

# 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 can be partially correct and even factual but misleading. It undermines public trust in science and democracy, particularly on social media, where it can spread rapidly. High-quality and timely correction of misinformation that identifies and explains its (in)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 also describes images and conducts multimodal searches to verify and correct multimodal content. Fact-checking experts 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. We propose and evaluate 13 dimensions of misinformation correction quality, ranging from the accuracy of identifications and factuality of explanations to the relevance and credibility of references. The results demonstrate MUSE’s ability to promptly write high-quality responses to potential misinformation on social media—overall, MUSE outperforms GPT-4 by 37% and even high-quality responses from laypeople by 29%. This work reveals LLMs’ potential to help combat real-world misinformation effectively and efficiently.

## Introduction

Misinformation, which broadly includes partially incorrect and factual but misleading content<sup>1,2</sup>, 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>2,6,7</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 of 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 one to search for related and trusted articles and write justifications—whereas creating and spreading misinformation is easy. As a consequence, a 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; Supplementary 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 such as LLMs (e.g., OpenAI’s GPT series) 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>. For instance, GPT-4<sup>27</sup> (as of March 2023) and MisinfoCorrect<sup>28</sup> lack access to timely updated knowledge and 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>26,29</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>30–37</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>31,38,39</sup>. These LLMs, as well as MisinfoCorrect<sup>28</sup>, also struggle with counteracting multimodal misinformation due to their nonacceptance of visual inputs. Finally, understanding the quality of a correction necessitates a comprehensive evaluation of fact-checking experts due to the complexity of real-world misinformation<sup>26</sup>. However, prior evaluation of LLMs was conducted by non-experts on limited aspects of their generated responses (e.g., the general quality of responses<sup>28,37</sup>), which is insufficient to comprehensively understand their performance in identifying and explaining (in)accuracies, generated text, and references.

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. 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 effectively and promptly responding to potential misinformation. Results demonstrate MUSE’s advance when the content is textual with or without images 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.



## Approach

MUSE is designed to automatically respond to content that potentially is misinformation. In other words, the content might be inaccurate, partially accurate, or even factually accurate but misleading, all of which are misinformation, or fully accurate as non-misinformation. The content can contain text with or without visual information. The response should identify what part(s) of the content is (in)accurate, explain why that part of the content is (in)accurate, and provide links as references. The pipeline of MUSE is illustrated in Fig. 1.

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 (Supplementary 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 (Supplementary 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 additional features, including identification of celebrities and embedded text—information crucial for making accurate verification and corrections<sup>40</sup>. For example, even a state-

of-the-art image captioning model<sup>41</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 Supplementary 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> (Supplementary Fig. S6). We included the tweets from Community Notes with at least two responses ( $n=247$ ); 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 ( $p = 4 \times 10^{-12}$ ;  $N = 462$ ). In particular, 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).

Furthermore, we observed that *the quality of MUSE-generated responses to textual and multimodal (textual and visual) content that potentially is misinformation is higher than responses by GPT-4 and even high-helpfulness responses by laypeople* (Fig. 3a; Supplementary Fig. S7-S8). The quality of MUSE-generated responses to text-based tweets is 29% higher than GPT-4-generated responses ( $p = 6 \times 10^{-28}$ ;  $N = 310$ ), 33% higher than laypeople's high-helpfulness responses ( $p = 10^{-44}$ ;  $N = 310$ ), and 65% higher than laypeople's average-helpfulness responses ( $p = 10^{-67}$ ;  $N = 310$ ). The quality of MUSE-generated responses to multimodal tweets is generally 21% higher than laypeople's high-helpfulness responses ( $p = 3 \times 10^{-8}$ ;  $N = 154$ ), 39% higher than laypeople's average-helpfulness responses ( $p = 10^{-16}$ ;  $N = 154$ ),

and 56% higher than GPT-4 ( $p = 6 \times 10^{-18}$ ;  $N = 152$ ).

Finally, we found that *the quality of MUSE-generated responses to potential misinformation across domains including politics and international affairs, economy and business, crime and law, social issues and human rights, and health and medicine is higher than responses by GPT-4 and even high-helpfulness responses by laypeople* (Methods; Fig. 3b; Supplementary Fig. S9-S13). The quality of MUSE-generated responses is 35% (political and international affair;  $p = 10^{-16}$ ;  $N = 160$ ), 47% (economy and business;  $p = 4 \times 10^{-11}$ ;  $N = 76$ ), 44% (crime and law;  $p = 3 \times 10^{-8}$ ;  $N = 76$ ), 31% (social issues and human rights;  $p = 0.0002$ ;  $N = 60$ ), and 27% (health and medicine;  $p = 3 \times 10^{-5}$ ;  $N = 48$ ) higher than GPT-4-generated responses. The quality of MUSE-generated responses is 33% (political and international affair;  $p = 10^{-21}$ ;  $N = 160$ ), 27% (economy and business;  $p = 10^{-8}$ ;  $N = 76$ ), 28% (crime and law;  $p = 3 \times 10^{-9}$ ;  $N = 76$ ), 38% (social issues and human rights;  $p = 7 \times 10^{-10}$ ;  $N = 60$ ), and 25% (health and medicine;  $p = 4 \times 10^{-5}$ ;  $N = 48$ ) higher than even high-helpfulness responses by laypeople.

## Discussion

While concerns have arisen about LLMs in facilitating the creation of misinformation<sup>42,43</sup>, our work demonstrates LLMs' potential to improve the online information ecosystem by correcting misinformation<sup>1,2</sup>. Real-world misinformation can combine (in)accurate, factually accurate but misleading, and unverifiable claims. Identifying a piece of content as misinformation without further interventions, done by many previous AI-driven approaches<sup>44</sup>, has limited impact on reducing misinformation's detrimental effects. Showing the identification results without explanations might even increase false beliefs<sup>45</sup>. Correcting misinformation has been shown to reduce its spread and false beliefs, whose effectiveness is affected by quality, timeliness, and scalability<sup>4,22</sup>. However, research on scaling up and accelerating misinformation correction—which may include automation—is still at the early stage. Existing AI models struggle to identify and correct misinformation, especially on social media, where misinformation is not restricted to narrow domains and spreads rapidly. Correcting misinformation requires comprehending content that can be multimodal and the context beyond it that often involves emerging events. It requires identifying what part(s) of the content is (in)accurate and explaining why that part of the content is (in)accurate with trustworthy references. We propose MUSE and demonstrate the high quality of its automatically generated responses to social media posts that potentially are misinformation. MUSE augments existing powerful LLMs (here, it is GPT-4) with the capabilities of addressing images, accessing up-to-date knowledge, and finding accurate references. 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.

We provide MUSE as a solution to assist social media users and platforms in accurately, scalably, promptly, and transparently responding to suspicious content. MUSE is end-to-end and thereby simple to use. It is nonparametric, i.e., does not need training or fine-tuning an AI model, and thereby easily and cheaply updated, especially compared to parametric models<sup>32</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 and domains (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>28,38,39</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 two minutes after suspicious content (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'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 (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.

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>46</sup> and where misinformation has been shown to diffuse faster than the truth<sup>13</sup>. Fourth, experts assessed and compared MUSE against one other LLM, GPT-4, which can also be seen as an ablation study, since MUSE augments GPT-4 (Methods). 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>47,48</sup>.

## Conclusion

We proposed MUSE, a nonparametric LLM to scale up and accelerate misinformation corrections on social media. 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 produce high-helpfulness responses on X Community Notes in identifying and explaining inaccuracies, generating high-quality text, providing high-quality references, and producing the highest quality of responses overall. MUSE also excels when the responded content is textual or multimodal (textual and visual) and when it is related to politics and international affairs, economy and business, crime and law, social issues and human rights, or health and medicine. This study demonstrates the potential of LLMs in responding to online misinformation effectively, scalably, promptly, and transparently.

## References

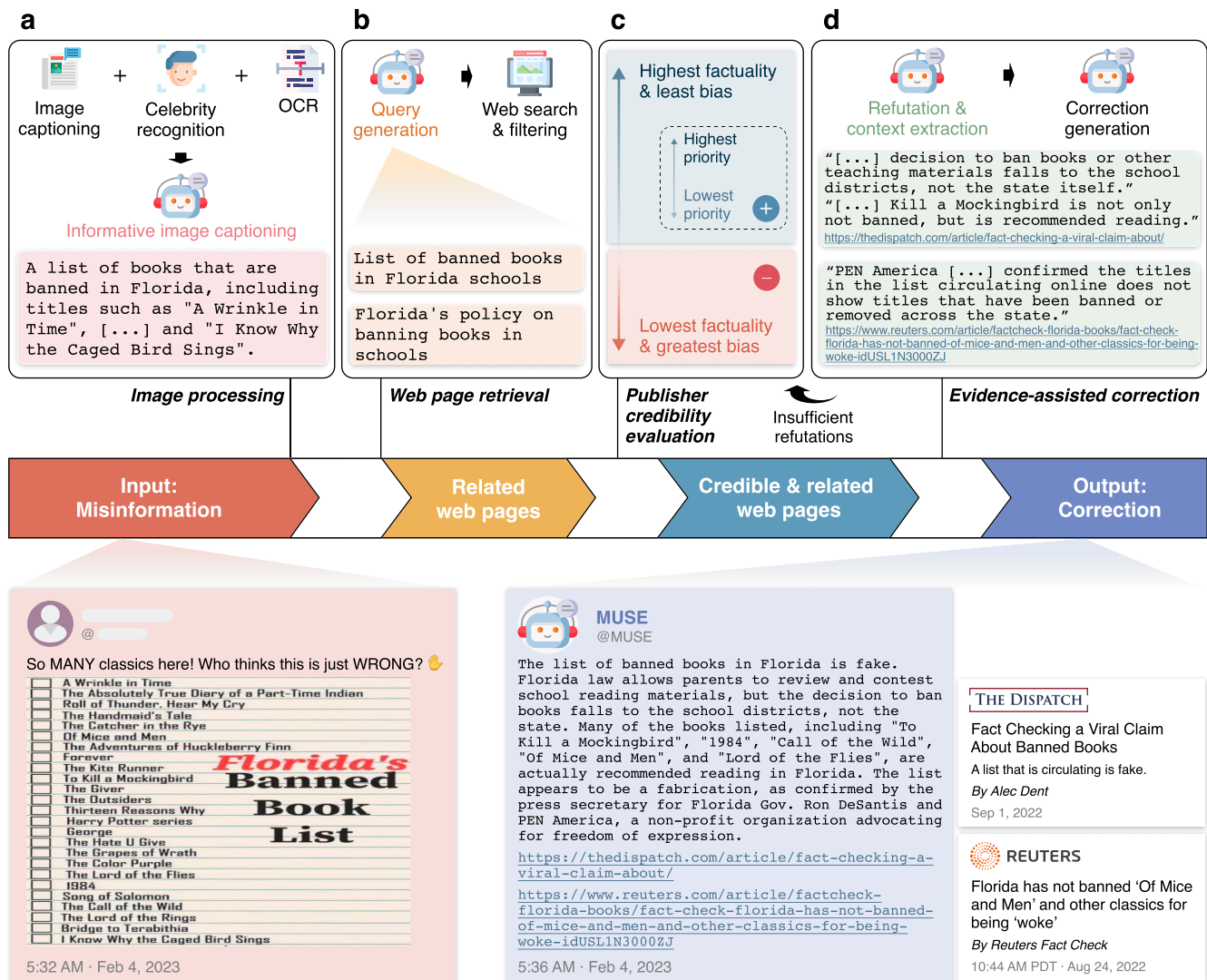
1. Traberger, C. S. Misinformation: broaden definition to curb its societal influence. *Nature* **606**, 653–653 (2022).
2. Watts, D. J., Rothschild, D. M. & Mobius, M. Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci.* **118**, e1912443118 (2021).
3. Lazer, D. M. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
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).



