Customized large language models can outperform Community Notes in correcting misinformation

Xinyi Zhou[1,2,*], **Ashish Sharma**[1,3], **Amy X. Zhang**[1,*], **and Tim Althoff**[1,*]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
[2]Current affiliation: Computer Science Department, Boise State University, Boise, ID 83702, USA
[3]Current affiliation: Microsoft Corporation, Redmond, WA 98052, USA
[*]To whom correspondence should be addressed. Email: xinyizhou@boisestate.edu, axz@cs.washington.edu, althoff@cs.washington.edu

arXiv:2403.11169v6 [cs.CL] 10 Sep 2026

Abstract

Addressing misinformation in real-world settings is challenging: content is often multimodal; factuality judgments are nuanced and context-dependent; new events emerge rapidly across domains; corrections must be timely, trustworthy, and politically impartial; and multidimensional, multistakeholder frameworks remain lacking. Crowdsourced fact-checking systems such as Community Notes have gained broad adoption, but timely, scalable coverage remains difficult. We introduce MUSE, which augments large language models (LLMs) with trust-aware retrieval of up-to-date evidence and task-specific multimodal reasoning. Given a piece of content, MUSE identifies whether and which parts may be false or misleading and provides explanations grounded in credible references. We also develop an evaluation framework that assesses expert-rated response quality—including identification accuracy, explanation factuality, and the relevance and credibility of supporting references—as well as user perceptions. Across social media posts spanning modalities, domains, political leanings, misinformation tactics, and popularity, MUSE consistently produces high-quality responses, including for content not previously fact-checked online, and outperforms even highly rated Community Notes by 29%. It also improves participants' recognition of misinformation by 10%. Our work establishes a general methodological and evaluative framework for timely, scalable, and trustworthy correction of misinformation.

Keywords Misinformation, large language model, retrieval-augmented generation, multimodality, social media

Significance Statement

False and misleading content is pervasive and difficult to address at scale, with serious implications for democratic processes, economic stability, public health, and social cohesion. We present a framework for timely, scalable, and trustworthy correction of misinformation. MUSE identifies potentially misleading claims in multimodal content and explains its assessments using evidence-grounded references from credible publishers, enabling auditable verification. Evaluations with fact-checking experts and laypeople across diverse real-world content demonstrate its effectiveness.

Introduction

Misinformation can undermine democratic processes, economic stability, public health, and social cohesion[1–6]. These risks are especially acute on social media, where false content can spread rapidly and the

surrounding interface and habitual sharing can impede truth discernment[7–9]. Correcting misinformation in this setting is difficult: posts may be multimodal; claims can be partly accurate yet misleading through omitted context or spurious causal framing; new events emerge quickly across domains; and useful corrections must be timely, explanatory, evidence-grounded, trustworthy, and politically impartial[3,10–12]. Context-rich corrections can improve discernment, whereas delayed or selectively applied corrections have less impact and may create an implied-truth effect for unlabeled content[4,13–15].

Professional fact-checkers and crowdsourced systems have become prominent responses to online misinformation, with aggregated crowd judgments often matching professional fact-checkers' assessments[16,17]. X's Community Notes is a leading example: explanatory notes are perceived as more trustworthy than context-free warnings and can reduce subsequent engagement with misleading posts, while citations to external, less politically biased sources are associated with higher helpfulness ratings[18–21]. Yet these benefits do not guarantee timely or comprehensive coverage. In our February 2023 snapshot, 88% of suspicious posts received no response and 93% received no highly rated response within the first hour (Supp. Fig. S1). Scalable methods that preserve the strengths of evidence-rich, auditable notes while reducing these coverage and latency constraints are therefore needed.

