides significantly more reachable links that are relevant to the generated text (**l-m**). MUSE’s references are more credible than the references offered in high- and average-helpfulness responses by laypeople (**n**). Note that laypeople’s responses were created on average 14 hours after the social media post. Here, MUSE only retrieved web pages published *before* the post (Methods).



**Figure 3.** Quality of responses to social media posts across modalities, fact-checking statuses, political divides, domains, and tactics used to make them or part of them false or misleading. **a:** MUSE consistently outperforms GPT-4 and laypeople who produce even high-helpfulness responses by at least 21% when responding to textual content (n=155) and multimodal content (n=77). **b:** MUSE outperforms GPT-4 and laypeople who produce even high-helpfulness responses by at least 28% even when responding to content that has not been fact-checked online (n=195). **c:** MUSE consistently outperforms GPT-4 and laypeople who produce even high-helpfulness responses by at least 26% when responding to liberal content (n=110) and conservative content (n=50). **d:** MUSE consistently outperforms GPT-4 and laypeople who produce even high-helpfulness responses by at least 25% when responding to content about politics and international affairs (n=80), economy and business (n=38), crime and law (n=38), social issues and human rights (n=30), and health and medicine (n=24). **e:** MUSE consistently outperforms GPT-4 and laypeople who produce even high-helpfulness responses by at least 19% when responding to misinformation that includes misinterpretations or misrepresentations (n=51), false or oversimplified causation (n=35), lack of context (n=33), fabrications (n=31), loaded language (n=30), false or biased data (n=29), and improper analogies or equivalences (n=19). Note that laypeople's responses were created on average 14 hours after the social media post. Here, MUSE only retrieved web pages published *before* the post (Methods).

## Methods

### Implementation Details of MUSE

**Informative image captioning.** We employed pretrained BLIP-2<sup>46</sup> for image captioning with “*A photo of*” as the prompt, Amazon Rekognition API ([aws.amazon.com/rekognition](https://aws.amazon.com/rekognition)) for celebrity recognition, and Amazon Textract API ([aws.amazon.com/textract](https://aws.amazon.com/textract)) for OCR. GPT-4 (gpt-4-0613) was leveraged with in-context learning to integrate image captioning, celebrity recognition, and OCR results into informative image descriptions. As examples, we selected eight images with quotes, photos, screenshots of posts, articles, and charts from social media, and manually generated their informative descriptions (see the example images in Supplementary Fig. S10 and the prompt in Supplementary Fig. S11). These example images do not appear in the dataset we used.

**Query generation.** We applied GPT-4 (gpt-4-0613) to generate queries. The prompt was “*Given a tweet, you are required to generate N different queries from the tweet for the Google search engine to get the most relevant web content to fact-check the tweet. If the given tweet is not informative enough to generate a query, you should answer "none".*”;  $N = 3$  for text-only tweets, and  $N = 5$  for tweets with images. The tweet information concatenated its textual content, informative descriptions of images (for tweets with images), time, and the poster’s name.

**Web search.** Google Programmable Search Engine ([programmablesearchengine.google.com](https://programmablesearchengine.google.com)) was utilized for text-only misinformation. We limited its search scope to the target publishers. The maximum number of retrieved web links with the same priority was set as 10. For misinformation with images, we used Google Reverse Image API provided by SerpApi ([serpapi.com/google-reverse-image](https://serpapi.com/google-reverse-image)). Since Google Reverse Image API does not have access to customizing sites to search, we started with collecting the first page of retrieval results by the reverse image search engine (i.e., the first ten retrieved web links) and selected the web pages from the target publishers. We set the maximum number of pages as five (i.e., the maximum number of retrieved web links as 50) and the maximum number of retrieved web links with the same priority as 10.

**Relevance of web pages to misinformation content.** First, we obtained the web content from each retrieved web link based on news-please, a generic and open-source web content extractor that works for a large variety of websites<sup>55</sup>. To compute the relevance between a piece of text-only misinformation and a retrieved web page, we first applied a pretrained Sentence-Transformer (msmarco-distilbert-base-tas-b)<sup>56</sup> to embed the misinformation (URLs and emojis were removed) and the web page’s main text. Then, we measured their relevance by the dot product of two embeddings, following the guidance from Reimers and Gurevych<sup>56</sup>. The web page was relevant to the misinformation only if their dot product was equal to or above a threshold value. To determine this threshold value, we randomly selected ten pieces of text-only misinformation excluded in MUSE’s evaluation, collected the top ten web pages for each piece of misinformation after searching the web, and manually checked their actual relevance and computed relevance scores to the misinformation. We set this threshold value as 90 such that the removed web pages were indeed irrelevant. For misinformation with images, we further adopted a pretrained Vision-Transformer (facebook/dino-vitb8)<sup>57</sup> to embed each image of the misinformation and the web page’s main image and measured their relevance by the cosine similarity of two embeddings. The web page was relevant to the misinformation only if the textual relevance was equal to or above 95 or the visual relevance was equal to or above 0.7. We determined the threshold values in the same way as for text-only misinformation, which ensured the selected web pages were indeed relevant.

**Credibility evaluation of publishers.** We used the professional human ratings from Media Bias/Fact

Check (MBFC, [mediabiasfactcheck.com](https://mediabiasfactcheck.com)) to determine the factuality and bias of web pages. MBFC is a widely accepted independent and transparent website offering a large-scale evaluation of more than 5,000 publishers<sup>58–60</sup>. It provides six factuality categories: “very high,” “high,” “mostly factual,” “mixed,”, “low,” and “very low” and 11 bias categories: “least biased,” “left-center,” “right-center,” “left,” “right,” “extremely left,” “extremely right,” “pro-science,” “questionable,” “satire,” and “conspiracy-pseudoscience” (see their definitions and statistics in Supplementary Table S1). MUSE only considered as references the web pages whose factuality was annotated as one of “very high,” “high,” and “mostly factual”, and bias was annotated as one of “least biased,” “left-center,” “right-center,” and “pro-science,” where “pro-science” publishers are defined as consisting of *least biased* legitimate science publishers (Supplementary Table S1). In this way, MUSE explicitly excluded moderately to strongly biased publishers (e.g., Daily Beast and Fox News). It also explicitly excludes the publishers whose factuality is mixed (e.g., Wikipedia and Twitter) or low, including those rejecting established scientific consensus on issues such as climate change or vaccines and identified as overt propaganda by reputable third-party evaluators (e.g., Infowars and Natural News; Supplementary Table S1). MUSE further divided the publishers considered as potential references into three priorities. High-priority publishers (e.g., CDC, Science, Pew Research, and Reuters; n=118) have “very high” factuality and are either “least biased” or “pro-science.” Of the remaining, publishers whose factuality is at least “high” were labeled medium priority (e.g., Britannica, Statista, Psychology Today, The Economist, The Dispatch, NPR, Propublica, and Know Your Meme; n=2,123), and publishers who do not have high- or medium priority were low priority (e.g., Forbes and USA Today; n=204).

**Evidence extraction.** We leveraged GPT-4 (gpt-4-0613) for evidence extraction. The prompt is “*Given an article: 1. Quote its paragraphs, at most two, that explicitly and completely refute the given tweet. 2. Quote its paragraphs, at most two, that implicitly refute the given tweet. Such paragraphs often provide the tweet’s context that can imply the tweet is cherry-picking by showing the full picture. If the article does not have such content or is irrelevant to the tweet, you should answer ‘none.’*” The article information included the article’s content and published date. The article’s content has the maximum number of characters, which we set as 20,000 considering gpt-4-0613’s context window is 8,192 tokens. The tweet information concatenated its textual content, informative image captions (for tweets with images), time, the poster’s name, the poster’s screen name, and the poster’s description.

**Response generation.** We utilized GPT-4 (gpt-4-0613) for response generation. The prompt is “*You are required to respond to a tweet, given some facts as references. Your response should satisfy all the following requirements: - Your response should explain where and why the tweet is or is not misinformed or potentially misleading. - You should prioritize the facts very close to the date the user tweeted, very recently, and listed at the beginning of the facts. - You should show the URLs that support your explanation. You should not number the URLs. - Your response should be informative and short. - Your response should start with ‘This tweet is.’*” The tweet information concatenated its textual content, informative image captions (for tweets with images), time, the poster’s name, the poster’s screen name, and the poster’s description. The facts listed every piece of extracted evidence with its source link and published date. The pieces of evidence were sorted by their publishers’ priorities (from highest to lowest). Pieces of evidence with the same priority were further sorted by their relevance to the tweet in descending order, which has been shown to increase GPT-4’s accuracy<sup>36</sup>.

## Evaluation

**Helpfulness classification.** The helpfulness of laypeople’s responses in Community Notes is positively associated with their helpfulness scores, normally distributed from -0.3 to 0.6 with an average score of 0.17 (standard deviation: 0.17; Supplementary Fig. S12; as of February 2023). We viewed laypeople’s responses

whose helpfulness scores are equal to or above 0.35 as having high helpfulness, as the average helpfulness score of these responses is 0.44, which is above 0.4—X’s suggested threshold value to differentiate helpful responses, often displayed on the corresponding tweets on X and visible to the public (Supplementary Fig. S6), from the others<sup>17</sup> (Supplementary Fig. S12). We considered laypeople’s responses whose helpfulness scores are in [0.05, 0.25) to be average helpfulness, as the average helpfulness score of these responses is 0.17, same as the average helpfulness score of all laypeople’s responses in Community Notes (Supplementary Fig. S12).

**Accuracy of social media posts.** We obtained the accuracy label of the tweets included in our evaluation based on their responses generated by MUSE and baselines along with the annotations of experts (specified later). Specifically, we selected the responses that identify a tweet’s (in)accuracies without mistakes. If a tweet has more than one such response, we further selected the response that has the highest overall quality score. Then, we determined a tweet’s accuracy by manually reviewing the corresponding response. We observed that 48% of the tweets are a combination of accurate claims and inaccurate or misleading claims, 46% are inaccurate or misleading, 3% are fully accurate and not misleading, and the remaining 3% cannot be determined are not unverifiable. Note that we neither presuppose the fine-grained accuracy labels of the tweets nor whether the tweets are misinformation in both MUSE’s design and evaluation.

**Response approaches.** We included laypeople, MUSE, and MUSE’s variants as the approaches evaluated in our study. For each tweet, laypeople have two responses: one has high helpfulness, and the other has average helpfulness. We further generated responses by MUSE. Note that laypeople’s responses were created in the past, where MUSE could potentially have an advantage by retrieving more recently published web pages. Therefore to have a fair comparison, we constrained MUSE to only retrieve older web pages. Responses from Community Notes range from seven minutes to three years (median: 14 hours) after the tweet was originally posted on social media. We generated one response by MUSE to each tweet by only retrieving web pages published thirty minutes *before* the creation time of the corresponding laypeople’s high-helpfulness response (Supplementary Fig. S13). We also had MUSE generate an additional response to each tweet by only retrieving web pages published thirty minutes before the creation time of the corresponding laypeople’s average-helpfulness response (Supplementary Fig. S13). To evaluate MUSE’s capability for immediately responding to potential misinformation, we finally generated one response where MUSE only retrieved web pages published before the post time of the corresponding tweet. Moreover, we generated one response to each tweet by GPT-4 (gpt-4-0613), which can be seen as a variant of MUSE that is not augmented by credibility-aware retrieval and vision-enabled, i.e., only has the step of response generation in Fig. 1d. For tweets with images, we included two more variants of MUSE: one is augmented by credibility-aware retrieval but not vision-enabled (denoted as MUSE\vision), and the other is vice versa (denoted as MUSE\retrieval). For MUSE\vision, it generated one response to each tweet by only retrieving web pages published thirty minutes before the creation time of the corresponding laypeople’s high-helpfulness response.