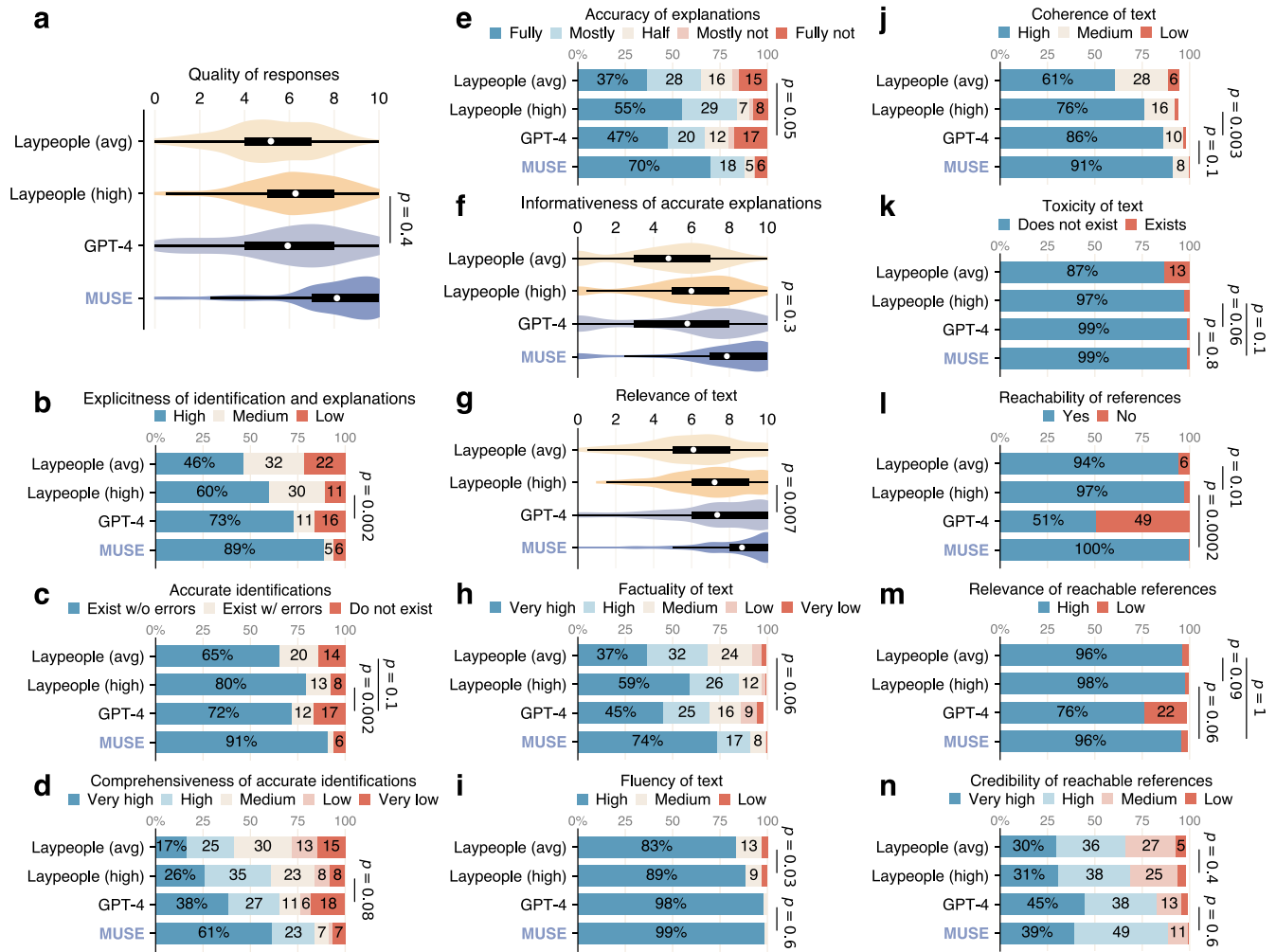
6. Moore, R. C., Dahlke, R. & Hancock, J. T. Exposure to untrustworthy websites in the 2020 US election. *Nat. Hum. Behav.* 1–10 (2023).
7. Forum, W. E. Global risks report 2024 (World Economic Forum, 2024).
8. Thompson, B. G. *et al.* *Final Report of the Select Committee to Investigate the January 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**, eabo6169 (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. 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).
29. 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).
30. 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).
31. Menick, J. *et al.* Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147* (2022).
32. 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).
33. 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).
34. 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).
35. Vu, T. *et al.* FreshLLMs: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214* (2023).
36. 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).
37. 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).
38. Ecker, U. K. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).
39. 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).
40. 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).
41. 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).
42. Menczer, F., Crandall, D., Ahn, Y.-Y. & Kapadia, A. Addressing the harms of AI-generated inauthentic content. *Nat. Mach. Intell.* **5**, 679–680 (2023).

43. 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).
44. Zhou, X. & Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)* **53**, 1–40 (2020).
45. Guo, Z., Schlichtkrull, M. & Vlachos, A. A survey on automated fact-checking. *Transactions Assoc. for Comput. Linguist.* **10**, 178–206 (2022).
46. Walker, M. & Matsa, K. E. News consumption across social media in 2021. *Pew Res. Cent.* (2021).
47. Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. GPT-4 passes the bar exam. *Philos. Transactions Royal Soc. A* **382**, 20230254 (2024).
48. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).

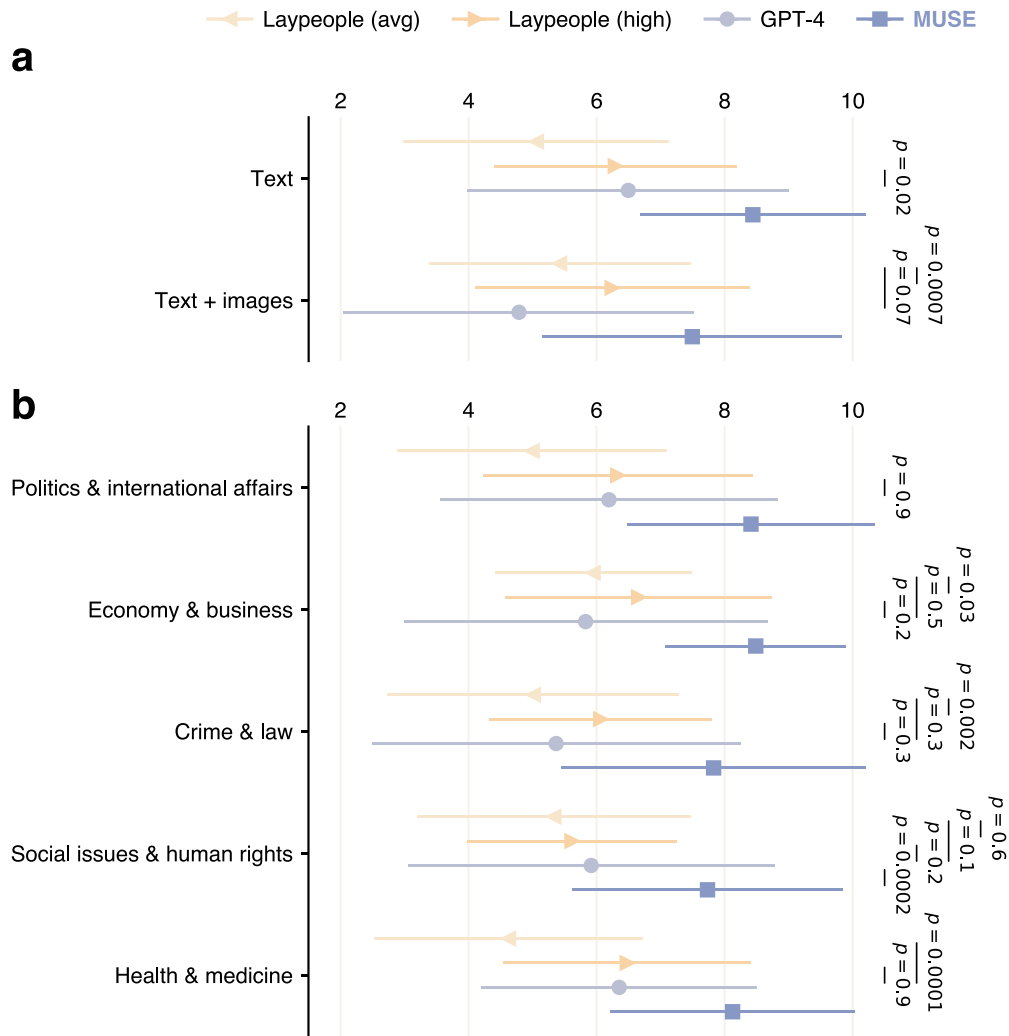


**Figure 1.** Overview of MUSE, an LLM augmented by addressing images and accessing timely knowledge from credible publishers to enable identifying and explaining (in)accuracies in a piece of multimodal content with accurate and trustworthy references. Given a piece of content that may or may not be 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=132). **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 and domains ( $p < 4 \times 10^{-5}$  for each approach pair respectively in **a-b** by Mann-Whitney U test; experiments=132). **a:** The quality of MUSE-generated responses to textual misinformation (mean: 8.4;  $n = 155$ ) is more than 29% higher than responses by GPT-4 (6.5; 155) and laypeople (high-helpfulness: 6.3; 155, average-helpfulness: 5.1; 155). For multimodal (textual and visual) responses (7.5; 77), the quality is more than 21% higher than responses by GPT-4 (4.8; 77) and laypeople (high: 6.2; 77, average: 5.4; 75). **b:** The quality of MUSE-generated responses to politics and international-affair misinformation (mean: 8.4;  $n = 80$ ) is more than 33% higher than responses by GPT-4 (6.2; 80) and laypeople (high-helpfulness: 6.3; 80, average-helpfulness: 5.0; 79). For economy and business misinformation (8.5; 38), the quality is more than 27% higher than responses by GPT-4 (5.8; 38) and laypeople (high: 6.7; 38, average: 5.9; 38). For crime and law misinformation (7.8; 38), the quality is more than 28% higher than responses by GPT-4 (5.4; 38) and laypeople (high: 6.1; 38, average: 5.0; 38). For social issues and human rights misinformation (7.7; 30), the quality is more than 31% higher than responses by GPT-4 (5.9; 30) and laypeople (high: 5.6; 30, average: 5.3; 30). For health and medicine misinformation (8.1; 24), the quality is more than 25% higher than responses by GPT-4 (6.4; 24) and laypeople (high: 6.5; 24, average: 4.6; 24). 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>41</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. S14 and the prompt in Supplementary Fig. S15). These example images do not appear in the dataset we used. The temperature of GPT-4 was set as 0.