LLMs offer a complementary path to scale. Prior work has explored their use in detecting and verifying claims, generating rationales for model judgments, synthesizing crowd-written notes, and drafting full-length fact-checks[22–29]. These advances are important, but existing systems often address only one stage of the fact-checking workflow or operate under restricted conditions, such as isolated claims, text-only inputs, a single domain or deception type, or reliance on existing fact-checks or crowd-written notes. Evaluation poses a related challenge. Static misinformation benchmarks can be susceptible to training-data contamination and memorization, making it difficult to assess performance on emerging events[30]. Moreover, lexical-overlap metrics such as ROUGE, and even automated model-based grading, cannot on their own determine whether a correction accurately identifies the misleading content, provides a factual and complete explanation, and cites relevant and credible sources[12]. End-user effects likewise cannot be inferred from benchmark accuracy. AI-based dialogues can reduce conspiracy beliefs, but imperfect LLM-generated fact-checks may fail to improve, and may even reduce, headline discernment; user perceptions can also vary depending on whether AI authorship is disclosed[31–33]. A practical evaluation should therefore combine expert assessment of complete responses and their supporting evidence with measurement of end-user perceptions.

Here, we introduce MUSE (multimodal misinformation correction), a customized LLM framework for timely, scalable, and trustworthy misinformation correction. MUSE augments LLMs with trust-aware retrieval of up-to-date evidence and task-specific multimodal reasoning. Given a social media post, it identifies (in)accurate and potentially misleading elements and explains its assessment using auditable references from credible sources. We evaluate MUSE on real posts spanning modalities, domains, political viewpoints, misinformation tactics, and user popularity including posts for which we found no prior online fact-check. Fact-checking experts assess 13 dimensions spanning identification accuracy, explanation factuality, the relevance and credibility of references, and overall response quality; a study with 988 participants assesses end-user perceptions. Across this diverse setting, MUSE achieves an approximately 29% higher expert-rated response-quality score than highly rated Community Notes. It also improves participants' recognition of misinformation by approximately 10%. These findings demonstrate the promise of customized LLMs for correcting real-world misinformation and provide a methodological and evaluative framework for designing and assessing such systems.

Approach

MUSE is an end-to-end system that automatically responds to content containing text, with or without visuals. It identifies accurate and inaccurate claims, explains its judgments, and provides reference links. MUSE comprises three interconnected modules: (1) an LLM for response generation, (2) a hierarchical, credibility-aware web retriever for evidence retrieval, and (3) a multimodal integrator that equips the first two modules with multimodal capabilities. Fig. 1 illustrates the pipeline.

Given text-only content, MUSE first generates a set of search queries from the input. Each query is submitted to a web search engine, which returns links to up-to-date web content relevant to the query. After scraping the linked pages, MUSE computes the relevance of each page to the input and removes irrelevant pages. It then evaluates the credibility of the remaining pages using their publishers' factuality and bias ratings and prioritizes pages from publishers with high factuality and low bias. Next, MUSE extracts evidence from the selected pages. This evidence may refute false claims or provide context supporting some or all of the input claims. Finally, MUSE provides the extracted evidence and corresponding source links to an LLM, which generates a response to the input.

Social media content often includes extraneous material that neither requires verification nor contributes to correction, such as unverifiable opinions or emojis, as in Fig. 1. Generating targeted queries rather than submitting the post verbatim therefore denoises the input and improves retrieval. Multiple queries also help decompose posts with several claims and reduce incomplete assessments. Although retaining all retrieved pages could increase the available evidence, it can substantially raise the computational cost of MUSE because of the volume of page content that often needs to be processed. We also find that low-relevance pages increase hallucinations in generated responses (Supp. Fig. S2). Moreover, MUSE ranks pages by publisher factuality and bias and extracts evidence from the most factual and least biased sources first. It stops after obtaining sufficient refuting evidence—defined as at least two pages that refute the misinformation—or exhausting all credible pages.