**Expert recruitment.** We worked with Hacks/Hackers ([hackshackers.com](https://hackshackers.com)), an international grassroots journalism organization, to recruit fact-checking and journalism experts. Hacks/Hackers helped send our recruitment materials to the people in its email list. Recruitment started in May 2023 and continued until August 2023. Among the 15 respondents, we selected the 12 respondents who had the highest experience in fact-checking or journalism and whose proficiency in English is at least fluent. Specifically, five (41.7%) of the selected respondents had 1–3 years, three (25%) had 4–6 years, one (8.3%) had 7–9 years, and three (25%) had 9+ years of experience in fact-checking or journalism. Nine (75%) of the selected respondents are native speakers, and three (25%) are fluent in English. The study was approved by the University of Washington’s Institutional Review Board (determined to be exempt; IRB ID STUDY00017831). We have

also obtained informed consent from all respondents.

**Study workflow.** We divided our study into two phases:

- **Phase I: Onboarding.** First, we scheduled and hosted an onboarding remote meeting with every participant. We explained our data annotation protocol (Supplementary Fig. S14-S20) and demonstrated the use of our web interface for data annotation (Supplementary Fig. S21). Every participant was asked to complete three annotation tasks (i.e., annotate the order-randomized responses made by various approaches to three tweets) after the meeting. Phase-I annotation was designed for the participants to enhance their understanding of the protocol and to familiarize themselves with the interface where they were required to provide explanations. Finally, we manually reviewed the explanations and sent each participant feedback to resolve any potential confusion and misunderstanding. Two participants who are native speakers in English dropped out of the study during Phase I. One of the participants had 1–3 years, and the other had 9+ years of experience in fact-checking or journalism. We removed their data from the final analyses and moved their annotation tasks to Phase II. There were 15 tasks completed in this training session; the same task can be assigned to more than one participant. In our final analyses, we excluded any data from this training session.
- **Phase II: Annotation.** We randomly divided the remaining ten participants into five groups, with two participants in each group. Every participant was randomly assigned 26 or 27 tasks for Phase II annotation. Seven, around 30% of these tasks, were the same as those assigned to another participant within the same group, which allowed us to evaluate inter-annotator agreement. The remaining 19 or 20 tasks were different from those assigned to the other participants. No participants dropped out of the study during Phase II. Finally, all the 232 tasks at this phase were completed and included in our final analyses. In our final analyses, the weight of each annotation for the tasks assigned to two participants was 0.5 and that for the tasks assigned to one participant was 1 to avoid bias towards the tasks assigned to two participants.

We compensated each participant who completed the study with a 450 USD Amazon gift card.

**Evaluation criteria.** Recruited experts evaluated each response from the following perspectives:

- **Quality of identifying and explaining (in)accuracies.** Such quality was measured by the response's 1) *explicitness*, i.e., whether the response explicitly, implicitly, or unclearly identifies and explains (in)accuracies; 2) *existence of (in)correct identifications*, i.e., whether the correction precisely identifies any (in)accuracies, with or without falsely identifying any inaccurate claims as accurate or an accurate claim as inaccurate; 3) *comprehensiveness of correct identifications*, which is five-scaled, ranging from no comprehensiveness (the response does not precisely identify any (in)accuracies in the tweet) to extremely high comprehensiveness (the response precisely identifies every (in)accuracy in the tweet); 4) *accuracy of explanations*, which is five-scaled, ranging from completely inaccurate to fully accurate; and 5) *informativeness of accurate explanations*, ranging from score 0 (the response does not provide any context in explaining the (in)accuracies) to 10 (the response provides completely sufficient context that helps a person understand why the content is inaccurate).
- **Quality of generated text.** Such quality was measured by the generated text's 1) *relevance to the tweet*, ranging from score 0 (the generated text is completely irrelevant to the responded tweet) to 10 (the generated text catches at least the most critical point in the responded tweet); 2) *factuality*, which is five-scaled, ranging from completely false, inaccurate, or unverifiable to completely factual and accurate; 3) *fluency*, i.e., whether the generated text had mistakes in the use of English, such

as capitalization errors, misspelled words, and sentence fragments<sup>61,62</sup>—the fluency had three levels: high (the generated text does not have any mistakes), medium (the generated text has minor mistakes barely causing confusion and reducing the text’s readability), and low (the generated text has mistakes leading to confusion and reducing the text’s readability); 4) *coherence* (logical consistency and correct and valid reasoning)<sup>61</sup>, i.e., whether the generated text is barely, partially, or fully coherent and logical; and 5) *toxicity*, i.e., whether the generated text is impolite, provoking, or biased.

- **Quality of references.** Such quality was measured by the reference’s 1) *reachability*, i.e., whether the web page is found; 2) *relevance to the generated text*, i.e., whether the web page is relevant to or supports the generated text; and 3) *credibility*, ranging from low (the page content and its publisher are both questionable), medium, high, to very high (the page content is backed up by facts with minimal bias, and its publisher always publishes high-quality information with minimal bias).
- **Overall quality of corrections.** Such quality was measured by taking all 13 aforementioned evaluation criteria into account, ranging from 0 (very low quality) to 10 (very high quality).

**Inter-annotator agreement.** We adopted the weighted Cohen’s kappa coefficient ( $\kappa$ ) to compute the agreement between two experts in every group, except for the toxicity of generated text, the fluency of generated text, and the relevance of references to the generated text because of their highly skewed distributions. Such distributions significantly underestimate the inter-annotator agreement and may cause the coefficient calculation to be not applicable<sup>63</sup>; for example, when two experts in a group annotated all pieces of generated text as not toxic. Instead, we reported the average observed agreement for the toxicity of generated text, the fluency of generated text, and the relevance of references to the generated text, which is 0.96 (vs  $\kappa$  is not applicable), 0.86 (vs  $\kappa = 0.02$ ), and 0.81 (vs  $\kappa = 0.02$ ), respectively. According to  $\kappa$ ’s interpretation<sup>64</sup>, experts achieved substantial agreement on the reachability of references (mean: 0.79). They achieved moderate agreement on the overall quality of responses (0.51), the informativeness of accurate explanations (0.50), the comprehensiveness of correct identifications (0.46), and the relevance of generated text to the responded responses (0.41). They achieved fair agreement on the accuracy of explanations (0.40), the factuality of generated text (0.39), the existence of (in)correct identifications (0.39), the explicitness of identifying and explaining (in)accuracies (0.34), the credibility of references (0.31), and the coherence of generated text (0.28), consistent with prior observations that even fact-checking experts can disagree on misinformation<sup>19,65</sup>.

**Impact of time.** We assessed the impact of time on MUSE’s performance from two perspectives. First, we compared three responses by MUSE to each tweet, which simulated responding the tweet under different starting times. Results in Supplementary Fig. S22 show that MUSE performs similarly (mean $\pm$ SD of the overall quality of responses:  $8.1\pm2.0$ ) when it starts responding the tweet right after appearing on social media, when it follows the starting times of laypeople who produce high-helpfulness responses (median: 13 hours after the tweet was posted; Supplementary Fig. S13), and when it follows the starting times of laypeople who produce average-helpfulness responses (median: 16 hours after the tweet was posted; Supplementary Fig. S13). Second, we separated tweets posted after September 2021 (n=207) from all tweets (n=232), considering that GPT-4 (gpt-4-0613)’s training data is up to September 2021. In other words, tweets posted before and in September 2021, along with its Community Notes data, might have been included in the training data of the GPT-4 that MUSE augments. If this information was available during GPT-4 training, it may lead to artificially inflated performance that is unlikely to generalize to future tweets, where such information is not available. Results in Supplementary Fig. S23 show that

MUSE performs stably (mean $\pm$ SD of the overall quality of corrections:  $8.1\pm2.0$ ) when responding to all tweets and when responding to tweets posted after September 2021, consistently outperforming GPT-4 and even high-helpfulness responses made by laypeople.

**Impact of retrieval and vision.** We have demonstrated that MUSE outperforms GPT-4, which can be seen as a variant of MUSE that is not augmented by credibility-aware retrieval and vision-enabled (Approach). Results in Supplementary Fig. S24 further demonstrate that both the retrieval and vision components are valuable. Overall, MUSE outperforms its variant that is not augmented by the retrieval by 25% and its variant that is not vision-enabled by 33% in the quality of generated responses.

**Importance of evaluation criteria.** We calculated the Spearman correlation coefficient between experts' annotation results on overall response quality and each criterion that specifies overall response quality. Results demonstrate that each of the 13 criteria we proposed and defined is significantly associated with overall response quality ( $p < 10^{-5}$ ;  $n = 1124$ ; Supp. Fig. S25). We further separated responses into two groups based on their overall quality annotated by experts. One group has relatively low quality (i.e., quality is in  $[0, 0.6]$ ), and the other has relatively high quality (i.e., quality is in  $(0.6, 1]$ ). The threshold of 0.6 is close to the median quality rating, leading to two similarly sized groups. Calculating the correlation coefficient for each group reveals the following insights (Supp. Fig. S25). First, to enhance the overall quality from low to medium, a response should accurately identify and explain inaccuracies, ensure the generated text is factual and not toxic, and use references relevant to the text. Second, to further improve the response's overall quality from medium to high, it is not sufficient to identify just an inaccuracy; instead, *every* inaccuracy in the misinformation needs to be identified accurately. Additionally, the correction should be explicit, fluent, and coherent, which enhances readability. The explanation should be informative and supported by credible references, thereby improving the response's transparency and trustworthiness. The generated text should be relevant to the misinformation, such as by capturing its multimodal content. Finally, it is always crucial to generate factual text that accurately explains inaccuracies.

**Runtime and cost.** The average runtime of MUSE in responding to a social media post was two minutes. Note that the experiments were conducted on 16G memory M1 CPU, running five parallel processes. The runtime could be further optimized with GPUs and more memory. Our focus in designing MUSE was in maximizing the quality of corrections. The total cost of MUSE in responding to a social media post was roughly 0.5 USD, almost all from the GPT-4 (gpt-4-0613) that MUSE augmented. In particular, evidence extraction cost the most, and increases with the number of retrieved web pages used to extract evidence and their content length, which is often substantial. We reduced the cost by removing the retrieved web pages with relatively low relevance to misinformation, which also helped reduce GPT-4's hallucinations assessed through qualitative evaluation (Approach; Supplementary Fig. S3). The cost could be further reduced significantly with an open-sourced LLM, such as Llama-3 (Discussion; Supp. Fig. S9).

## Data availability

Data used in this study are available at <https://github.com/Social-Futures-Lab/MUSE>. We comply with X/Twitter Terms of Service by only releasing the IDs of tweets. The experts' names are anonymized.

## Code availability

Code used for analyzing the study data is available at <https://github.com/Social-Futures-Lab/MUSE>. Source code of MUSE will be made available with publication.

## References

55. Hamborg, F., Meuschke, N., Breitinger, C. & Gipp, B. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, 218–223, DOI: [10.5281/zenodo.4120316](https://doi.org/10.5281/zenodo.4120316) (2017).
56. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992 (2019).
57. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660 (2021).
58. Weld, G., Glenski, M. & Althoff, T. Political bias and factualness in news sharing across more than 100,000 online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 796–807 (2021).
59. Zhou, X., Mulay, A., Ferrara, E. & Zafarani, R. ReCOVery: A multimodal repository for COVID-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3205–3212 (2020).
60. Bozarth, L., Saraf, A. & Budak, C. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 48–59 (2020).
61. Yuan, W., Neubig, G. & Liu, P. BARTScore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* **34**, 27263–27277 (2021).
62. Fabbri, A. R. *et al.* SummEval: Re-evaluating summarization evaluation. *Transactions Assoc. for Comput. Linguist.* **9**, 391–409 (2021).
63. Xu, S. & Lorber, M. F. Interrater agreement statistics with skewed data: Evaluation of alternatives to cohen's kappa. *J. Consult. Clin. Psychol.* **82**, 1219 (2014).
64. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Medica* **22**, 276–282 (2012).
65. Bhuiyan, M. M., Zhang, A. X., Sehat, C. M. & Mitra, T. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proc. ACM on Human-Computer Interact.* **4**, 1–26 (2020).