**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. The temperature of GPT-4 was set as 0.

**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>49</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>50</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>50</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>51</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>52-54</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. It also explicitly excludes the publishers whose factuality is low, including those rejecting established scientific consensus on issues such as climate change or vaccines, identified as overt propaganda, and designated as hate groups by reputable third-party evaluators (Supplementary Table S1). MUSE further divided the publishers considered as potential references into three priorities. High-priority publishers (#=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 (#=2,123), and publishers who do not have high- or medium priority were low priority (#=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. We set the temperature of GPT-4 as 0.

**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>35</sup>. We set the temperature of GPT-4 as 0.

## 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. S16; 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. S16). 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. S16).

**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 accurate, 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. S17). 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. S17). 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, including the informed consent form, 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).

**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. S18-S24) and demonstrated the use of our web interface for data annotation (Supplementary Fig. S25). 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>55,56</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>55</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>57</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>58</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>19,59</sup>.

**Domain classification.** We started by asking GPT-4 (gpt-4-0613) an open-ended question “Which topic is the tweet content most related to? Your answer should be a word or phrase.” Then, we asked GPT-4 to “Cluster all the topics into  $N$  domains”; we determined  $N = 6$  after manual review. These six domains are (i) politics and international affairs, (ii) economy and business, (iii) crime and law, (iv) social issues and human rights, (v) health and medicine, and (vi) other that includes entertainment, sports, astronomy, and more. Finally, we asked GPT-4 a close-ended question “Which of the six domains is the tweet content most related to? Your answer should only contain the domain’s name.” The tweet content concatenated its textual content and informative descriptions of images (for tweets with images). The temperature of GPT-4 was set as 0. Among 232 tweets included in our final analyses, 80 (34%) are related to politics and international affairs, 38 (16%) are related to economy and business, 38 (16%) are related to crime and law, 30 (13%) discuss social issues and human rights, 24 (10%) are in health and medicine, and the remaining 22 (9%) are in other domains.

**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. S26 show that MUSE performs similarly (mean $\pm$ SD of the overall quality of responses: 8.1 $\pm$ 2.0) 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. S17), and when it follows the starting times of laypeople who produce average-helpfulness responses (median: 16 hours after the tweet was posted; Supplementary Fig. S17). Second, we separated tweets posted after September 2021 (#=207; Supplementary Fig. S27) from all tweets (#=232; Supplementary Fig. S27), 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. S28 show that MUSE performs stably (mean $\pm$ SD of the overall quality of corrections: 8.1 $\pm$ 2.0) 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. S29 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.

**Runtime and cost.** We conducted experiments on 16G memory M1 CPU. The program ran in five parallel processes. The average runtime of MUSE in responding a social media post was two minutes. The total cost of MUSE to respond 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).

## 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

49. Hamborg, F., Meuschke, N., Breiting, 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).

50. 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).
51. 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).
52. 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).
53. Zhou, X., Mulay, A., Ferrara, E. & Zafarani, R. ReCOVary: A multimodal repository for COVID-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3205–3212 (2020).
54. 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).
55. Yuan, W., Neubig, G. & Liu, P. BARTScore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* **34**, 27263–27277 (2021).
56. Fabbri, A. R. *et al.* SummEval: Re-evaluating summarization evaluation. *Transactions Assoc. for Comput. Linguist.* **9**, 391–409 (2021).
57. 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).
58. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Medica* **22**, 276–282 (2012).
59. 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) and Amy X. Zhang (axz@cs.uw.edu).

## Supplementary materials

### List of supplementary materials

Table [S1](#)

Figures [S1](#) to [S29](#)



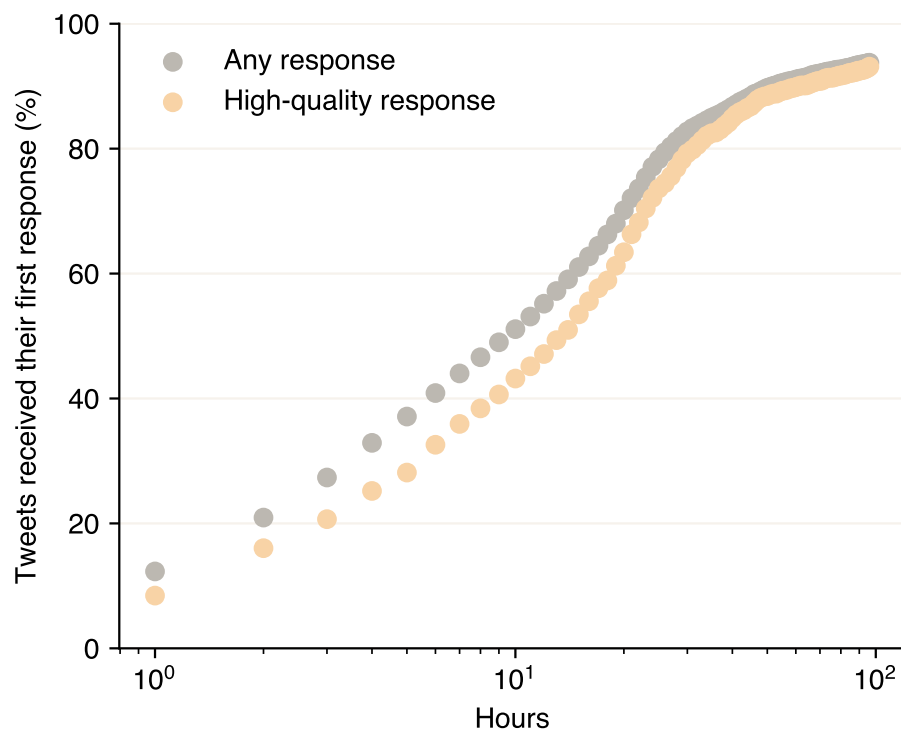
**Table S1.** Definitions and statistics of factuality and bias categories offered by Media Bias/Fact Check ([mediabiasfactcheck.com](https://mediabiasfactcheck.com)).

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

---

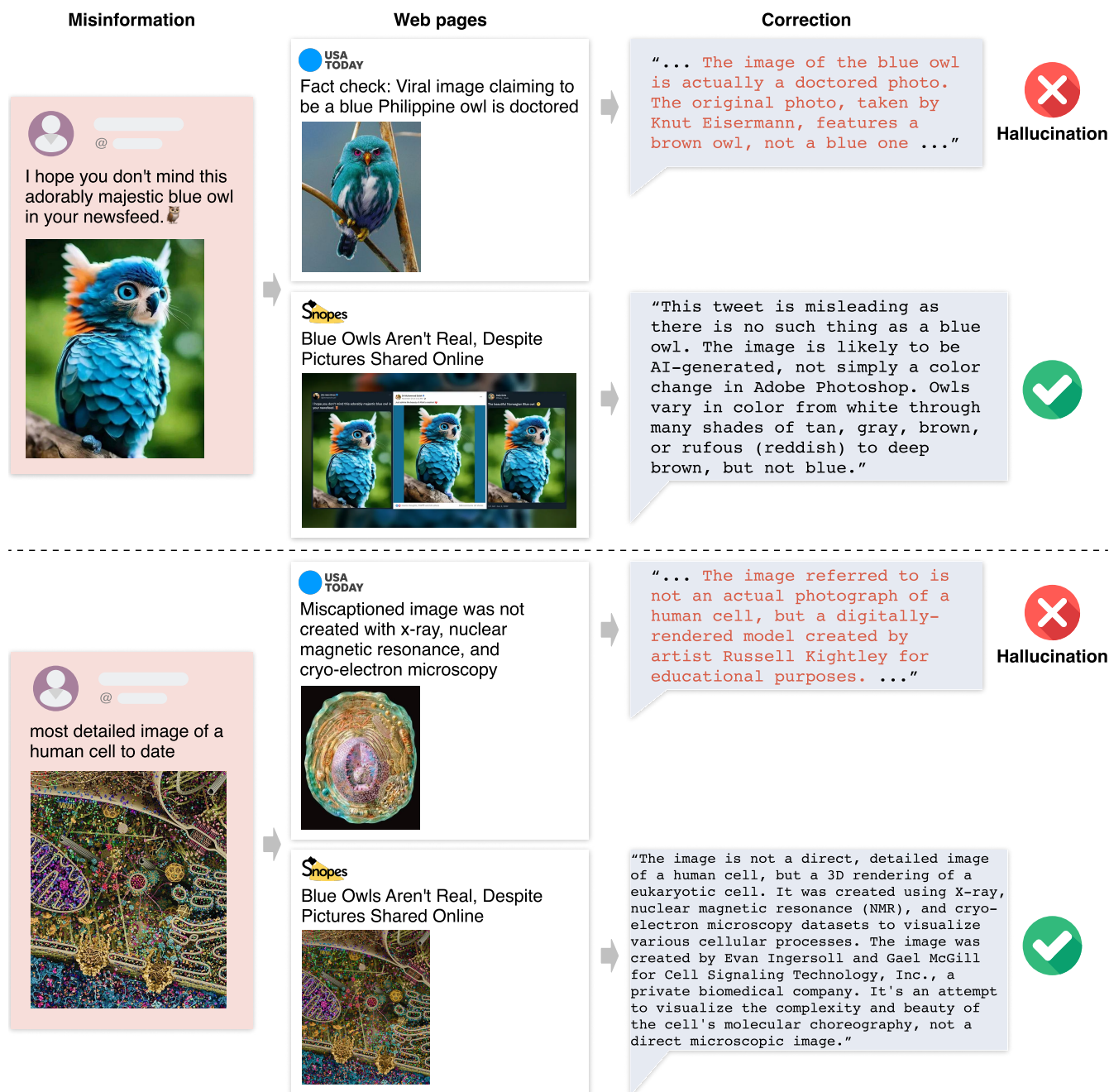
**Total: 5,272**



**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
 <p>MAJOR BREAKING: The Pentagon confirms that at least THREE Chinese spy balloons flew over the US during Trump's presidency and <b>he HID THEM from the public and never shot them down.</b></p>	 Pentagon confirmation on Chinese spy balloons over US during Trump's presidency	 Pentagon confirmation on Chinese spy balloons over US during Trump's presidency  <b>Trump hiding information about Chinese spy balloons</b>  <b>Incidents of Chinese spy balloons over US not being shot down</b>
 <p>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>	 Paul Whelan dishonorable discharge reasons	 Paul Whelan dishonorable discharge reasons  Paul Whelan larceny, dereliction of duty, lying, social security fraud, and writing bad checks  <b>George Floyd fake \$20 bill incident details</b>
 <p>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>	 Teenager driven by mother across state lines to fire into crowd of protestors	 <b>Actor accidentally discharges gun on movie set</b>  Teenager fires into crowd of protestors  Teenager driven by mother to protest across state lines