For inputs containing both text and images, MUSE performs informative image captioning to generate a textual description of each image, enabling downstream LLM-based processing. Even when the foundation LLM supports multimodal input, a general-purpose model may not support the identification of public figures (Supp. Fig. S3). Although such restrictions can be motivated by privacy concerns, identifying public figures can be important for misinformation verification and correction. Our end-to-end design seeks to balance these privacy considerations with correction effectiveness. Specifically, MUSE augments image-captioning models with capabilities that existing models often fail to capture[34], including public-figure identification and optical character recognition (OCR). For example, a strong image-captioning model[35] may describe the visual misinformation in Fig. 1 only as "a list of banned books in Florida," omitting the book titles needed for verification and correction. MUSE also performs multimodal web searches and computes multimodal relevance scores to remove irrelevant pages.

Experiments

Data MUSE was evaluated using X Community Notes, a crowdsourced fact-checking system shown to mitigate misinformation and now expanding to other platforms[18,36]. Each free-response fact-check, often written by a lay contributor, receives a helpfulness score aggregated from ratings by users with diverse backgrounds, including different political ideologies. Responses with sufficiently high scores are displayed publicly on the corresponding post[18] (Supp. Fig. S4). We included 247 posts published between February 2021 and February 2023, each with at least one high- and one average-helpfulness response as of February 2023. Helpfulness scores ranged from -0.3 to 0.6 and were approximately normally distributed (mean: 0.17, SD: 0.17; Supp. Fig. S5). We defined high helpfulness as a score ≥ 0.35; these responses

had a mean of 0.44, above X's suggested threshold of 0.4[18]. Average-helpfulness responses had scores in $[0.05, 0.25)$ and a mean of 0.17, matching the overall mean. The expert annotations described below indicated that 48% of posts contained both accurate and inaccurate or potentially misleading claims, 46% were inaccurate or potentially misleading, 3% were fully accurate and not misleading, and the remaining 3% were unverifiable (Supp. Fig S6).

Baselines We compared responses from Community Notes contributors, MUSE, and its variants. For each post, the evaluation set included one high- and one average-helpfulness Community Notes response. Because these responses were written before our evaluation, MUSE could otherwise benefit from retrieving more recently published web pages. We therefore generated two temporally matched responses per post, limiting retrieval to pages published no later than 30 minutes before the corresponding high- or average-helpfulness response was created (Supp. Fig. S7). To evaluate MUSE's ability to immediately respond to misinformation, we also generated one response per post using only web pages published before the post's post time. As ablations, we evaluated MUSE's foundation LLM alone and, for posts containing images, two additional variants: MUSE\vision, which retained credibility-aware retrieval but disabled vision, and MUSE\retrieval, which retained vision but disabled retrieval.

Implementation Details Our primary implementation of MUSE used GPT-4, whose September 2021 cutoff for most training data predates over 89% of the evaluation posts, mitigating the risk of training-data contamination. We also evaluated MUSE with GPT-5.5, a recent multimodal reasoning model, and the open-weight Llama 3 (Discussion). In the prompting-based pipeline, the LLM generated search queries, extracted evidence, produced evidence-grounded responses, and combined BLIP-2 image captions[35], Amazon Rekognition celebrity labels, and Amazon Textract OCR outputs into informative image descriptions using in-context learning (Supp. Figs. S8–S12). Google Search served as the web retriever.

To estimate relevance, we encoded the text of each input item and candidate webpage using the pretrained Sentence Transformer `msmarco-distilbert-base-tas-b`[37,38] and their visual content using the pretrained Vision Transformer `dino-vitb8`[39], then computed textual and visual similarity scores. We calibrated the relevance thresholds on 100 webpage–input pairs comprising the top 10 search results for each of 10 randomly selected input items excluded from the evaluation set. We manually labeled the pairs and selected thresholds yielding no false-positive inclusions in the calibration set.

We operationalized publisher-level factuality and bias using human-curated ratings from Media Bias/Fact Check, a media-rating organization with broad coverage that has been used in prior research[40,41]. We retained publishers rated "very high," "high," or "mostly factual" and classified as "least biased," "left-center," "right-center," or "pro-science" (Supp. Table S1). We then assigned the 2,445 eligible publishers to three priority tiers: high ("very high" factuality and "least biased" or "pro-science"; $n = 118$; e.g., CDC, Nature, Pew Research Center), medium (remaining publishers with at least "high" factuality; $n = 2{,}123$; e.g., Britannica, Statista), and low (remaining eligible publishers; $n = 204$; e.g., Forbes, USA Today).