## Acknowledgements

We would like to thank Hacks/Hackers for advertising the study and helping with the recruitment. We also thank members of the UW Social Futures Lab and Behavioral Data Science Group for their suggestions and feedback. This work was supported in part by the National Science Foundation's Convergence Accelerator program under Award No. 49100421C0037. T.A. and A.S. were supported in part by the Office of Naval Research (#N00014-21-1-2154), NSF IIS-1901386, and NSF CAREER IIS-2142794.

## Author contributions

X.Z., A.X.Z., and T.A. designed the study. X.Z. led, and A.S. assisted in developing the model. A.S. built the web interface for evaluation. X.Z. prepared and analyzed the data. X.Z. drafted and all authors revised the manuscript. A.X.Z. and T.A. supervised the study.

## **Competing interests**

The authors declare no competing interests.

## **Additional information**

**Supplementary information** is available for this paper.

**Correspondence and requests for materials** should be addressed to Tim Althoff ([althoff@cs.washington.edu](mailto:althoff@cs.washington.edu)) and Amy X. Zhang ([axz@cs.uw.edu](mailto:axz@cs.uw.edu)).

## **Supplementary materials**

### **List of supplementary materials**

Table S1

Figures S1 to S25

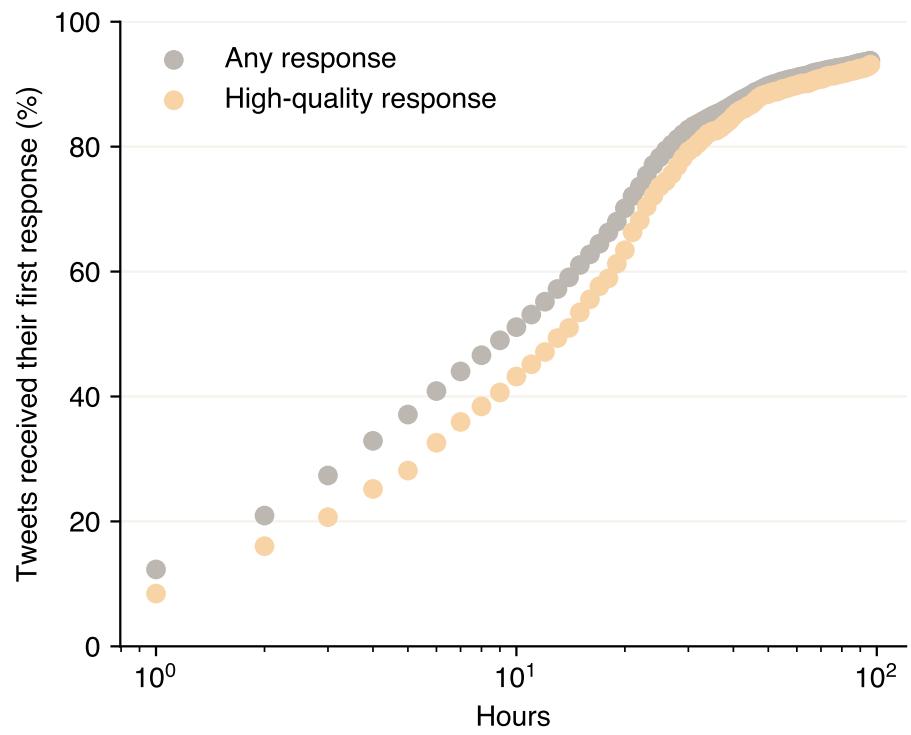
**Table S1.** Definitions and statistics of factuality and bias categories offered by Media Bias/Fact Check ([mediabiasfactcheck.com](http://mediabiasfactcheck.com)).

Category	Definition	# of sources
<b>Factuality:</b>		
- Very high:	The source is consistently factual, relies on credible information, promptly corrects errors, and has never failed any fact checks in news reporting or opinion pieces.	118
- High:	The source is mostly factual and uses mostly credible, low-biased, or high-factual sources. It corrects errors quickly and has failed only one news fact check and up to two op-ed fact checks.	2,313
- Mostly factual:	The source is generally accurate but may have a few uncorrected fact-check failures. It can fail up to three op-ed fact checks, especially if it is a low-volume site. While it may use biased sources occasionally, it mostly links to factual content. It is usually pro-science but may sometimes use misleading wording or offer alternative viewpoints. It is reasonably transparent and trustworthy most of the time, but caution is advised.	326
- Mixed:	The source may rely on improper sourcing or link to other biased or mixed-factual sources. It often has multiple failed fact checks and does not correct false information or lacks transparency, including the absence of a disclosed mission statement or ownership details. Sources rejecting established scientific consensus on issues such as climate change or vaccines will receive this rating or lower. Sources identified as overt propaganda or designated as hate groups by reputable third-party evaluators will receive this rating or lower due to their inherent bias and potential spread of misleading information.	1,437
- Low:	The source is often unreliable and should be fact-checked for fake news, conspiracy theories, and propaganda.	677
- Very low:	The source is almost always unreliable and should always be fact-checked for intentional misinformation.	252
<b>Total: 5,123</b>		
<b>Bias:</b>		
- Least biased:	The source has minimal bias and uses very few loaded words (i.e., wording that attempts to influence an audience by using an appeal to emotion or stereotypes). It is factual and usually sourced.	1,054
- Left-center:	The source has a slight to moderate liberal bias. It often publishes factual information that utilizes loaded words to favor liberal causes. It is generally trustworthy for information but may require further investigation.	850
- Right-center:	Similar to the definition of left-center bias but replacing liberal with conservative.	492
- (Extremely) left:	The source is moderately to strongly biased toward liberal causes through story selection or political affiliation. It may utilize strong loaded words, publish misleading reports, and omit reporting of information that may damage liberal causes. It may be untrustworthy.	402
- (Extremely) right:	Similar to the definition of (extremely) left bias but replacing liberal with conservative.	314

- Pro-science:	The source consists of legitimate science or is evidence-based through the use of credible scientific sourcing. Legitimate science follows the scientific method, is unbiased, and does not use emotional words. The source also respects the consensus of experts in the given scientific field and strives to publish peer-reviewed science. It may have a slight political bias but adheres to scientific principles.	189
- Conspiracy-pseudoscience:	The source may publish unverifiable information not always supported by evidence. It may be untrustworthy for credible or verifiable information, so fact-checking and further investigation are recommended on a per-article basis when obtaining information from it.	433
- Questionable:	The source exhibits one or more of the following: extreme bias, consistent promotion of propaganda or conspiracies, poor or no sourcing of credible information, a complete lack of transparency, or fake news (i.e., the deliberate attempt to publish hoaxes or disinformation for profit or influence). It may be very untrustworthy and should be fact-checked on a per-article basis.	1,390
- Satire:	The source exclusively uses humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, particularly in the context of contemporary politics and other topical issues. It does not attempt to deceive.	148

**Total: 5,272**

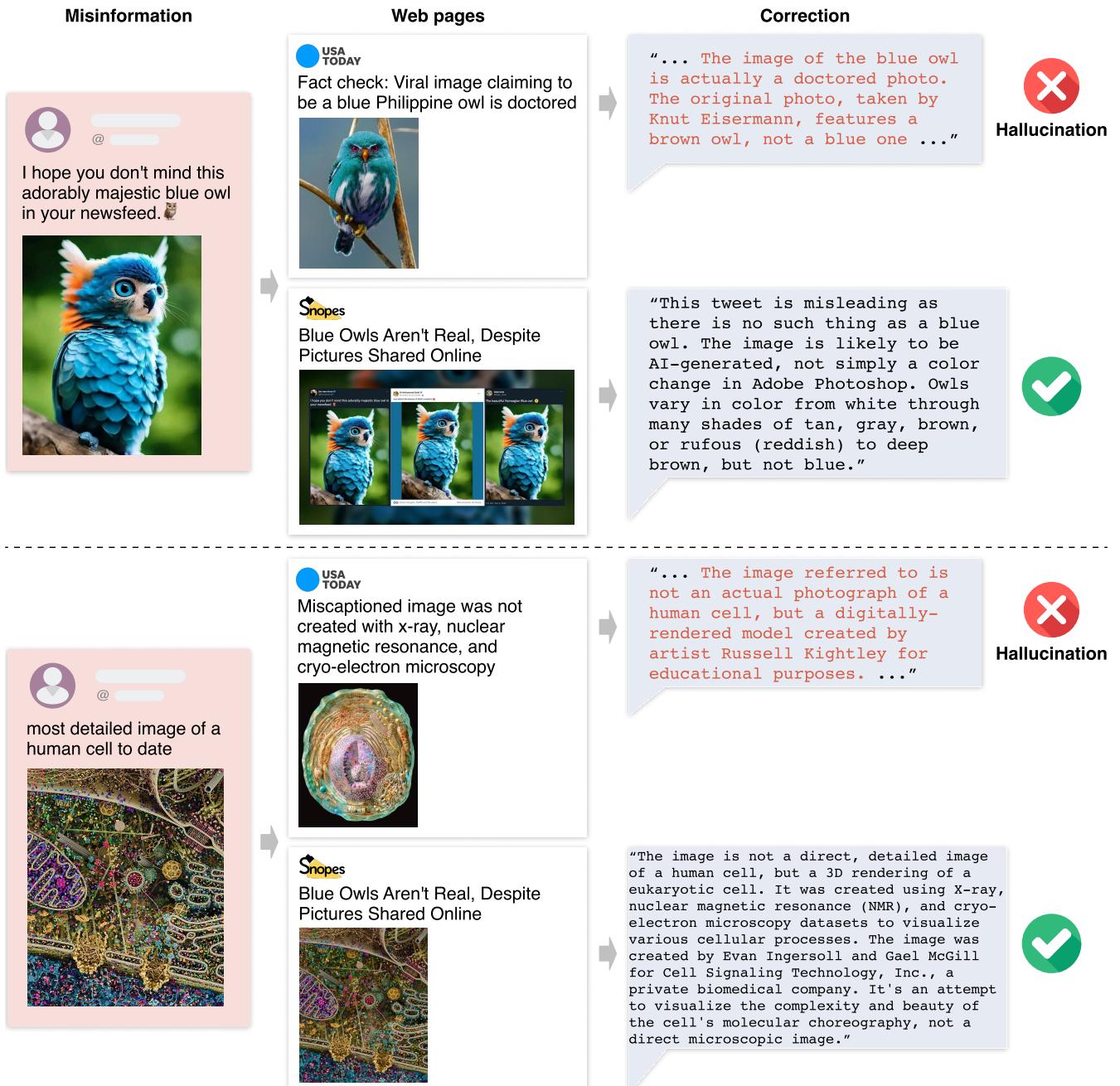
---



**Figure S1.** Distribution of potential misinformation in X Community Notes (as of February 2023) that received its first response (gray) or first high-quality response (orange) within a certain amount of time.