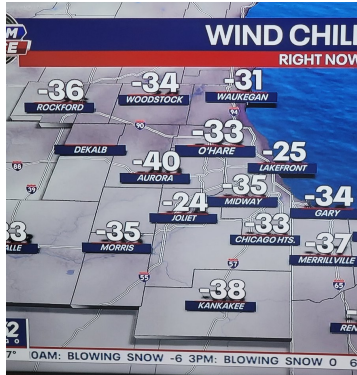
**Figure S2.** Examples that show that generating multiple queries helps decompose a post, which may have multiple claims that each needs to be verified, whereas generating one query may overlook some of them and hence lead to not comprehensive identifications of (in)accuracies. Bold text: the verification-needed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



### Image captioning

a weather map on a tv screen

elon musk's tweet on social media

a young person sitting on a chair

### Informative image captioning

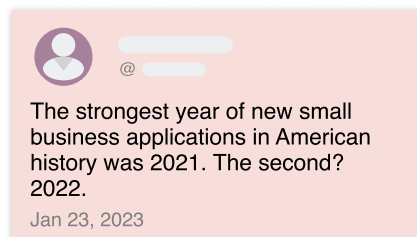
A weather map displayed on a TV screen showing wind chill temperatures in various locations such as Waukegan, Woodstock, Rockford, DeKalb, O'Hare, Lakefront, Aurora, Midway, Gary, Joliet, Chicago Heights, Morris, Merrillville, and Kankakee. The map indicates extreme cold temperatures with conditions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

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.

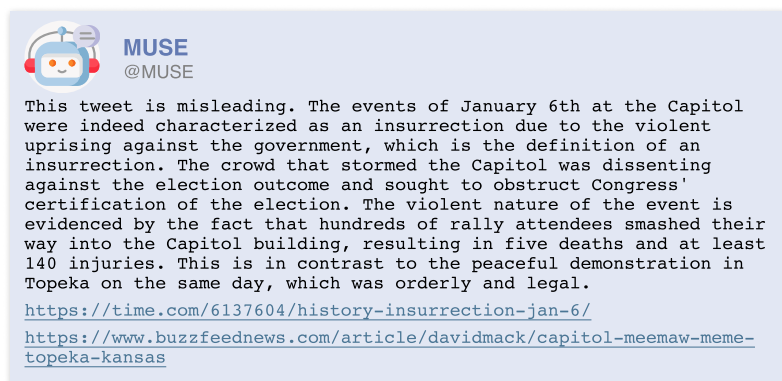
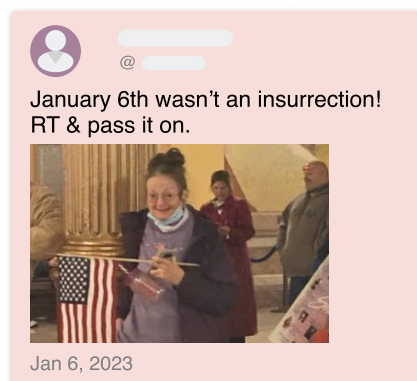
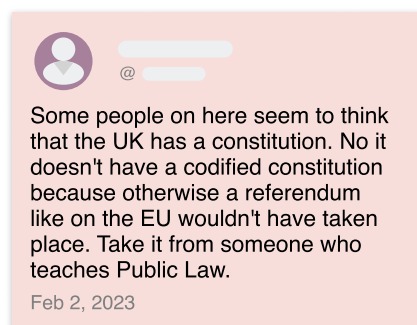
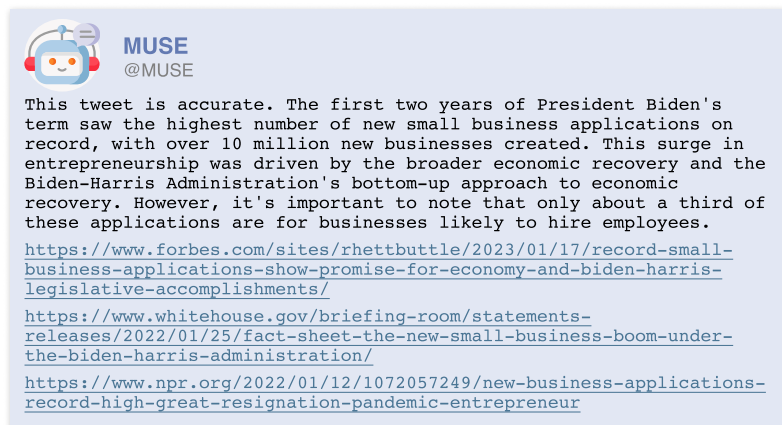
A photo of a young Aaron Swartz sitting on a chair.

**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



### MUSE-generated response



**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.

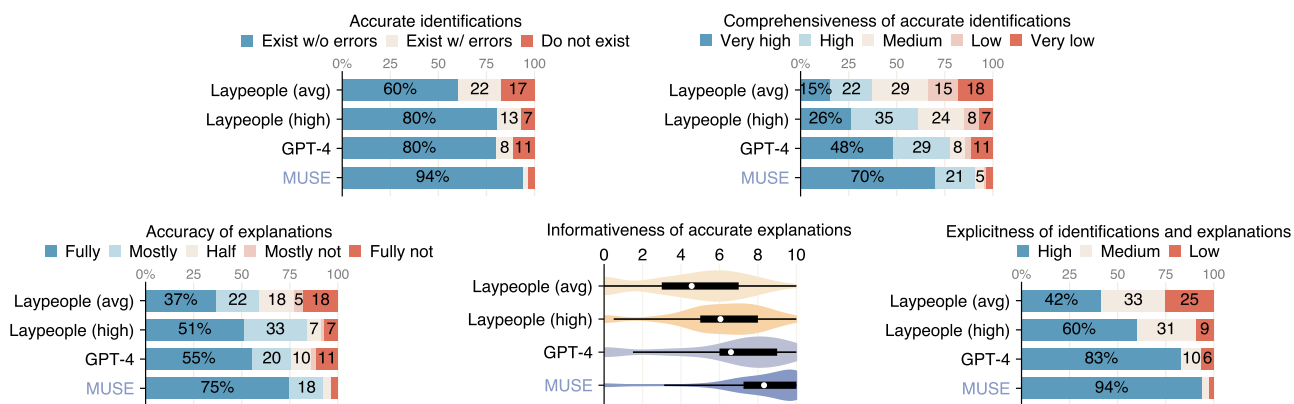
[buzzfeednews.com/article/davidm...](https://buzzfeednews.com/article/davidm...)

Do you find this helpful?

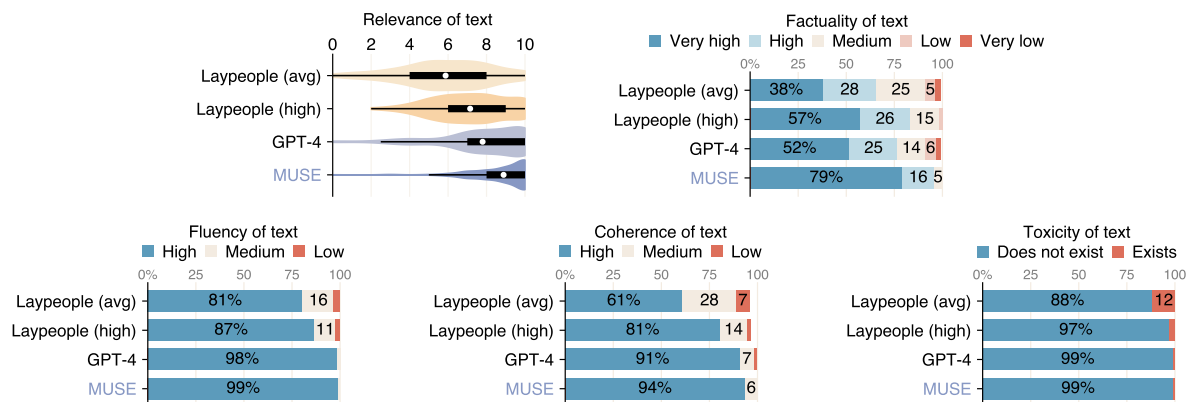
**Rate it**

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

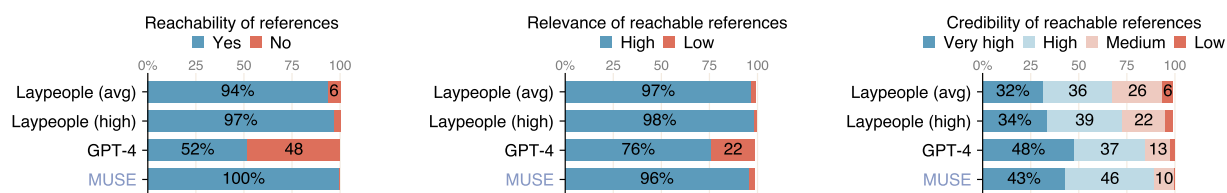
**Figure S6.** An example of a high-helpfulness response from Community Notes displayed on the corresponding tweet and visible to the public.



(a) Quality of responses in identifying and explaining (in)accuracies.

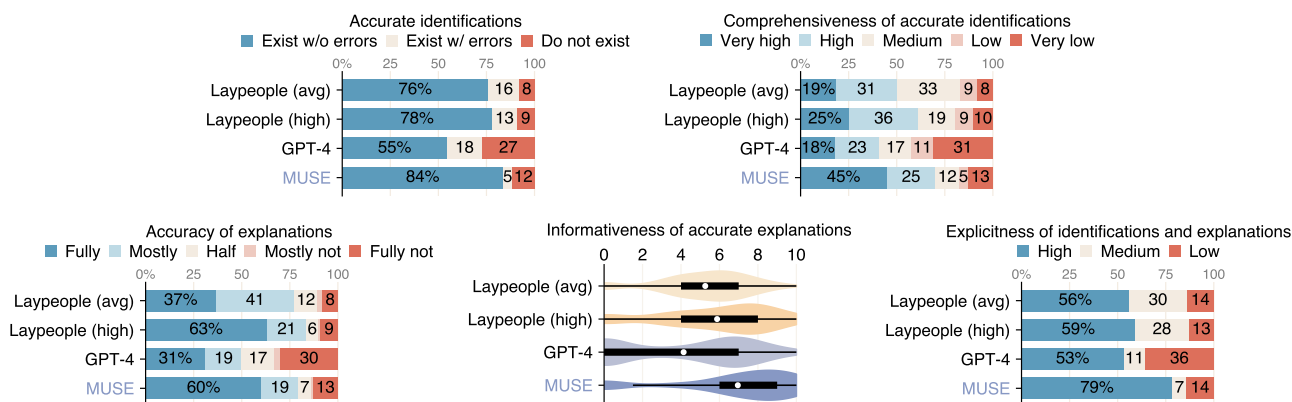


(b) Quality of responses in generated text.