Expert Evaluation With support from Hacks/Hackers, an international grassroots journalism organization, we recruited fact-checking experts through its email list. Of 15 respondents, we selected the 12 with the most relevant experience and at least fluent English proficiency. Five (42%) had 1–3 years of domain experience, three (25%) had 4–6 years, one (8%) had 7–9 years, and three (25%) had 9+ years. Nine (75%) were native English speakers, and three (25%) were fluent. The UW IRB determined the study exempt (STUDY00017831).

The study had two phases. In Phase I, we remotely explained the annotation protocol (Supp. Fig. S13–S19), demonstrated the annotation interface (Supp. Fig. S20), and asked participants to complete three practice tasks with explanations. We reviewed their responses and provided feedback to resolve misun-

derstandings. Two participants withdrew, and all Phase-I data were excluded from analysis. We divided the remaining ten participants into five pairs and randomly assigned each 26 or 27 Phase-II tasks. Seven tasks (approximately 30%) overlapped within each pair to measure inter-annotator agreement. All ten completed Phase II and received a $450 Amazon gift card.

Recruited experts evaluated responses to the same post across four dimensions. For *identifying and explaining factual (in)accuracies*, they assessed explicitness, identification correctness and comprehensiveness, and explanation accuracy and informativeness. For *generated-text quality*, they rated relevance, factuality, fluency[42,43], coherence[42], and toxicity. For *reference quality*, they assessed reachability, relevance to the generated text, and credibility. Finally, considering all 13 criteria, they rated *overall correction quality* on a 10-point scale. Evaluators were blinded to the approach used to generate each response.

We measured inter-rater agreement using weighted Cohen's κ, except for toxicity, fluency, and reference relevance, whose skewed labels warranted observed agreement (0.96, 0.86, and 0.81)[44]. Under standard interpretations[45], agreement was substantial for reference reachability (0.79), moderate for overall quality, informativeness, comprehensiveness, and text relevance (0.41–0.51), and fair for all other κ-scored criteria (0.28–0.40), consistent with known expert disagreement in misinformation assessment[16,46].

Measuring End User Perceptions of Corrections Previous research suggests that bias against AI-generated content can reduce its effectiveness and perceived trustworthiness[33]. To assess MUSE's potential real-world impact, we conducted a Prolific study with 988 U.S. participants, following Costello et al.[31] (Supp. Fig. S21–S22). The sample was representative of the U.S. population by sex, age, and political affiliation. Each participant viewed a social media post from the expert-evaluation dataset and a corresponding response, then rated their belief in and intention to share the post before and after the response, as well as the response's trustworthiness (all on 1–7 scales). They were randomly assigned to either no source disclosure or explicit disclosure that the correction was human-written or AI-generated.

Results

MUSE-generated responses exceed even high-helpfulness Community Notes in overall quality. MUSE achieved a mean quality score of 8.1/10, exceeding laypeople's high-helpfulness responses by 29% (6.3; $p = 3 \times 10^{-48}$; $N = 464$), GPT-4 by 37% (5.9; $p = 4 \times 10^{-42}$; $N = 464$), and laypeople's average-helpfulness responses by 56% (5.2; $p = 5 \times 10^{-81}$; $N = 462$; Mann–Whitney U tests unless noted; Fig. 2a). Here, lay responses were written an average of 14 hours after posting, whereas MUSE used only web pages published before the post.

MUSE better identifies and explains (in)accuracies. Overall, 89% of MUSE responses explicitly identified and explained (in)accuracies, 16% more than GPT-4, 29% more than high-helpfulness notes, and 43% more than average-helpfulness notes (Fig. 2b). Moreover, 91% contained at least one correct identification and no mistakes, compared with 80%, 72%, and 65%, respectively (Fig. 2c); 61% identified all (in)accuracies, versus 38%, 26%, and 17% (Fig. 2d). Fully accurate explanations occurred in 70% of MUSE responses, compared with 55% for high-helpfulness notes, 47% for GPT-4, and 37% for average-helpfulness notes (Fig. 2e). Its mean informativeness score was 7.9, respectively 32%, 36%, and 65% higher (Fig. 2f).