Misinformation	Number of generated queries = 1	Number of generated queries > 1
 @ [redacted] <b>MAJOR BREAKING:</b> The Pentagon confirms that at least THREE Chinese spy balloons flew over the US during Trump's presidency and he HID THEM from the public and never shot them down.	<p>Q Pentagon confirmation on Chinese spy balloons over US during Trump's presidency</p>	<p>Q Pentagon confirmation on Chinese spy balloons over US during Trump's presidency</p> <p>Q <b>Trump hiding information about Chinese spy balloons</b></p> <p>Q <b>Incidents of Chinese spy balloons over US not being shot down</b></p>
 @ [redacted] Paul Whelan was dishonorably discharged for larceny, dereliction of duty, lying, social security fraud, and writing bad checks. Never forget: Many of the people calling Whelan a hero thought <b>George Floyd deserved what happened to him for a fake \$20 bill</b> .	<p>Q Paul Whelan dishonorable discharge reasons</p>	<p>Q Paul Whelan dishonorable discharge reasons</p> <p>Q Paul Whelan larceny, dereliction of duty, lying, social security fraud, and writing bad checks</p> <p>Q <b>George Floyd fake \$20 bill incident details</b></p>
 @ [redacted] holy fucking shit, if <b>an actor accidentally discharging a gun while rehearsing for a movie</b> has you shitting bricks, wait until you hear about the teenager who asked his mommy to drive him across state lines so he could deliberately fire into a crowd of protestors	<p>Q Teenager driven by mother across state lines to fire into crowd of protestors</p>	<p>Q <b>Actor accidentally discharges gun on movie set</b></p> <p>Q Teenager fires into crowd of protestors</p> <p>Q Teenager driven by mother to protest across state lines</p>

**Figure S2.** Examples that show that generating multiple queries helps decompose a post, which may have multiple claims that each needs to be verified, whereas generating one query may overlook some of them and hence lead to not comprehensive identifications of (in)accuracies. Bold text: the verification-needed claims that are overlooked when generating one query but captured when generating more than one query.



**Figure S3.** Examples that show how retrieved web pages with relatively low relevance to potential misinformation can promote LLM (in this case, GPT-4) hallucinations when generating responses.

### Misinformation



 **Elon Musk**  @elonmusk

Twitter has had a massive drop in revenue, due to activist groups pressuring advertisers, even though nothing has changed with content moderation and we did everything we could to appease the activists. Extremely messed up! They're trying to destroy free speech in America.

9:28 AM · 11/4/22 · Twitter for iPhone



### Image captioning

a weather map on a tv screen

### Informative image captioning

A weather map displayed on a TV screen showing wind chill temperatures in various locations such as Waukegan, Woodstock, Rockford, DeKalb, O'Hare, Lakefront, Aurora, Midway, Gary, Joliet, Chicago Heights, Morris, Merrillville, and Kankakee. The map indicates extreme cold temperatures with conditions of blowing snow.

elon musk's tweet on social media

A screenshot of a post by Elon Musk on Twitter, "Twitter has had a massive drop in revenue, due to activist groups pressuring advertisers, even though nothing has changed with content moderation and we did everything we could to appease the activists. Extremely messed up! They're trying to destroy free speech in America." The post was made at 9:28 AM on 11/4/22.

a young person sitting on a chair

A photo of a young Aaron Swartz sitting on a chair.

**Figure S4.** Examples of informative image captions, which augment image captions with names of visually represented celebrities and embedded text (see Methods for implementation details).

### Social media post



@ [REDACTED]

The strongest year of new small business applications in American history was 2021. The second? 2022.

Jan 23, 2023

### MUSE-generated response



MUSE  
@MUSE

This tweet is accurate. The first two years of President Biden's term saw the highest number of new small business applications on record, with over 10 million new businesses created. This surge in entrepreneurship was driven by the broader economic recovery and the Biden-Harris Administration's bottom-up approach to economic recovery. However, it's important to note that only about a third of these applications are for businesses likely to hire employees.

<https://www.forbes.com/sites/rhettbuttle/2023/01/17/record-small-business-applications-show-promise-for-economy-and-biden-harris-legislative-accomplishments/>

<https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/25/fact-sheet-the-new-small-business-boom-under-the-biden-harris-administration/>

<https://www.npr.org/2022/01/12/1072057249/new-business-applications-record-high-great-resignation-pandemic-entrepreneur>



@ [REDACTED]

Some people on here seem to think that the UK has a constitution. No it doesn't have a codified constitution because otherwise a referendum like on the EU wouldn't have taken place. Take it from someone who teaches Public Law.

Feb 2, 2023



MUSE  
@MUSE

This tweet is partially correct. The UK does not have a codified constitution, which allows for referendums on constitutional reforms without a binding rule. However, while the UK was a member of the EU, it was effectively under a codified constitution - the constitution of the EU. This constrained both the government and Parliament. The question of whether the UK should adopt a codified constitution is a topic of ongoing debate.

<https://www.instituteforgovernment.org.uk/article/comment/referendum-and-british-constitution-strange-bedfellows>

<https://foreignpolicy.com/2019/10/08/brexit-means-britain-needs-a-constitution/>

<https://pubmed.ncbi.nlm.nih.gov/34876878/>



@ [REDACTED]

January 6th wasn't an insurrection! RT & pass it on.



Jan 6, 2023



MUSE  
@MUSE

This tweet is misleading. The events of January 6th at the Capitol were indeed characterized as an insurrection due to the violent uprising against the government, which is the definition of an insurrection. The crowd that stormed the Capitol was dissenting against the election outcome and sought to obstruct Congress' certification of the election. The violent nature of the event is evidenced by the fact that hundreds of rally attendees smashed their way into the Capitol building, resulting in five deaths and at least 140 injuries. This is in contrast to the peaceful demonstration in Topeka on the same day, which was orderly and legal.

<https://time.com/6137604/history-insurrection-jan-6/>

<https://www.buzzfeednews.com/article/davidmack/capitol-meemaw-meme-topeka-kansas>

**Figure S5.** Examples of MUSE-generated responses to accurate, partially accurate, and factually correct but misleading content on social media.



...

January 6th wasn't an insurrection!

RT & pass it on.



Readers added context they thought people might want to know

This picture was taken at a separate, peaceful rally in Kansas on the same day as the January 6 attack in Washington D.C. It shows the Kansas state house, not the conditions on the ground in the US Capitol.

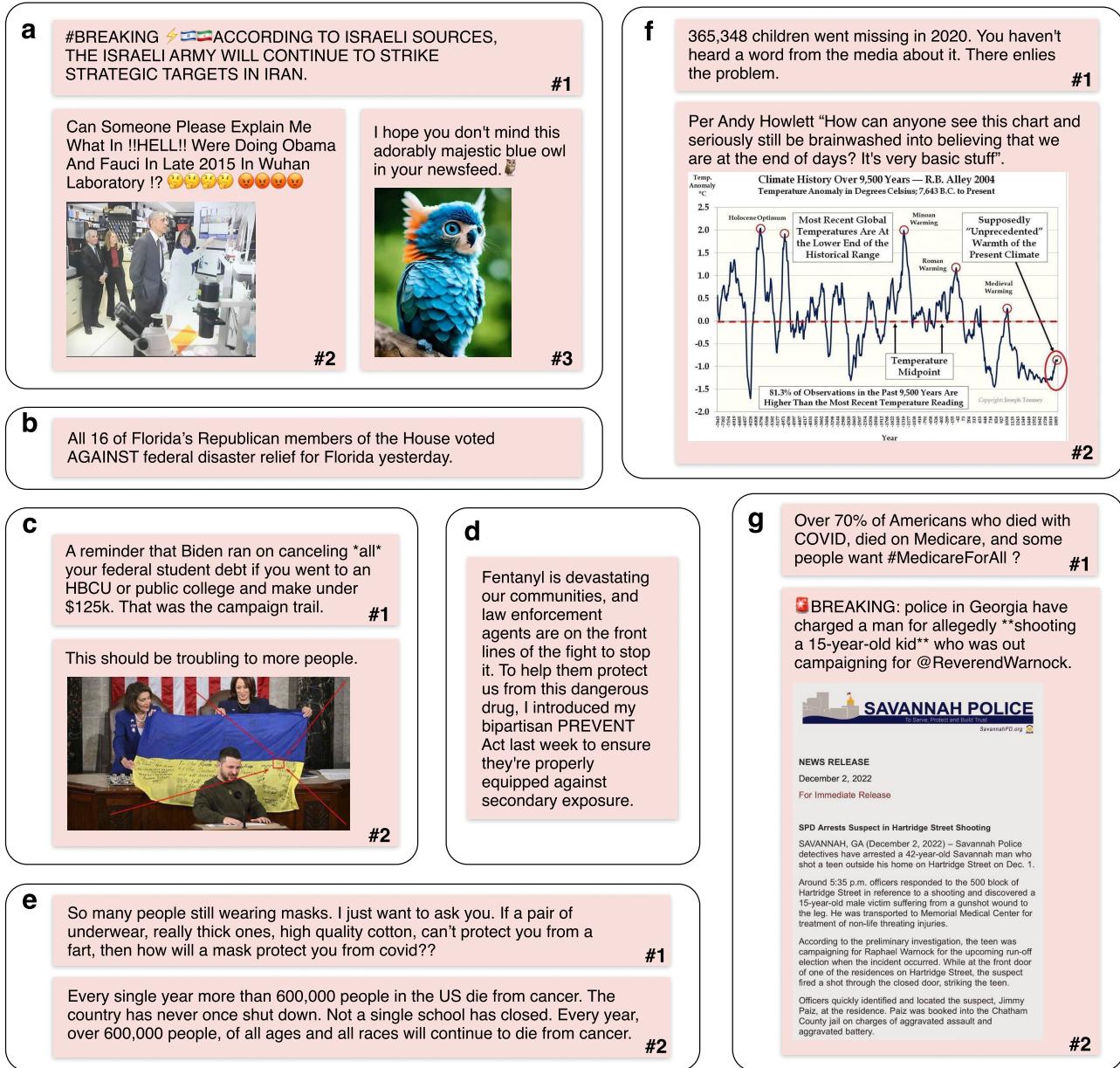
[buzzfeednews.com/article/davidm...](http://buzzfeednews.com/article/davidm...)

Do you find this helpful?

Rate it

Context is written by people who use X, and appears when rated helpful by others. [Find out more](#).

**Figure S6.** An example of a high-helpfulness response from Community Notes displayed on the corresponding tweet and visible to the public.



**Figure S7.** Examples of social media content that uses various tactics to make it or part of it false or misleading. These tactics include **a:** *Fabricating* a news event (#1), a story with an authentic image (#2, also related to conspiracy theory), and an image with AI (#3). **b:** *Lacking context* regarding the motivation behind behavior. Here, it is the consideration that the bill contained no funding for Florida, without which could mislead readers. **c:** *Misinterpreting or misrepresenting* a plan, where Biden did not promise to cancel all the debt (#1), or symbol, which is not Nazi SS runes but a shorthand for the 46th Separate Airmobile Brigade (#2). **d:** *Using loaded language*. **e:** *Improperly analogizing or equating* masks blocking respiratory droplets and underwear blocking gas molecules (#1), and COVID-19 and cancer (#2). **f:** *Presenting false, partial, or biased data*. 365348 includes multiple reports for the same child and runaways who may not be considered missing because of their guardians know their whereabouts (#1). The data ends in 1885 and is from a specific high elevation site in Greenland, not global temperatures (#2). **g:** *Implying false or oversimplified causation*. Many COVID-19 deaths occurred among Medicare beneficiaries because Medicare primarily serves the groups who are at greater risk of adverse outcomes from COVID-19 (#1). The boy was shot when—but not because—he was campaigning for Raphael Warnock (#2). Note that one content may apply more than one tactic intentionally or accidentally.

### Social media post

FBI Director Christopher Wray paid \$2,500 a night for prostitutes in Davos.