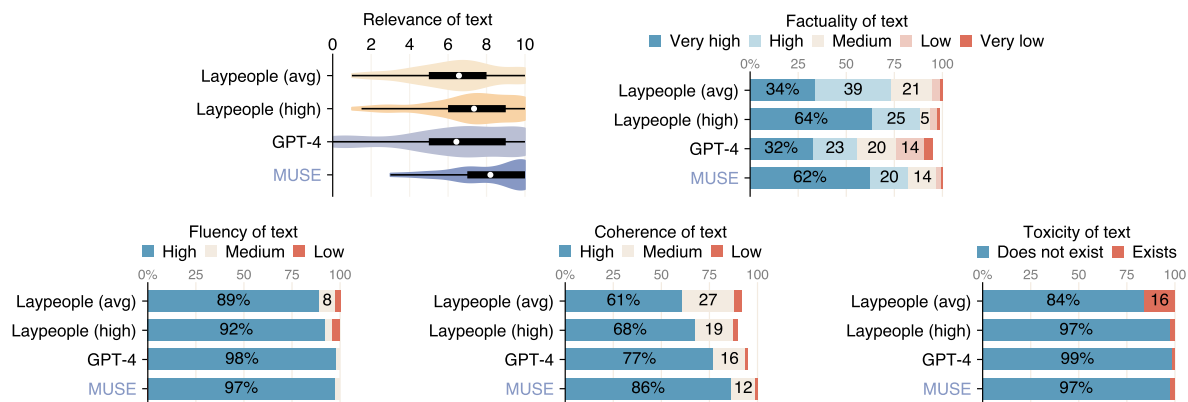


(c) Quality of responses in references.

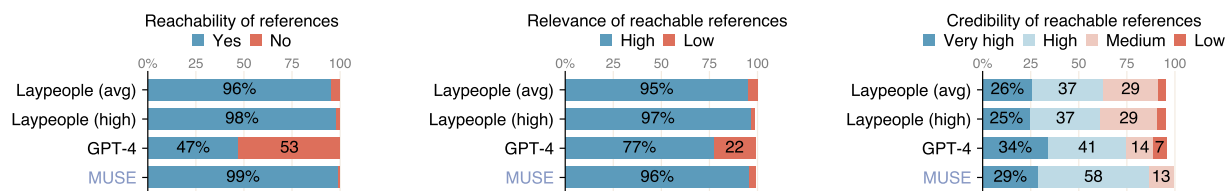
**Figure S7.** Expert evaluation results for textual posts (#=155).



(a) Quality of responses in identifying and explaining (in)accuracies.



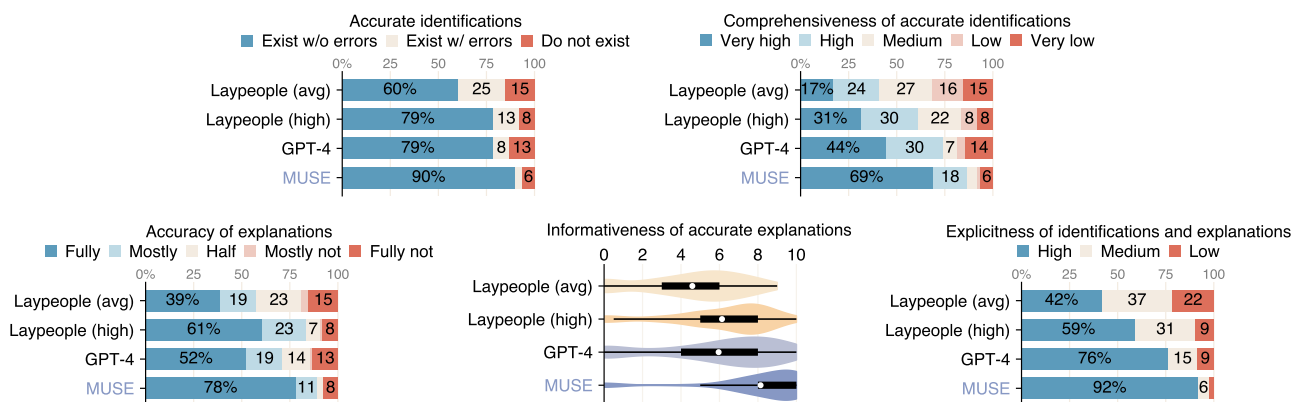
(b) Quality of responses in generated text.



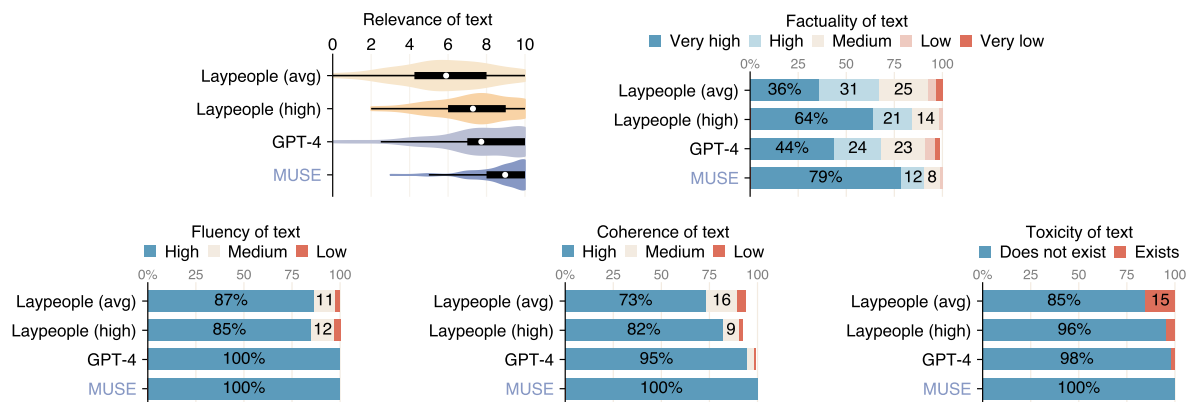
(c) Quality of responses in references.

**Figure S8.** Expert evaluation results for multimodal (textual and visual) posts (#=77).

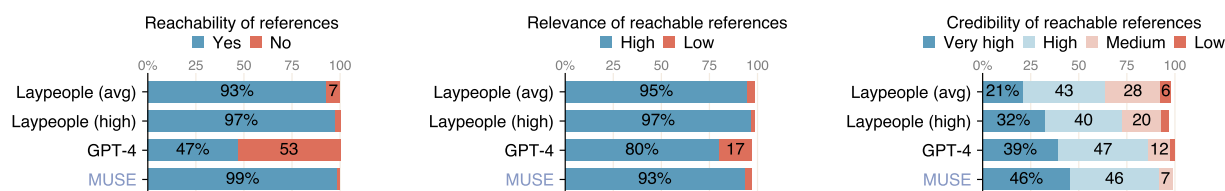




(a) Quality of responses in identifying and explaining (in)accuracies.

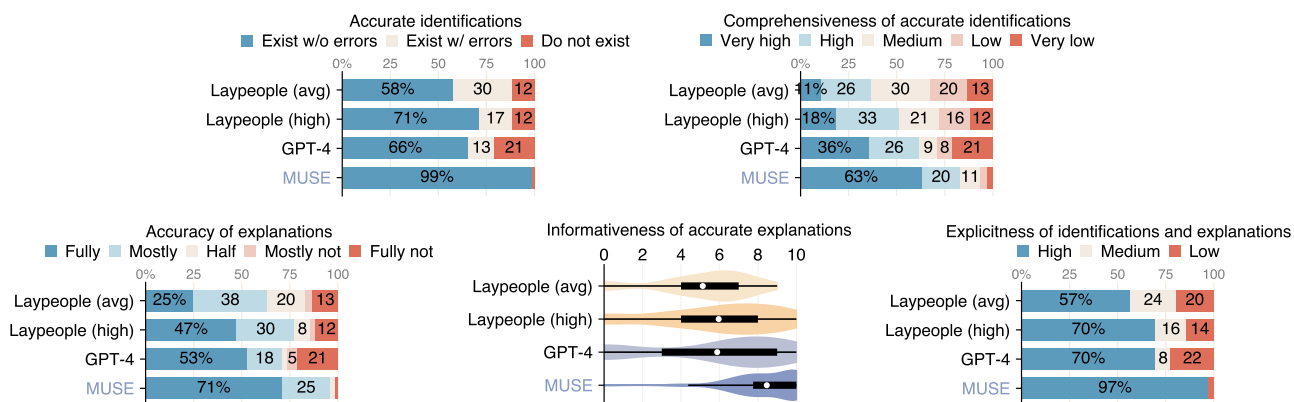


(b) Quality of responses in generated text.

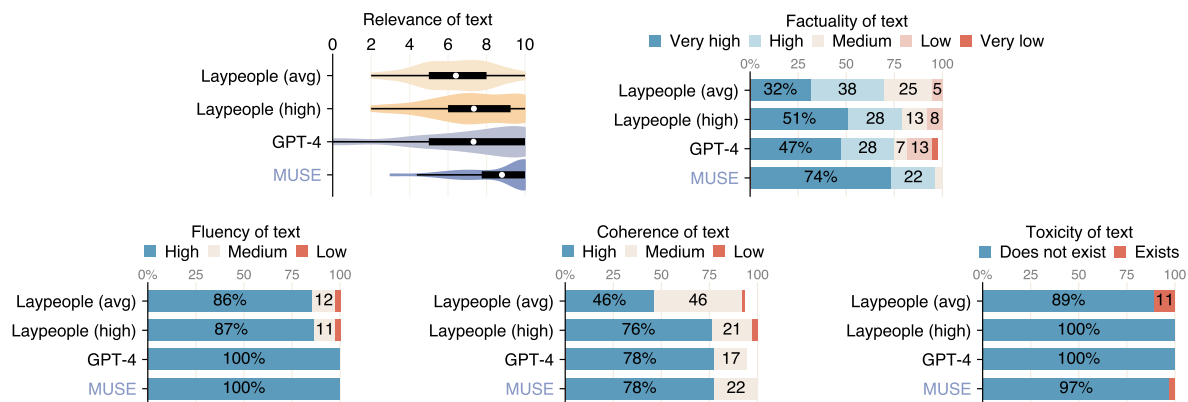


(c) Quality of responses in references.