MUSE generates higher-quality text. By augmenting GPT-4 with timely web access and visual analysis, MUSE improved relevance ($p = 10^{-15}$; $N = 460$) and factuality ($p = 2 \times 10^{-20}$; $N = 459$) without reducing fluency ($p = 0.6$; $N = 464$), coherence ($p = 0.1$; $N = 459$), or safety with respect to toxicity ($p = 0.8$; $N = 464$). Compared with high-helpfulness lay responses, its text was more relevant ($p = 2 \times 10^{-30}$; $N = 464$), factual ($p = 4 \times 10^{-6}$; $N = 463$), fluent ($p = 2 \times 10^{-10}$; $N = 464$), and coherent ($p = 10^{-5}$;

$N = 451$); it was also less toxic than average-helpfulness responses ($p = 4 \times 10^{-12}$; $N = 462$). MUSE attained a mean relevance score of 8.7—18% above GPT-4, 21% above high-helpfulness notes, and 43% above average-helpfulness notes (Fig. 2g). Its text was completely factual in 74% of cases, versus 59% for high-helpfulness notes and 45% for GPT-4 (Fig. 2h); nearly all responses were grammatically correct and free of bias, impoliteness, or provocation, and 91% were highly coherent and logical (Fig. 2i–k).

MUSE provides higher-quality references. Among GPT-4's references, 49% returned page-not-found errors, and only 76% of reachable links were relevant (Fig. 2l–m). For MUSE, nearly 100% of links were reachable and 96% were relevant. Its references were also more credible than those in high-helpfulness lay responses ($p = 4 \times 10^{-11}$; $N = 744$; Fig. 2n).

MUSE's superior performance is not driven by potential training-data contamination or reliance on existing fact-checks. Using September 2021 as GPT-4's assumed training-data cutoff, we compared performance on all tweets ($n = 232$) with performance on tweets posted after the cutoff ($n = 207$). Posts and Community Notes from before or during September 2021 might have appeared in GPT-4's training data, potentially inflating performance in ways that would not generalize to future content. Nevertheless, MUSE performed stably on both sets (mean±SD correction quality: 8.1±2.0) and consistently outperformed GPT-4 and even high-helpfulness lay responses (Fig. 3a).

MUSE cited at least one related fact-checking article for 16% of all posts and relied exclusively on such articles for only 6%. After excluding all posts for which MUSE cited a related fact-checking article, its mean response-quality score remained 8.0, close to its mean of 8.1 across all posts. For these posts, the quality of responses generated by MUSE was at least 28% higher than that of GPT-4 and high-helpfulness responses by laypeople (Fig. 3b). Thus, MUSE's advantage cannot be explained primarily by its ability to summarize retrieved professional fact-checks.

MUSE performs robustly across modalities, political leanings, domains, misinformation tactics, and source popularity. For both text-only and multimodal posts, the quality of MUSE-generated responses exceeded that of GPT-4 and even high-helpfulness lay responses by at least 21% (Fig. 3c). We also annotated the political leaning of each post. GPT-4, which has previously been reported to exhibit a liberal bias[47], scored 5% lower on liberal-leaning misinformation than on conservative-leaning misinformation (Fig. 3d). By contrast, MUSE achieved the same mean quality score of 8.3 for both groups and outperformed GPT-4 and high-helpfulness lay responses by at least 26%. We further annotated the topic domain of each post and the tactics that made it, or parts of it, false or potentially misleading. Across domains and misinformation tactics, responses generated by MUSE maintained a mean quality score of approximately 8 and outperformed GPT-4 and even high-helpfulness responses by laypeople by at least 19% (Fig. 3e–f; see example posts using various tactics in Supp. Fig. S23).