Mitch McConnell said he would resign if he doesn't become the Senate Majority Leader in November. I think Democrats should run this message in every state. Voters despise Moscow Mitch all across the country. If you want to end McConnell's political career, vote for the Democrats!

1200 dead of COVID yesterday in Florida. Not the total for a week or a month, but ONE SINGLE DAY.

The future republicans are working towards



⚡ Russian Chief of the General Staff, Valery Gerasimov, reportedly fired today. Gerasimov was very highly regarded, the most important military leader of the past generation, & the architect of today's Russian Armed Forces. He is served as the head of the military since 2012.



### MUSE-generated response

There's no credible source or evidence provided to support the claim about FBI Director Christopher Wray. It's important to verify such serious allegations from reliable sources before sharing.

There's no public record of Mitch McConnell stating he would resign if he doesn't become the Senate Majority Leader. It's important to verify such claims from reliable sources before sharing. Misinformation can lead to uninformed decisions.

According to the Florida Department of Health, the highest single-day death toll was 276 on August 11, 2021. The number mentioned **seems** to be a cumulative total over several days, not a single day count. Always refer to official health department data for accurate information.

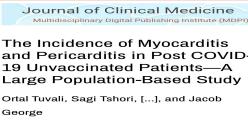
This tweet is potentially misleading as it **seems** to be based on hypothetical scenarios. The headlines in the attached image are from a future date and appear to be speculative, not factual. It's important to base political discussions on current, verifiable information to avoid misinformation.

This tweet is potentially misleading without further context or confirmation. As of now, there are no official reports or statements from the Russian government confirming that Valery Gerasimov has been fired. It's crucial to verify such significant information from reliable sources before sharing.

**Figure S8.** Examples of MUSE-generated responses to social media posts that lack evidence to verify.

#### Social media post

So Covid vaccines may cause myocarditis, but Covid itself causes myocarditis too, right? WRONG! A large-scale Israeli study of 196,992 unvaccinated adults after Covid infection showed that "Covid infection was not associated with either myocarditis or pericarditis."



were associated with pericarditis. Post COVID-19 infection was not associated with either myocarditis (aHR 1.08; 95% CI 0.45 to 2.56) or pericarditis (aHR 0.53; 95% CI 0.25 to 1.13). We did not observe an increased incidence of neither pericarditis nor myocarditis in adult patients recovering from COVID-19 infection.

Keywords: COVID-19, myocarditis, pericarditis

Louis DeJoy plans to slash 50,000 USPS jobs. Why in the world does Trump's postmaster general still have his job? And how is this not a massive national scandal?

#### Social media post

While we're distracted by pronouns, UFOs, & Woke Jesus, it's important to remember that over 100,000 gallons of burning vinyl chloride is destroying the air, water, soil, & living creatures of Palestine, Ohio. This is one of the biggest ecological disasters in American history.



#### Response by GPT4-based MUSE

While the Israeli study cited does indeed find no increased incidence of myocarditis or pericarditis in post-COVID-19 patients, other studies have found a higher risk of myocarditis in COVID-19 patients compared to those without COVID-19. For instance, a CDC report found that patients with COVID-19 had, on average, 15.7 times the risk for myocarditis compared to patients without COVID-19. However, the absolute risk difference was small due to the low risk of myocarditis in both groups. It's important to consider all available research and understand that results can vary based on different study designs and populations.

<https://pubmed.ncbi.nlm.nih.gov/35456309/>

<https://www.cdc.gov/mmwr/volumes/70/wr/mm7035e5.htm>

#### Response by Llama3-based MUSE

While the study cited in the tweet is correct that post-COVID-19 infection was not associated with either myocarditis or pericarditis in the specific population studied, this finding contradicts other recent studies that found a significant association between COVID-19 infection and myocarditis. For example, a CDC study published in 2021 found that patients with COVID-19 had nearly 16 times the risk for myocarditis compared with patients who did not have COVID-19, and risk varied by sex and age.

<https://www.cdc.gov/mmwr/volumes/70/wr/mm7035e5.htm>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9025013/>

While it's true that Louis DeJoy has plans to reduce USPS positions, it's not a direct job cut but rather a reduction through voluntary early retirements and expected attrition over the next few years. The aim is to reach a break-even point for USPS, which has been facing financial difficulties. The number of positions to be reduced represents only a quarter of the attrition the agency expects to see over a similar period. DeJoy has also stated that USPS will remain in a hiring posture in the coming years. As for why DeJoy still holds his position, the Postmaster General is not a political appointee and cannot be removed by a president. The power to remove the Postmaster General lies with the Postal Service's governing board, whose members were all appointed by President Trump.

<https://www.govexec.com/workforce/2022/07/usps-plans-slash-50000-positions-coming-years-reach-break-even-point/375096/>

<https://abcnews.go.com/Politics/controversial-postal-chief-poised-remain-biden-administration/story?id=74469137>

<https://www.reuters.com/business/us-postal-service-will-offer-early-retirements-consolidate-postal-districts-2021-03-03/>

The claim that Louis DeJoy plans to slash 50,000 USPS jobs is misleading. According to DeJoy's statement at an American Enterprise Institute event, the net reductions of 50,000 positions represent only one-quarter of the attrition the agency expects to see over a similar period, and USPS will remain in a hiring posture in the coming years. Additionally, DeJoy stated that the agency will be able to manage any headcount reduction through normal attrition, which results from employees retiring or leaving the agency.

<https://www.govexec.com/workforce/2022/07/usps-plans-slash-50000-positions-coming-years-reach-break-even-point/375096/>

<https://federalnewsnetwork.com/workforce/2022/07/usps-may-need-50k-fewer-employees-under-break-even-plan-dejoy-says/>

#### Response by GPT-4

There's no record of an ecological disaster involving 100,000 gallons of burning vinyl chloride in Palestine, Ohio. Vinyl chloride is indeed a hazardous substance, but there's no evidence of such an incident in the mentioned location. Please ensure to verify information before sharing.

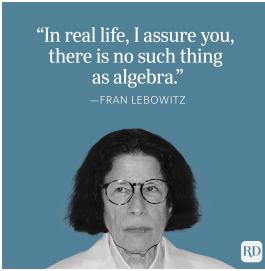
#### Response by Llama-3

The location of the ecological disaster is actually East Palestine, Ohio, not Palestine, Ohio. Palestine is not a town in Ohio. The rest of the tweet regarding the scale of the disaster appears to be accurate, highlighting the importance of attention to this environmental crisis.

**Figure S9.** Comparison between responses generated by GPT-4, Llama-3, and their based MUSE. We observed that GPT4-based MUSE and Llama3-based MUSE frequently generate similar responses to social media content (see the top example). GPT4-based MUSE occasionally generates more comprehensive corrections than Llama3-based MUSE (see the middle example). Note that the knowledge cutoff for Llama3 is December 2023, whereas that of GPT-4 is September 2021. In other words, Llama-3's training data include more recent events and knowledge than GPT-4's (see the bottom example, where the post was posted in February 2023, and Llama-3 more accurately identifies its falsehood than GPT-4).



(a)



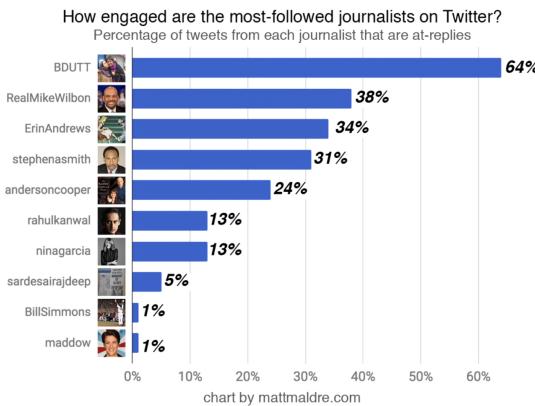
(b)



(c)



(d)



(e)



Yahoo Finance ✨  
@YahooFinance

Typically, the stock market bottoms four to five months before a recession ends, but RBC's research details that it has bottomed as early as nine months before the end of a recession.



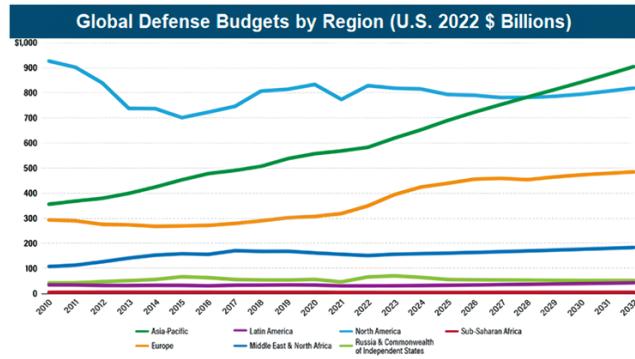
finance.yahoo.com  
Could the stock market power through a  
recession? 'This would be rare.'

09:57 · 22/4/2023 · 4,011 Views

1 Retweet 1 Quote 5 Likes 1 Bookmark



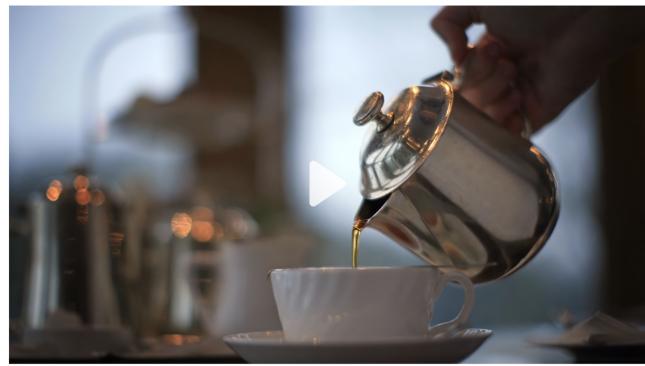
(g)



(f)

## New research reveals how coffee and tea can affect risk of early death for adults with diabetes

By Sandee LaMotte, CNN  
Updated 7:01 PM EDT, Wed April 19, 2023



(h)