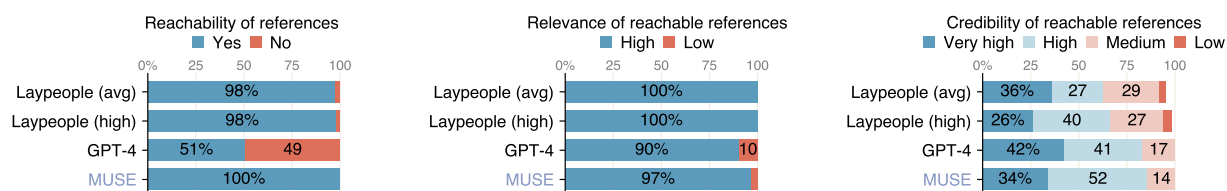
**Figure S9.** Expert evaluation results for political and international-affair posts (#=80).



(a) Quality of responses in identifying and explaining (in)accuracies.

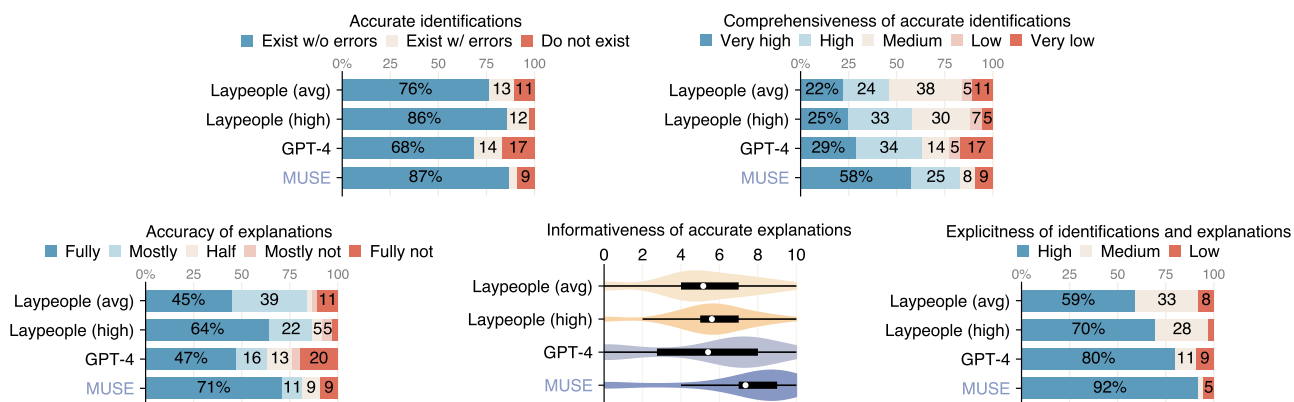


(b) Quality of responses in generated text.

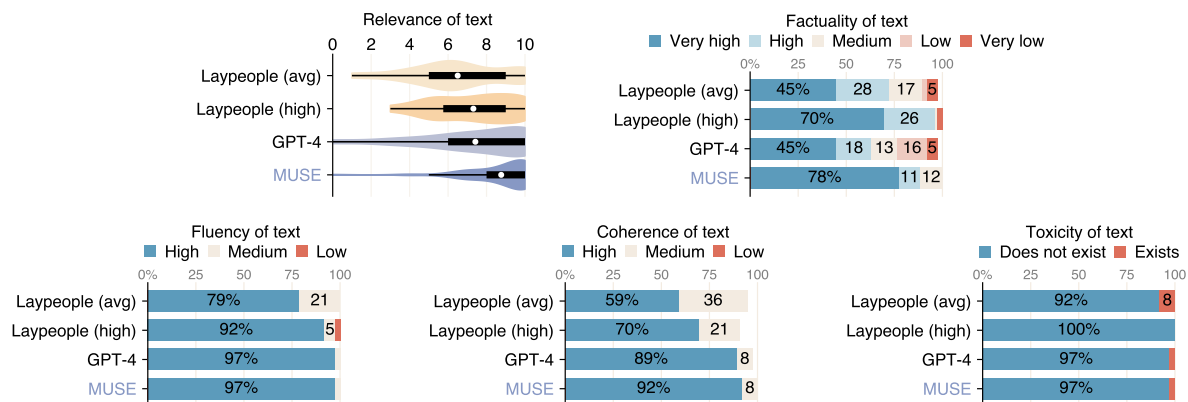


(c) Quality of responses in references.

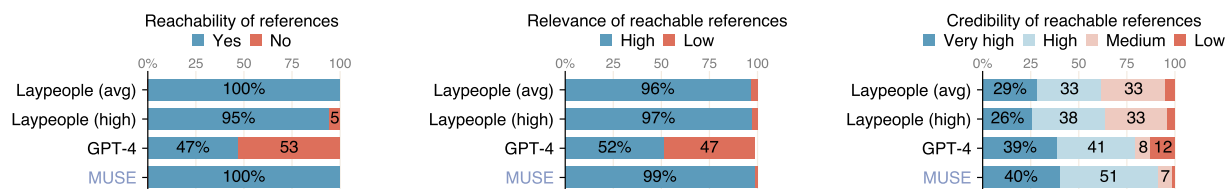
**Figure S10.** Expert evaluation results for economic and business posts (#=38).



(a) Quality of responses in identifying and explaining (in)accuracies.

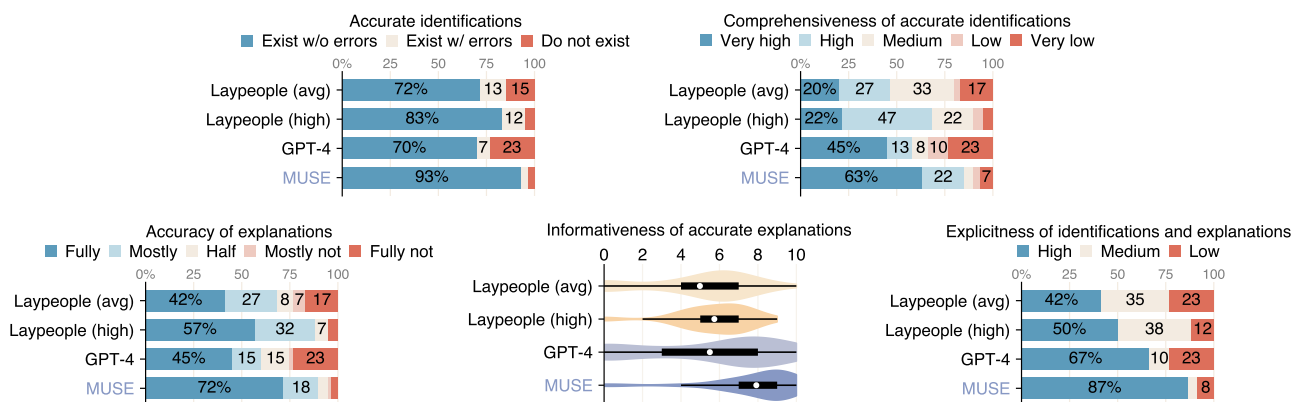


(b) Quality of responses in generated text.

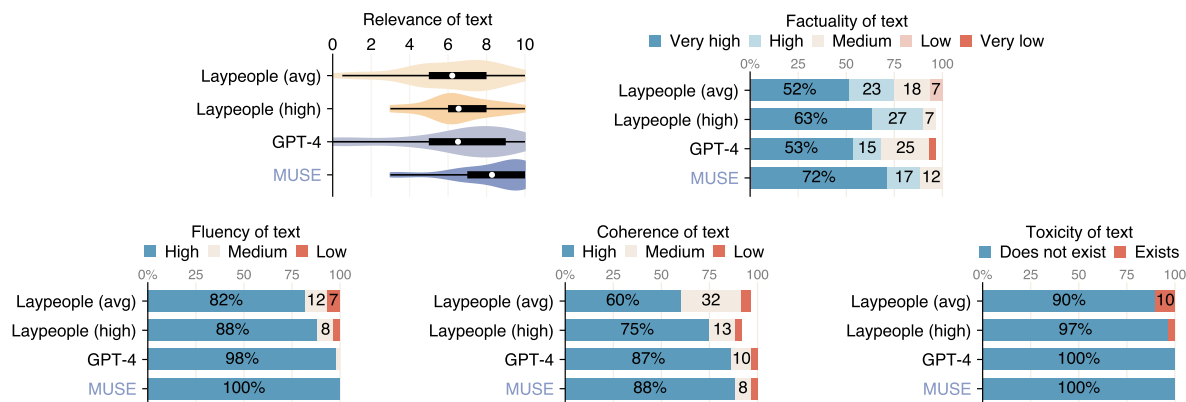


(c) Quality of responses in references.

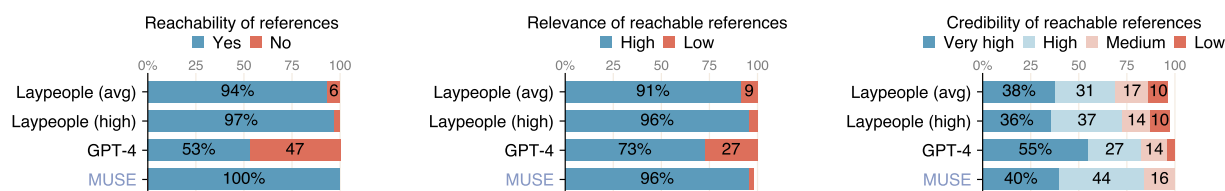
**Figure S11.** Expert evaluation results for crime and law posts (#=38).



(a) Quality of responses in identifying and explaining (in)accuracies.

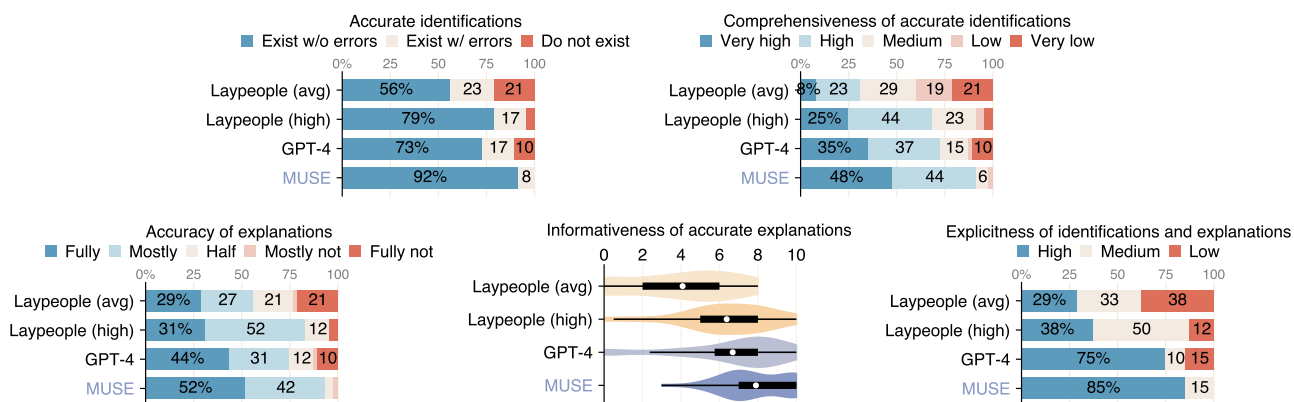


(b) Quality of responses in generated text.