Finally, we measured poster popularity by follower count. MUSE's response quality did not vary significantly with popularity and consistently exceeded that of GPT-4 and even high-helpfulness lay responses (Fig. 3g). Community Notes overrepresents relatively popular sources and, like any existing dataset, does not capture the full range of misinformation, including content from obscure accounts. However, it enables the study of high-reach misinformation, which may be especially persuasive and harmful[10] yet remains understudied. More broadly, no misinformation dataset is fully representative because selection effects shape both what content is created and what is annotated. Some biases are substantively meaningful: popular accounts reach larger audiences and may warrant faster assessment. Community Notes is therefore an important public dataset reflecting how social media platforms address potential misinformation[36].

Ablation studies demonstrate the value of retrieval and multimodal integration. GPT-4 can be viewed as an unaugmented variant of MUSE without the proposed retrieval and vision capabilities. Further ablations show that both components contribute substantially: MUSE outperformed the variant without retrieval by 25% and the variant without vision by 33% in overall response quality (Supp. Fig. S24).

Correction timing does not affect MUSE's response quality. For each post, we generated three MUSE responses simulating different start times. Performance was similar (mean±SD response quality: 8.1±2.0) whether correction began immediately after posting or at the median response times of laypeople producing high-helpfulness notes (13 hours) or average-helpfulness notes (16 hours; Supp. Figs. S25 and S7).

MUSE improves misinformation recognition regardless of source disclosure. MUSE-generated corrections significantly increased end-user participants' correct belief that the misinformation was misleading by 10% (from 4.5 to 4.9; Cohen's $d = 0.3$; $p = 0.0001$; Fig. 4a). No significant belief changes were observed for GPT-4 or lay responses of either helpfulness level. None of the corrections significantly affected participants' intention to share the misinformation, which was generally low (approximately 2/7; Fig. 4b). MUSE, GPT-4, and high-helpfulness lay responses were all rated significantly more trustworthy than average-helpfulness lay responses (Fig. 4c). MUSE received a mean trustworthiness score of 4.9, 8% higher than average-helpfulness lay responses (mean: 4.6; $p = 0.004$); the corresponding differences were 6% for GPT-4 ($p = 0.03$) and 10% for high-helpfulness lay responses ($p = 0.0002$).

Source disclosure (human or AI) did not significantly affect trustworthiness or sharing intention (Fig. 4e–f). However, withholding the source increased participants' correct belief by 8% (from 4.4 to 4.7; Cohen's $d = 0.2$; $p = 0.0001$; Fig. 4d), primarily for the less effective GPT-4 and average-helpfulness lay responses. By contrast, MUSE and high-helpfulness lay responses were similarly effective regardless of disclosure. Although prior work reports bias against disclosed AI-generated content[33], our results suggest that highly effective corrections remain robust to source disclosure.

Discussion

As Community Notes-like approaches are increasingly adopted by social media platforms[18,36], we propose MUSE as a practical end-to-end framework for helping users and platforms respond to misinformation reliably, promptly, and transparently. Its simple, training-free pipeline facilitates deployment and maintenance. Despite this simplicity, expert evaluation shows that MUSE produces high-quality corrections that accurately identify and explain inaccuracies, generate effective text, and provide credible supporting references, significantly outperforming even highly rated lay responses across misinformation modalities, political leanings, topic domains, tactics, popularity levels, and claims not previously fact-checked online.

In addition to achieving the highest accuracy and factuality, MUSE produces the most readable responses, with greater explicitness, fluency, and coherence. Its low-toxicity text and credible references may reduce correction backfire effects[48–50], while identifying inaccuracies comprehensively and at scale may mitigate the implied truth effect[15]. We further show that MUSE-generated responses improve people's recognition of misinformation, regardless of whether their source is disclosed. These corrections can be generated within minutes of suspicious content appearing: MUSE responds to a post in two minutes on average on a basic M1 CPU setup with 16 GB of memory. Our primary design objective was to maximize correction quality. By transparently linking retrieved evidence, MUSE enables users to verify responses while signaling insufficient evidence, providing accuracy nudges, and expressing uncertainty when appropriate (Supp. Fig. S26).