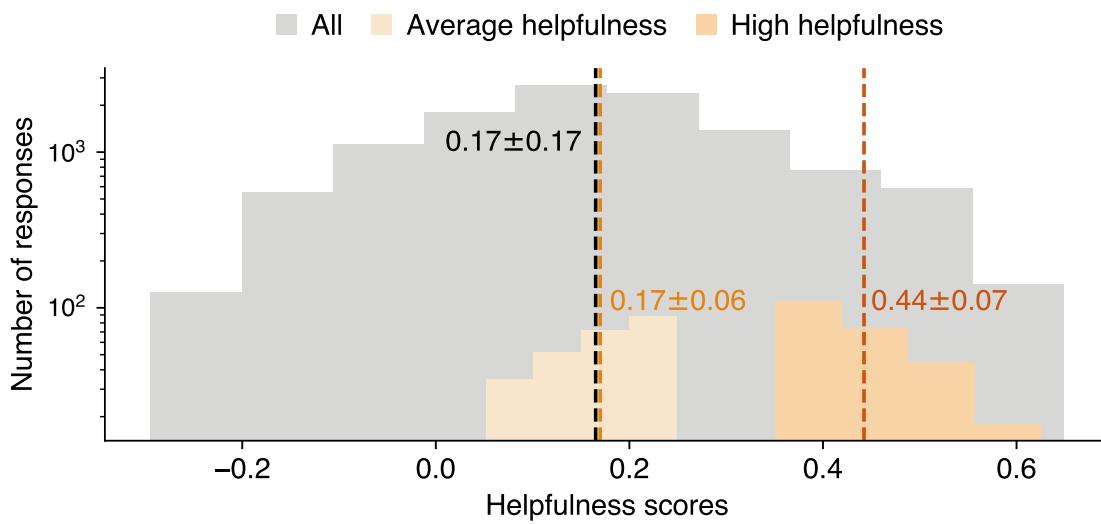
**Figure S10.** Images as examples used for informative image captioning with in-context learning.

```

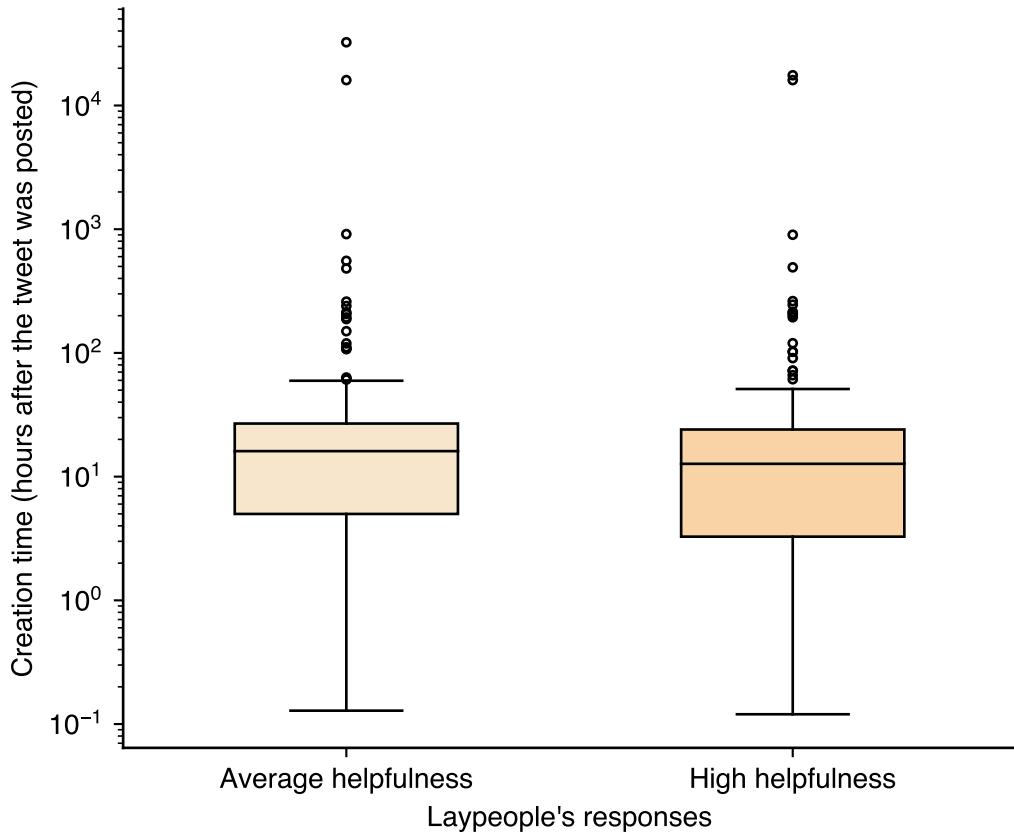
1  Describe an image in an informative way. Your description should be only based on the given {short
   ↵ caption}, {name of each person}, and {raw text}. If the image is from social media, you should
   ↵ start with "A screenshot of". If the image is a quote from someone, you should start with "A
   ↵ quote from" followed by this person's name if there is any, then by the quoted text. If the image
   ↵ is an article, you should start with "An article". If the image is a photo, you should start with
   ↵ "A photo of". If the image is a map, you should start with "A map of". {raw text} may contain
   ↵ nonsense data that are unnecessarily included in the image description; however, {name of each
   ↵ person} is not, and if the concept in {raw text} has a conflict with that in {short caption}
   ↵ (e.g., "Robbie Lemos" versus "robbie leems" shown later), {raw text} is often the right one.
2  short caption: {a woman with glasses and a quote that says, in real life, i assure you there is no
   ↵ such thing as algebra}
3  name of each person: {Fran Lebowitz}
4  raw text: {"In real life, I assure you, there is no such thing as algebra."}
5  image description: {A quote from Fran Lebowitz, "In real life, I assure you, there is no such thing
   ↵ as algebra."}
6  short caption: {two men in suits}
7  name of each person: {Jim Caviezel, Michael Emerson}
8  raw text: {}
9  image description: {A photo of Jim Caviezel and Michael Emerson in suits}
10 short caption: {robbie leems on twitter}
11 name of each person: {}
12 raw text: {Robbie Lemos @RobbieLemos 1d I'd like to congratulate my dear friend Deep Mind on a
   ↵ wonderful 1st day at work today at Google. Just in time for #EarthDay2023, cheers brother! 1 2
   ↵ 3,790}
13 image description: {A screenshot of a post of Robbie Lemos, "I'd like to congratulate my dear friend
   ↵ Deep Mind on a wonderful 1st day at work today at Google. Just in time for #EarthDay2023, cheers
   ↵ brother!" The post was posted on Twitter.}
14 short caption: {a moose}
15 name of each person: {}
16 raw text: {Yahoo Finance @YahooFinance Typically, the stock market bottoms four to five months before
   ↵ a recession ends, but RBC's research details that it has bottomed as early as nine months before
   ↵ the end of a recession. finance.yahoo.com Could the stock market power through a recession? 'This
   ↵ would be rare.' 09:57 22/4/2023 3,4,011 Views 1 Retweet 1 Quote 5 Likes 1 Bookmark}
17 image description: {A screenshot of a post from Yahoo Finance, "Typically, the stock market bottoms
   ↵ four to five months before a recession ends, but RBC's research details that it has bottomed as
   ↵ early as nine months before the end of a recession." The post shared an article from
   ↵ finance.yahoo.com claiming, "Could the stock market power through a recession? 'This would be
   ↵ rare.'" with a picture of a moose. The post was posted at 09:57 22/4/2023.}
18 short caption: {a person pouring tea into a cup}
19 name of each person: {}
20 raw text: {New research reveals how coffee and tea can affect risk of early death for adults with
   ↵ diabetes By Sandee LaMotte, CNN Updated 7:01 PM EDT, Wed April 19, 2023 f The health benefits of
   ↵ tea 01:10 - Source: CNN}
21 image description: {An article claiming, "New research reveals how coffee and tea can affect risk of
   ↵ early death for adults with diabetes." It attached a picture of a person pouring tea into a cup.
   ↵ It was written by Sandee LaMotte, published by CNN, and updated at 7:01 PM EDT, Wed April 19,
   ↵ 2023.}
22 short caption: {two people standing next to each other with the words love is blind}
23 name of each person: {Nick Lachey}
24 raw text: {"Love Is Blind\" co-host faceplants with a regressive line of questioning Hayley Miller
   ↵ MSNBC DAILY MSNBC}
25 image description: {An article claiming, "'Love Is Blind' co-host faceplants with a regressive line
   ↵ of questioning." It attached a picture of Nick Lachey and another person standing next to each
   ↵ other. It was written by Hayley Miller and published by MSNBC.}
26 short caption: {a bar graph that shows how engaged are the most followed journalists on twitter}
27 name of each person: {Rahul Kanwal}
28 raw text: {How engaged are the most-followed journalists on Twitter? Percentage of tweets from each
   ↵ journalist that are at-replies BDUTT 64%}
29 image description: {A bar graph showing how engaged the most followed journalists, including Rahul
   ↵ Kanwal, are on Twitter through the percentage of tweets from each journalist that are at-replies.
   ↵ The chart was made by mattmaldre.com.}
30 short caption: {a graph showing the global defense budget by region}
31 name of each person: {}
32 raw text: {Global Defense Budgets by Region ($ Billions) $1,000 800 600 400 200 0 2020 2021 2022 2023
   ↵ 2024 2025 Asia-Pacific Latin America North America Sub-Saharan Africa Europe Middle East & North
   ↵ Africa Russia & Commonwealth of Independent States Source: Aviation Week}
33 image description: {A graph showing the global defense budget by region. It is from Aviation Week.}
34 short caption: {[IMAGE_CAPTION]}
35 name of each person: {[CELEBRITIES]}
36 raw text: {[OCR]}
37 image description:

```

**Figure S11.** LLM prompt for informative image captioning.



**Figure S12.** Distribution of helpfulness scores of laypeople's responses in X Community Notes. All: All laypeople's responses in Community Notes. Average helpfulness: Laypeople's responses in Community Notes identified with average helpfulness and used in our study. High helpfulness: Laypeople's responses in Community Notes identified with high helpfulness and used in our study. For  $x \pm y$ ,  $x$ : mean,  $y$ : standard deviation. Community Notes data are regularly updated; ours are up until February 12, 2023.



**Figure S13.** Distribution of creation times of laypeople's responses in X Community Notes used in our study.  
Median of the creation time of laypeople's average-helpfulness responses: 16 hours after the tweet was posted.  
Median of the creation time of laypeople's high-helpfulness responses: 13 hours after the tweet was posted.

# Misinformation Response Study

Please evaluate and compare responses to misinformed or potentially misleading tweets from various aspects, such as factuality.

## Content Warning

The study may contain tweets with but not limited to abusive language, which may be disturbing you. If you have concerns or questions, please get in touch with us at [xzhou@cs.uw.edu](mailto:xzhou@cs.uw.edu) later!

## Prerequisite

To participate in this study, you should have a decent understanding of fact-checking and media bias.

## Notes before Starting

1. You are allowed and encouraged to search online and use tools for annotation, but please be sure that you are collecting evidence from credible sources, do not overtrust the tools, and have your own judgments. Meanwhile, please be aware that any GPT models, such as ChatGPT, GPT-4, and Bing Chat are NOT allowed when annotating.
2. Each response has a corresponding UTC time stamp when it was created. Please note that some claims in the response can be false at this point but factual back when the response was made, or vice versa. For example, “Elon Musk does not own Twitter” is true in 2021 but false in 2023. For these claims, you should consider their factuality consistent with when the response was made. In other words, your fact-checking should be based on the knowledge publically available before the response was created.
3. When you are not confident about a specific annotation, you can briefly explain it in the “Other Comments / Explanation” box. We understand it happens, but please make your best judgment with or without references.
4. Please be objective and politically neutral when annotating.

**Figure S14.** Annotation instructions (page 1/7, continued on the next page).

5. Please use your computer (laptop or desktop, rather than mobile device) for the annotation.

6. Any questions? Do not hesitate to contact us at [xzhou@cs.uw.edu](mailto:xzhou@cs.uw.edu)!

## What Will You Do?

You will be shown **26 or 27** tweets that can be misinformed or potentially misleading. For each tweet, you will be shown several responses — the number can vary from three to seven, with **four** as an average — that are supposed to be **corrections** in response to the tweet. In other words, each response aims to **explain where and why the tweet is misinformed or potentially misleading**. Each response consists of text as explanations and/or links as references.

You will be asked to evaluate various aspects of the responses that reflect how high-quality the explanation is. Specifically, you will need to answer the following questions for each response:

**Q1) What's the clarity of the response in identifying and explaining where and why the tweet is misinformed or potentially misleading?** Your answer should be one of the following options:

- A. The response **explicitly identifies and explains** where and why the tweet is misinformed or potentially misleading (regardless of whether the identification and explanation are correct). A typical example of such expressions can be, "Though it is true that X, the tweet is misinformed by claiming that Y because Z", where X and Y are from the tweet and Z is the explanation.
- B. Given the response, it is **hard to tell** where and why the tweet is misinformed or potentially misleading.
- C. Somewhere between A and B. For example, the response may only implicitly identify and explain where and why the tweet is misinformed or potentially misleading.