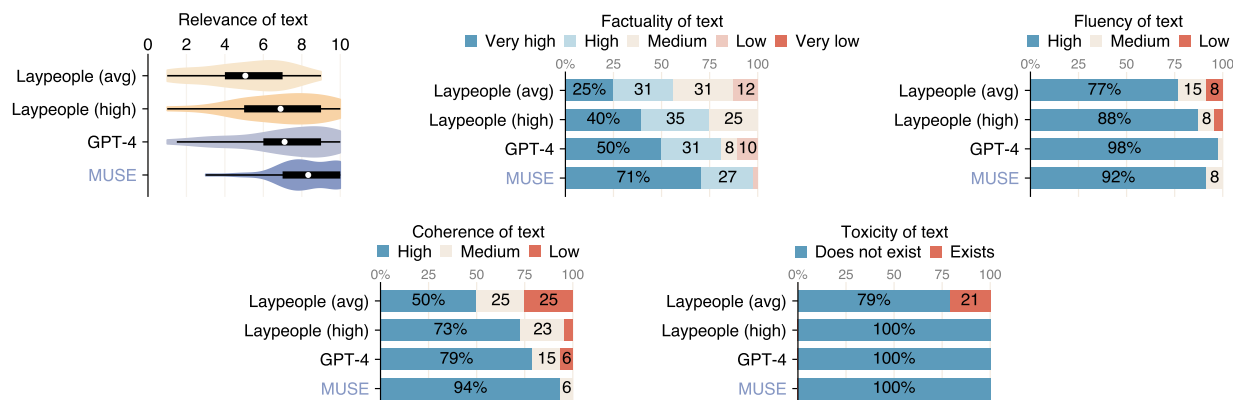


(c) Quality of responses in references.

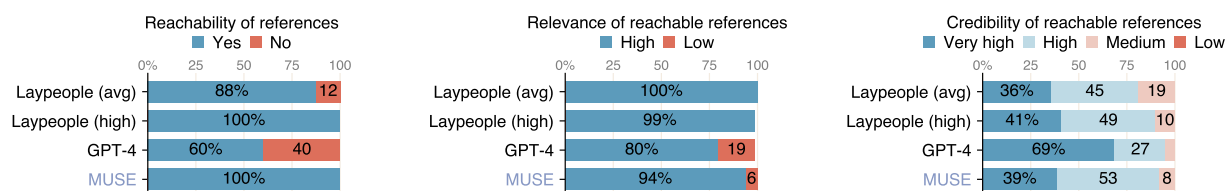
**Figure S12.** Expert evaluation results for social-issue and human-right posts (#=30).



(a) Quality of responses in identifying and explaining (in)accuracies.



(b) Quality of responses in generated text.



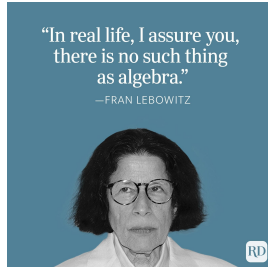
(c) Quality of responses in references.

**Figure S13.** Expert evaluation results for health and medicine posts (#=24).





(a)



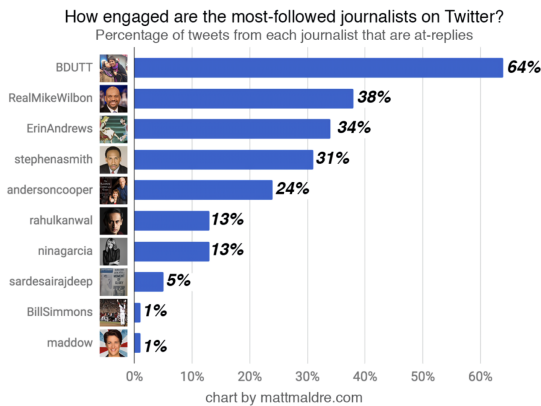
(b)



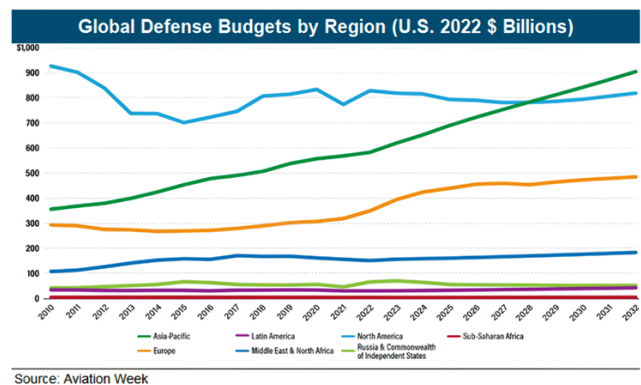
(c)



(d)



(e)



(f)



09:57 · 22/4/2023 · 4,011 Views

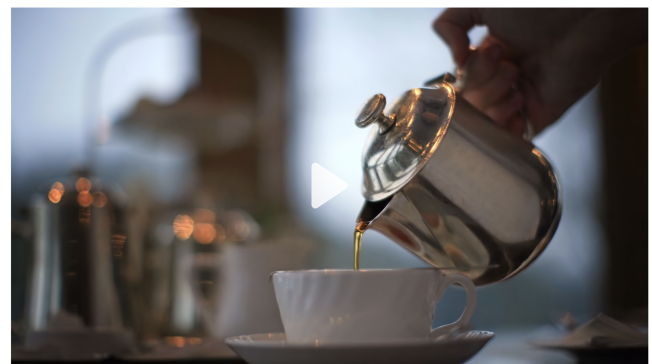
1 Retweet 1 Quote 5 Likes 1 Bookmark



(g)

## 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



The health benefits of tea

01:10 · Source: CNN

(h)

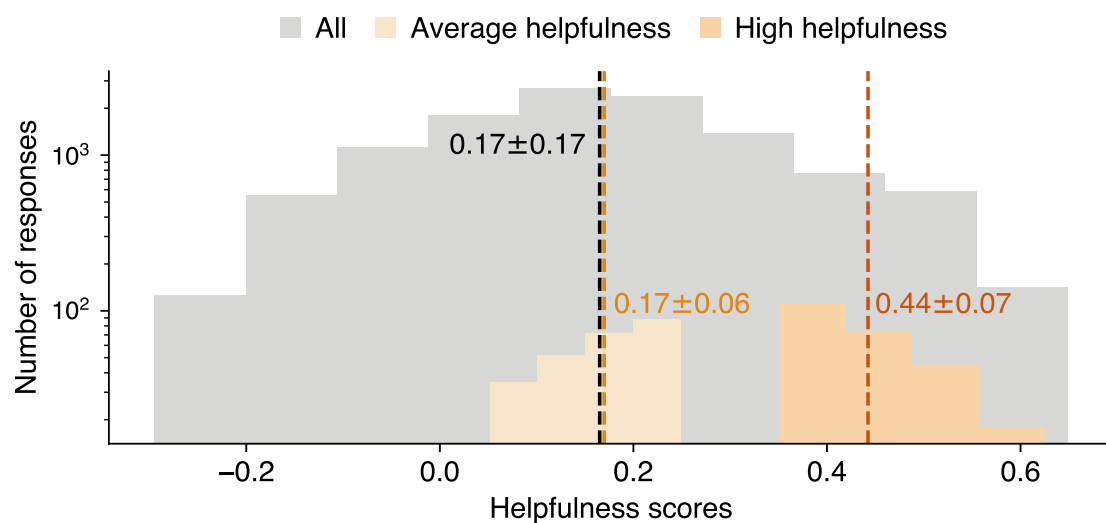
**Figure S14.** Images as examples used for informative image captioning with in-context learning.

```

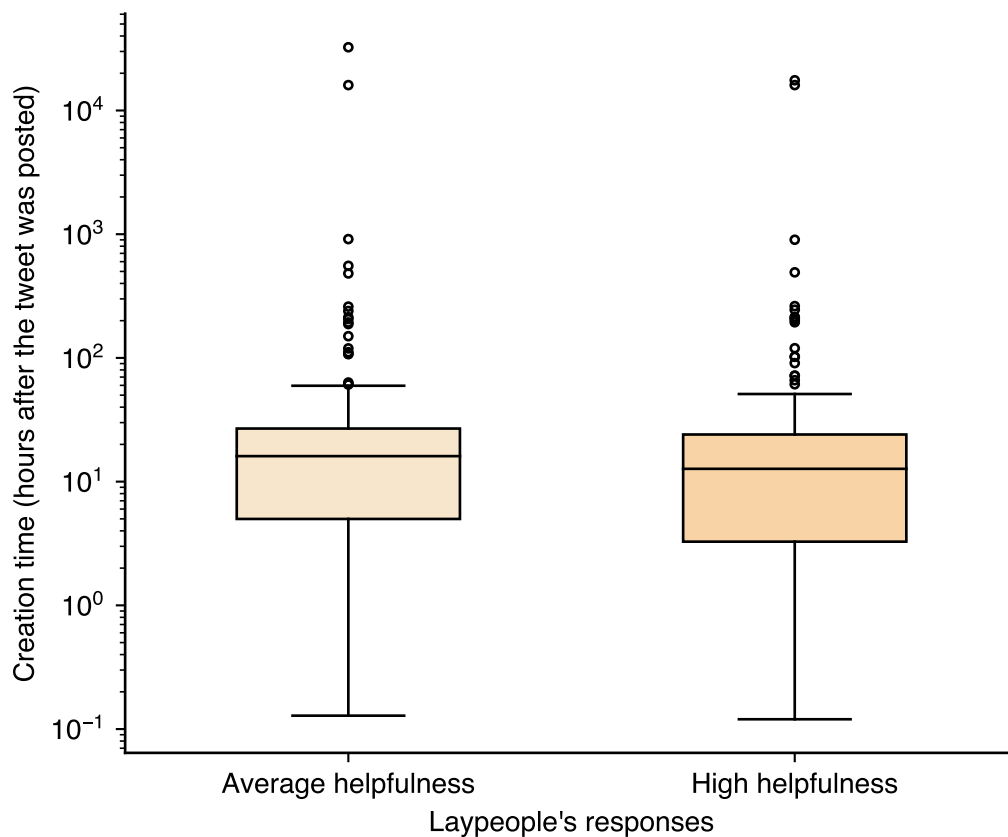
1 Describe an image in an informative way. Your description should be only based on the given {short
  ↳ caption}, {name of each person}, and {raw text}. If the image is from social media, you should
  ↳ start with "A screenshot of". If the image is a quote from someone, you should start with "A
  ↳ 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% RealMikeWilbon 38% 34% ErinAndrews 31% stephenasmith 24%
  ↳ andersoncooper rahulkanwal 13% 13% ninagarcia 5% sardesaiarajdeep 1% BillSimmons maddow 1% 0% 10%
  ↳ 20% 30% 40% 50% 60% chart by mattmaldre.com}
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 S15.** LLM prompt for informative image captioning.



**Figure S16.** 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 S17.** 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 S18.** 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 S19.** 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 S20.** 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 S21.** 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 S22.** 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 S23.** Annotation instructions (page 6/7, continued on the next page).

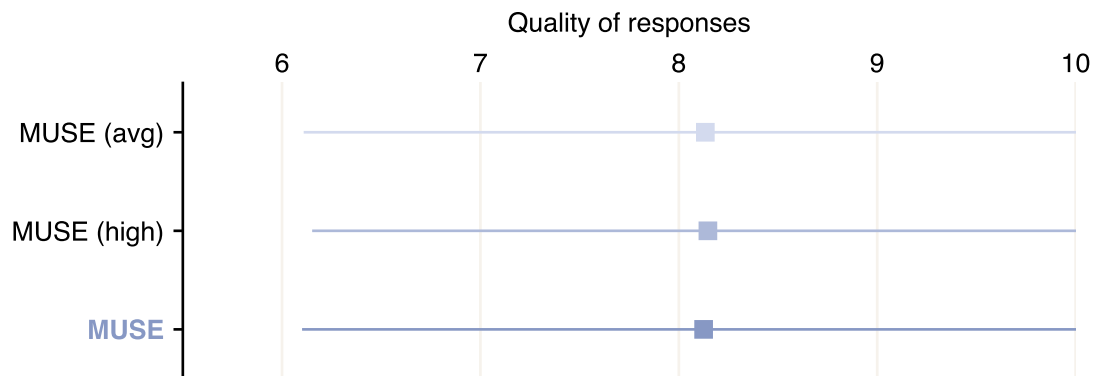
**Q9) How high-quality is the response in general?** Your answer should be a score between 0 and 10, where '0' refers to extremely low quality, and '10' refers to extremely high quality. If two or more responses to the same tweet are of similar quality, they can be scored the same, but **we encourage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

Continue

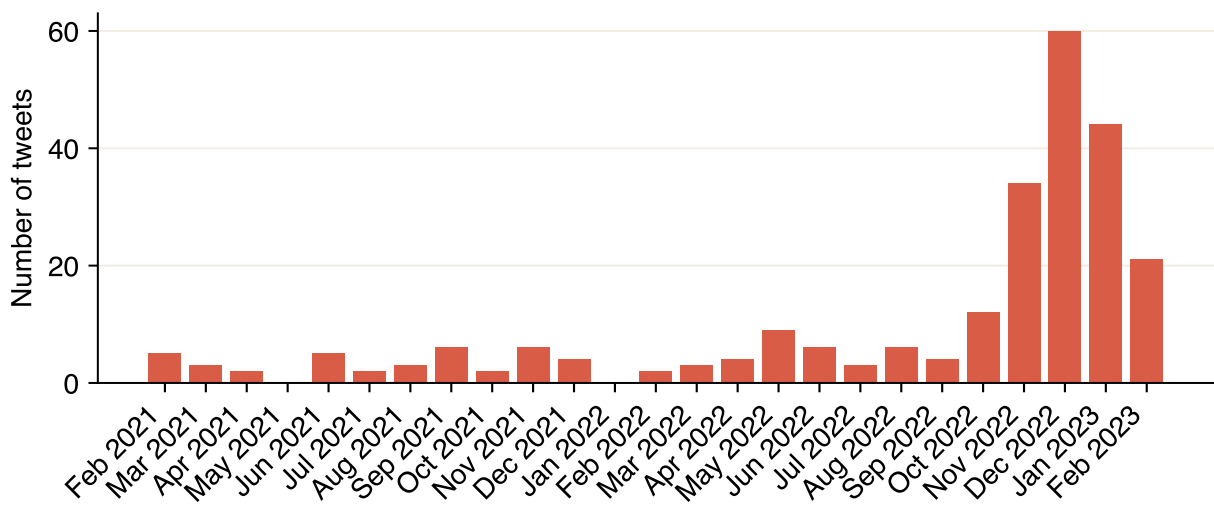
**Figure S24.** Annotation instructions (page 7/7).



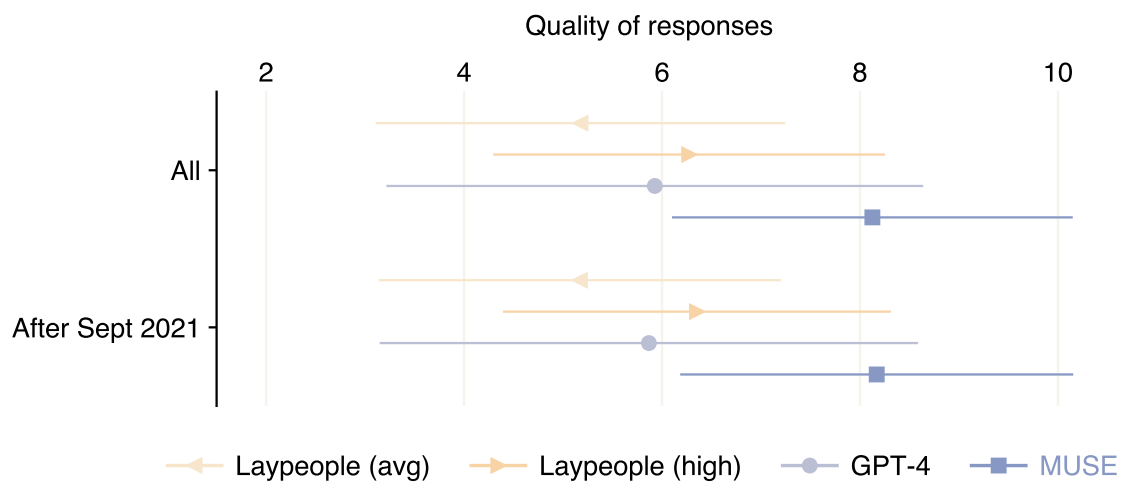




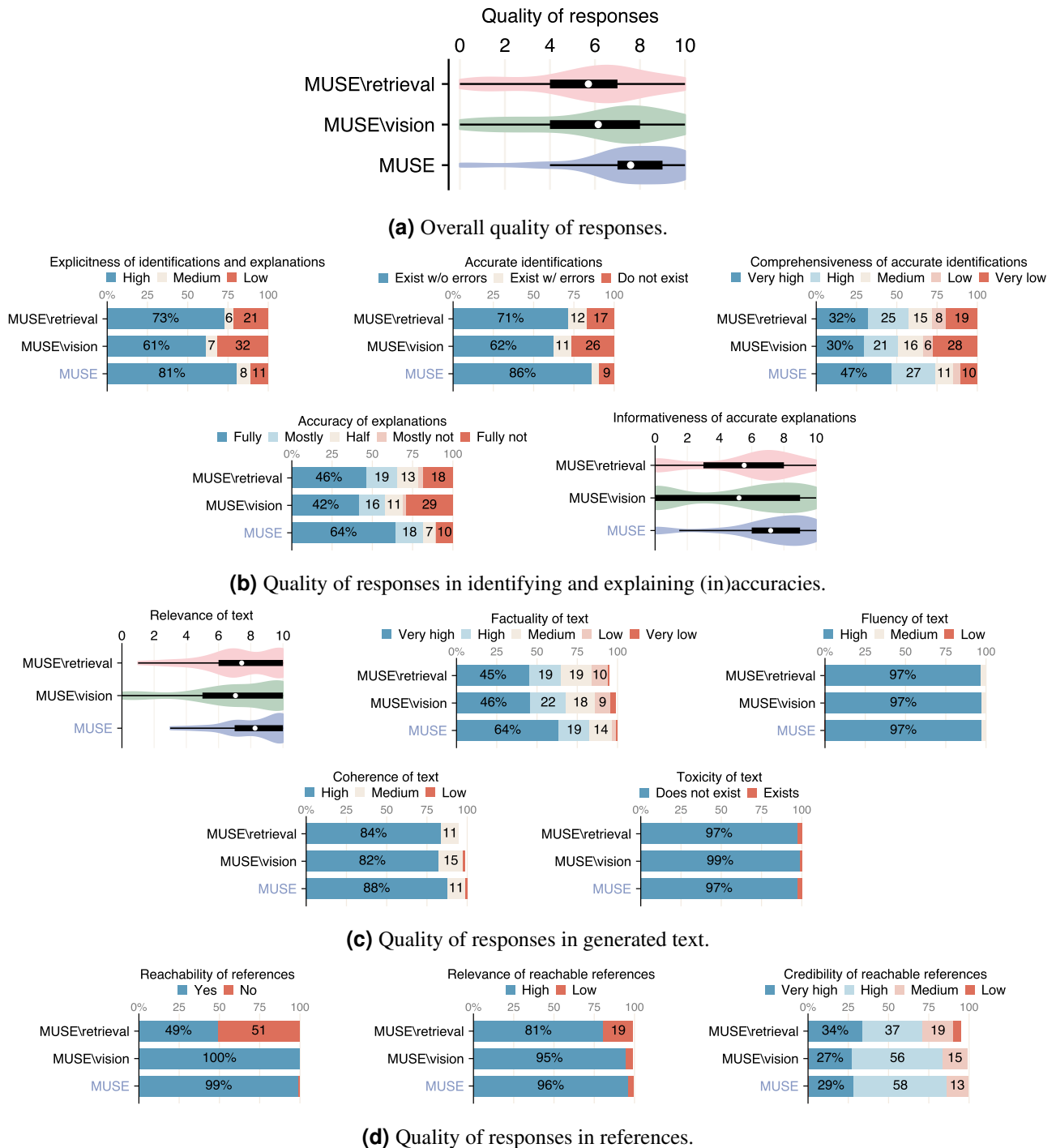
**Figure S26.** Impact of starting times of responding to tweets on MUSE’s performance. The simulated starting time for MUSE (avg): Thirty minutes before the corresponding laypeople’s average-helpfulness responses was created (median: 16 hours after the corresponding tweet was posted; Supplementary Fig. S17). 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. S17). The simulated starting time for MUSE: The post time of the corresponding tweet (i.e., 0 hours after the corresponding tweet was posted).



**Figure S27.** Distribution of post times of tweets in X Community Notes included in our analyses.



**Figure S28.** Impact of post times of tweets on the performance of MUSE and baselines.



**Figure S29.** 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.