**Q2) Does the response correctly identify where the tweet is misinformed or potentially misleading?** Your answer should be one of the following options:

- A. Yes. The response **correctly identifies at least one place** in the tweet that is misinformed or potentially misleading. The response may overlook the others, and the correctly identified place may not be the critical point of the tweet. However, the response **does not misidentify**, i.e., explicitly claim where the tweet should be misinformed or potentially misleading as accurate or factual or vice versa.

**Figure S15.** Annotation instructions (page 2/7, continued on the next page).

- B. No. The response **doesn't correctly identify** *any place* in the tweet that is misinformed or potentially misleading.
- C. Somewhere between A and B.

If your answer to Q2 was either A or C, please answer Q2.1 and Q2.2 below.

**Q2.1) What's the comprehensiveness of the response in correctly identifying where the tweet is misinformed or potentially misleading?** Your answer should be one of the following options:

- A. The response is of **extremely high comprehensiveness**, meaning it correctly identifies **every** place in the tweet that is misinformed or potentially misleading.
- B. The response is of **high comprehensiveness**, meaning it correctly identifies **most** places in the tweet that is misinformed or potentially misleading.
- C. The response is of **medium comprehensiveness**, meaning it correctly identifies **half** places in the tweet that is misinformed or potentially misleading.
- D. The response is of **low comprehensiveness**, meaning it correctly identifies **few** places in the tweet that is misinformed or potentially misleading.
- E. The response is of **no comprehensiveness**, meaning it correctly identifies **no** places in the tweet that is misinformed or potentially misleading.

**Q2.2) For the places in the tweet which the response correctly identified as misinformed or potentially misleading, does the response also correctly explain why they are misinformed or potentially misleading by showing the facts refuting or providing the context around them (regardless of the language style)?** Your answer should be one of the following options.

- A. The response is **fully correct** in explaining why they are misinformed or potentially misleading.
- B. The response is **mostly correct** in explaining why they are misinformed or potentially misleading while having **minor** mistakes.
- C. The response is about **half correct and half incorrect** in explaining why they are misinformed or potentially misleading.
- D. The response is **mostly incorrect** in explaining why they are misinformed or potentially misleading with **significant** mistakes.
- E. The response is **completely incorrect** in explaining why they are misinformed or potentially misleading.

If your answer to Q2.2 is among A-D, please answer Q2.2.1.

**Figure S16.** Annotation instructions (page 3/7, continued on the next page).

**Q2.2.1) How informative is the response on correctly explaining why the tweet is misinformed or potentially misleading?** Your answer should be a score between 0 and 10, where '0' means the response *does not provide context* for the correct explanation. '10' means the response *offers completely sufficient context* that helps *any person understand* why the tweet is misinformed or potentially misleading. Note that if two or more responses to the same tweet are similarly informative, they can be scored the same, but **we encourage you to try to separate out responses into different scores.**

*Note that your answer to the following questions (Q3-Q7) should only be based on the text of the responses.*

**Q3) How relevant is the response text to the tweet?** Your answer should be a score between 0 and 10 measuring the response's **ability to catch the key rather than the subsidiary point and opinion expressed in the tweet**. '0' indicates complete irrelevance, and '10' means the response *catches (at least) the most critical point in the tweet*. Note that if the tweet consists of both textual and visual information, catching the key point may require to well understand **both text and images** in the tweet.

**Q4) What's the overall factuality of the response text?** Your answer should be one of the following options:

- A. The response is **completely factual and accurate**. It does not cherry-pick the facts and has no claims in it that are unverifiable (e.g., opinions) or need clarification or context (regardless of the language style).
- B. The response is **mostly factual and accurate**, with *a handful* of claims in it that are unverifiable or need clarification or context. Overall, however, the response is **barely misleading**.
- C. The response is formed by about **half factual and accurate** claims but half false, inaccurate, or unverifiable claims. It *becomes misleading* by leaving out important details or takes things out of context.
- D. The response is **mostly false, inaccurate, or unverifiable**. It contains *an element of truth*. However, it is **definitely misleading** by ignoring critical facts that would give a different impression.
- E. The response is **completely false, inaccurate, or unverifiable**.
- F. Not applicable; the response does not have text.

**Q5) Does the response text have mistakes in the use of English?** Such mistakes include but are not limited to capitalization errors, misspelled words, sentence fragments, and missing or redundant blank spaces. Note that if a piece of text does not follow the rules of American English but non-American English or

**Figure S17.** Annotation instructions (page 4/7, continued on the next page).

vice versa, it does NOT count as a mistake. Your answer should be one of the following options.

- A. Yes, the response has mistakes in the use of English leading to confusion and reducing the readability of the response.
- B. Yes, but the mistakes are minor. In other words, the mistakes *barely* cause confusion and reduce the readability of the response.
- C. No. The response does not have mistakes in the use of English.
- D. Not applicable; the response does not have text.

**Q6) What is the overall coherence (logical consistency) and logicalness (correct and valid reasoning) of the response text?** Your answer should be one of the following options:

- A. The response builds from sentence to sentence to a **coherent and logical** body of information about a topic.
- B. The response is **partially coherent and logical**. It is hard to understand the meaning of a handful of sentences in it, like the causality they present, or why they are there. Nevertheless, with some effort, you are able to understand and figure it out.
- C. The response is **barely coherent and logical**. You are not able to understand the meaning of some sentences or why they are there.
- D. **Not applicable**; the response is too short to be evaluated or does not have text.

**Q7) Is the response text impolite, provoking, or biased toward who posted the tweet, a third individual, or a certain group (e.g., Democrats or Republicans)?** Your answer should be one of the following options:

- A. Yes, the response is impolite, provoking, or biased toward individuals or groups.
- B. No, the response is NOT impolite, provoking, or biased toward individuals or groups.
- C. Not applicable; the response does not have text.

*Now, let's focus on the links included within responses...*

**Q8) Can this link be accessed?** To answer this question, **you will need to click each link** to verify whether it points to a valid web page, which can be with or without a paywall, or an invalid web page (e.g., "Page not found"). Your answer should be one of the following options:

- A. Yes, the link can be accessed.
- B. No, the link can NOT be accessed.

---

**Figure S18.** Annotation instructions (page 5/7, continued on the next page).

If your answer to Q8 is A, please answer Q8.1-Q8.2.

**Q8.1) What's the credibility of this link?** Your answer should be one of the following options:

- A. **Very high credibility.** The link's content appears to be backed up by facts with minimal bias, e.g., in politics and language. The source always publishes high-quality information with minimal bias.
- B. **High credibility.** The link's content appears to be backed up by facts, though it can be slightly biased, e.g., in politics and language. The source leans towards a certain group (e.g., a political party), but overall it publishes information backed up by facts.
- C. **Medium credibility.** The link's content appears to be backed up by facts, though it can be biased, e.g., in politics and language. However, the source has a mix of high- and low-quality information, or the source has a clear bias toward a certain group (e.g., a political party), often publishing information favoring it and information negative to the other group.
- D. **Low credibility (informed in the response).** The link's content and its source are both questionable. However, the response informs readers of its low credibility; typical examples of such expressions can be "[LINK] in the tweet is false..." and "The image attached in the tweet originates from a satire website ([LINK])..."
- E. **Low credibility (not informed in the response).** The link's content and its source are both questionable. Meanwhile, the response doesn't inform readers of its low credibility.
- F. **Can't determine;** the link's content is behind a paywall or uses non-English language, or the link cannot be accessed.

**Q8.2) Is this link relevant to the response text (note: not the tweet)?** Note that if the link is provided after some sentences of the response text rather than at the end of the response, evaluating the relevance should be conducted between the content that the link points to and these sentences rather than the whole response text. Your answer should be one of the following options:

- A. Yes, the link's content is relevant to or supports the response text.
- B. No, the link's content is *barely* relevant to the response text.
- C. Can't determine; the link's content is behind a paywall or uses non-English language, or the link cannot be accessed, or the response does not have text.

Finally, you will need to answer one last question (Q9) based on the **text and links** in the responses as well as your answers to the previous questions.

**Figure S19.** Annotation instructions (page 6/7, continued on the next page).

**Q9) How high-quality is the response in general?** Your answer should be a score between 0 and 10, where '0' refers to extremely low quality, and '10' refers to extremely high quality. If two or more responses to the same tweet are of similar quality, they can be scored the same, but **we encourage you to try to separate out responses into different scores.**

Continue

**Figure S20.** Annotation instructions (page 7/7).

[<< Back to Assigned Tasks](#)[Instructions](#)[Example Task](#)

## Task #: 1

**Jeff Zients (@WHCOS)**

See User Description

Time: Dec. 16, 2022, 9:32 p.m.

**Tweet:** By selling oil from the SPR at \$96/barrel, we helped reduce the price of gas when it was at its highest. Now, by refilling the SPR at today's much lower prices, we turn a profit for the taxpayers.



**Note:** You will see the same responses across all questions.

\* indicates a required field.

**Q1\*** What's the clarity of the response in explaining where and why the tweet is misinformed or potentially misleading?

R1 — Time: Dec. 18, 2022, 2:24 p.m.

The Strategic Petroleum Reserve has been drawn down over 200 million barrels over the past year (~1/3 of the previously stored amount). <https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=mcsstus1&f=m> The refill being promoted is to pilot the purchase of up to 3 million barrels in January & deliver in February. <https://www.spr.doe.gov/posting/PURCHASE/FY23%20Winter%20Oil%20Purchase%203MMB/1-Request%20for%20Proposal/FY23%20Winter%20Oil%20Purchase%203MMB%20Request%20for%20Proposal.pdf>

- Explicitly explains — The response explicitly explains where and why the tweet is misinformed or potentially misleading (regardless of whether the explanation are correct). A typical example of such expressions can be, "Though it is true that X, the tweet is misinformed by claiming that Y because Z", where X and Y are from the tweet and Z is the explanation.
- Unclearly explains — Given the response, it is hard to tell where and why the tweet is misinformed or potentially misleading.
- Somewhere in between — For example, the response may only implicitly explain where and why the tweet is misinformed or potentially misleading.

R2 — Time: Dec. 18, 2022, 2:24 p.m.

This tweet is potentially misleading. While it's true that the Strategic Petroleum Reserve (SPR) was sold at a high price and is being refilled at a lower price, this doesn't necessarily equate to a profit for taxpayers. The SPR is a government-owned reserve meant to provide a buffer against potential disruptions in oil supplies, not a profit-making entity. The money from the sale goes back into the government's general fund, not directly to taxpayers. Furthermore, the price of gas is influenced by a multitude of factors, not just the sale or purchase of oil from the SPR. <https://www.energy.gov/eere/services/petroleum-reserves/strategic-petroleum-reserve> <https://www.eia.gov/energyexplained/oil-and-petroleum-products/prices-and-outlook.php>

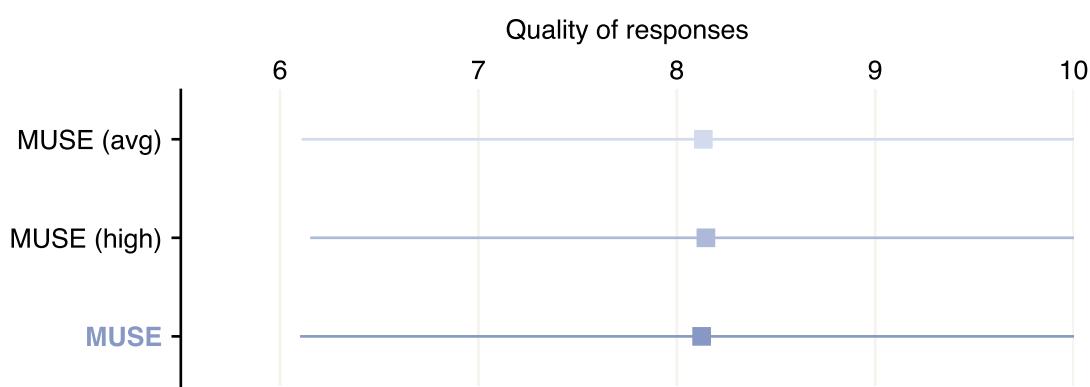
- Explicitly explains — The response explicitly explains where and why the tweet is misinformed or potentially misleading (regardless of whether the explanation are correct). A typical example of such expressions can be, "Though it is true that X, the tweet is misinformed by claiming that Y because Z", where X and Y are from the tweet and Z is the explanation.
- Unclearly explains — Given the response, it is hard to tell where and why the tweet is misinformed or potentially misleading.
- Somewhere in between — For example, the response may only implicitly explain where and why the tweet is misinformed or potentially misleading.

R3 — Time: Dec. 18, 2022, 2:24 p.m.

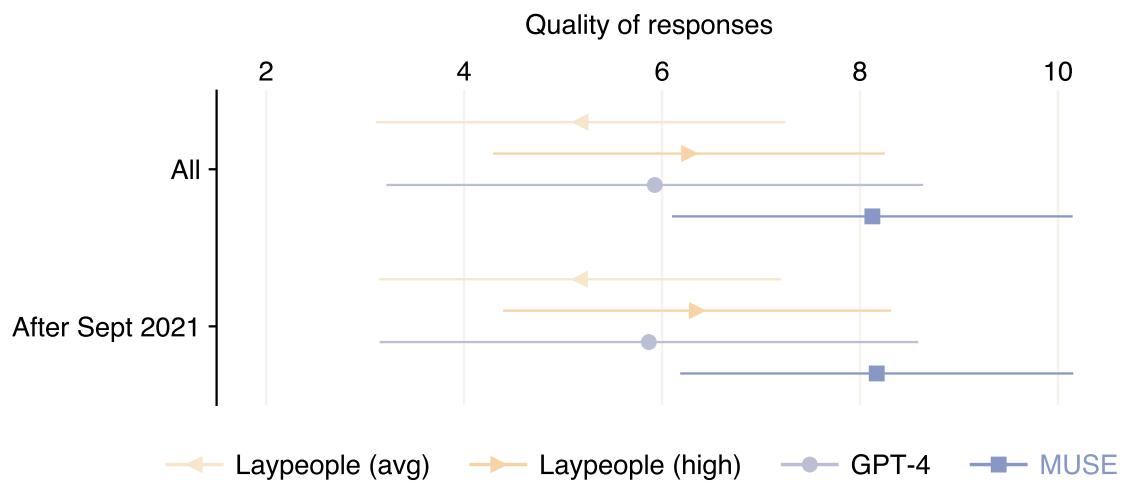
The Democrats blocked Republicans from filling the SPR when oil prices were at \$24, significantly below current prices. <https://www.bloomberg.com/news/articles/2022-09-13/biden-may-buy-oil-just-below-80-democrats-stymied-trump-at-24>

- Explicitly explains — The response explicitly explains where and why the tweet is misinformed or potentially misleading (regardless of whether the explanation are correct). A typical example of such expressions can be, "Though it is true that X, the tweet is misinformed by claiming that Y because Z", where X and Y are from the tweet and Z is the explanation.
- Unclearly explains — Given the response, it is hard to tell where and why the tweet is misinformed or potentially misleading.

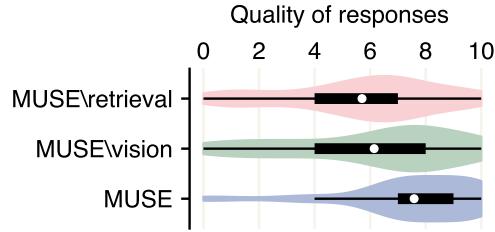
**Figure S21.** Annotation task page.



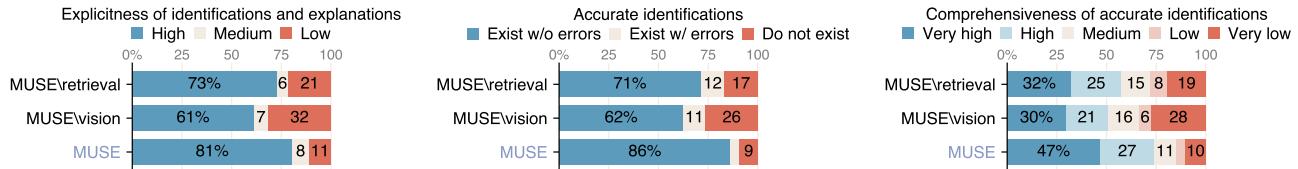
**Figure S22.** Impact of starting times of responding to tweets on MUSE's performance. The simulated starting time for MUSE (avg): Thirty minutes before the corresponding laypeople's average-helpfulness responses was created (median: 16 hours after the corresponding tweet was posted; Supplementary Fig. S13). The simulated starting time for MUSE (high): Thirty minutes before the corresponding laypeople's high-helpfulness responses was created (median: 13 hours after the corresponding tweet was posted; Supplementary Fig. S13). The simulated starting time for MUSE: The post time of the corresponding tweet (i.e., 0 hours after the corresponding tweet was posted).



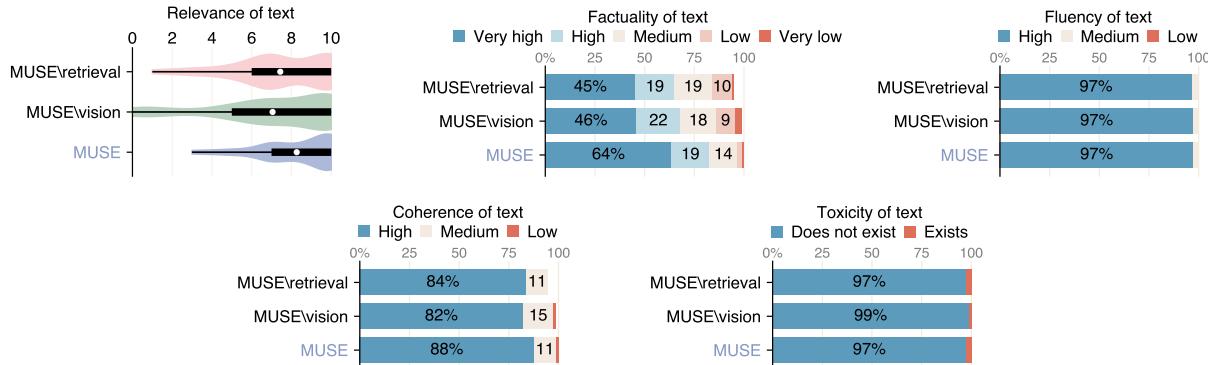
**Figure S23.** Impact of post times of tweets on the performance of MUSE and baselines.



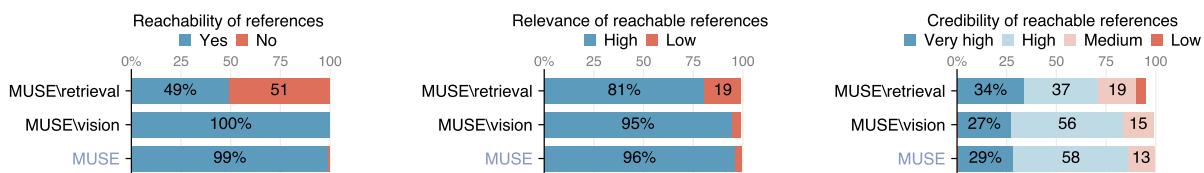
**(a)** Overall quality of responses.



**(b)** Quality of responses in identifying and explaining (in)accuracies.

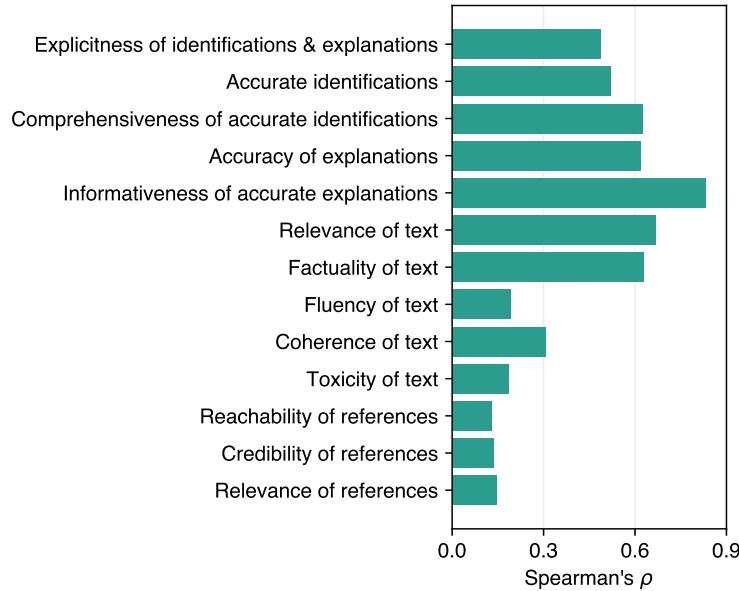


**(c)** Quality of responses in generated text.

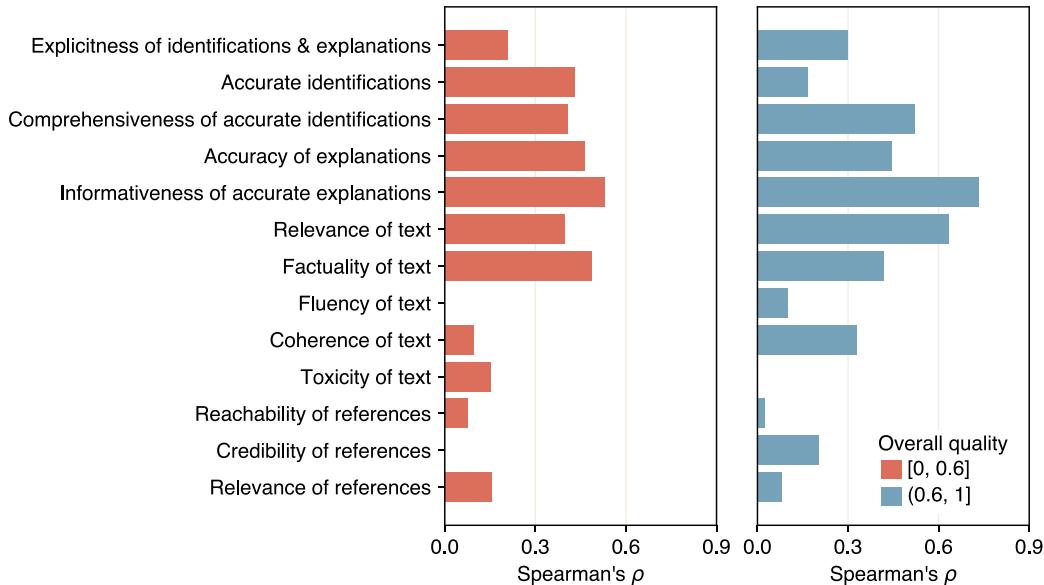


**(d)** Quality of responses in references.

**Figure S24.** Impact of retrieval and vision on MUSE's performance. Here, MUSE and MUSE\vision responded to tweets by only retrieving web pages published thirty minutes before the creation time of the corresponding laypeople's high-helpfulness response.



(a) Correlation for all responses.



(b) Correlation for responses whose overall quality is relatively low vs high.

**Figure S25.** Spearman correlation coefficient between overall response quality and each criterion that specifies overall response quality.