Responses generated by MUSE cost approximately 0.2 USD per post, with correction quality as our primary design objective. Given the task's complexity and the substantial quality improvements, this

cost is modest relative to alternatives. For example, crowdsourced lay judgments cost about 0.9 USD per article merely to assess whether its headline and lede contain misinformation, without providing an explanation or correction[16]. We also replaced GPT-4 with the open-weight Llama 3 70B model. Holding all other implementation details fixed, the variants performed comparably in our qualitative analysis (Supp. Fig. S27), suggesting that MUSE is not tied to a specific foundation model and may support lower-cost deployment.

We further replaced GPT-4 with GPT-5.5, an advanced multimodal reasoning model, and found that such models alone remain insufficient. Vanilla GPT-5.5 still hallucinated references: only 70% of links were reachable, based on an HTTP check that achieved an F1 score of 0.91 on expert-annotated URLs. Of the reachable links, 81% were relevant and supported the associated claims, yielding a reference-hallucination rate of 43%, vs 2% for MUSE. We assessed relevance by embedding the generated text and crawled reference content using OpenAI's `text-embedding-3-small` and classifying each pair by cosine similarity. A threshold of 0.1 achieved the highest validation F1 score of 0.86 on expert-annotated URLs. Although recent commercial models integrate web search, they remain opaque about which sources are retrieved and how evidence is selected, limiting trustworthiness[21]. Llama 3 and GPT-5.5 have knowledge cutoffs of December 2023 and 2025, respectively, so their training data may include evaluation posts, corrections, or Community Notes. Direct comparisons among Llama 3-, GPT-5.5-, and GPT-4-based versions of MUSE are therefore not strictly fair. Our goal is not to propose a state-of-the-art LLM, but to develop a scalable misinformation-response system whose corrections outperform or are comparable to highly rated Community Notes. Stronger future models may make MUSE even more effective.

This work has several limitations. First, MUSE supports English text and image inputs but not video. Second, we evaluated it on X, a single but widely used platform where many users consume news and false information can spread faster than truth[7,51]. We selected X because its transparent Community Notes system has been shown to reduce misinformation diffusion[18–20]. Third, as a retrieval-based framework, MUSE depends on sufficiently relevant and credible online evidence. In information voids, such as breaking news, insufficient evidence may prevent reliable verification; the system should therefore communicate uncertainty rather than make definitive judgments. Although Supp. Fig. S26 illustrates this behavior, we did not systematically evaluate information-void detection or abstention reliability. Fourth, our modest Community Notes sample consisted largely of posts containing inaccurate or potentially misleading elements, while often alongside accurate claims, limiting broader generalizability and assessment of truth discernment; end-user outcomes were immediate self-reports[52]. Larger expert evaluations and field studies are needed.

References

1. Lazer, D. M. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
2. Zhou, X. & Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)* **53**, 1–40 (2020).
3. Watts, D. J., Rothschild, D. M. & Mobius, M. Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci.* **118**, e1912443118 (2021).
4. Van Der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat. Medicine* **28**, 460–467 (2022).
5. Moore, R. C., Dahlke, R. & Hancock, J. T. Exposure to untrustworthy websites in the 2020 US election. *Nat. Hum. Behav.* 1–10 (2023).
6. Forum, W. E. Global risks report 2024 (World Economic Forum, 2024).

7. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).

8. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Sci. Adv.* **9**, eabo6169 (2023).

9. Ceylan, G., Anderson, I. A. & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. *Proc. Natl. Acad. Sci.* **120**, e2216614120 (2023).

10. Traberg, C. S. Misinformation: broaden definition to curb its societal influence. *Nature* **606**, 653–653 (2022).

11. Dufour, N. *et al.* AMMeBa: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697* (2024).

12. Augenstein, I. *et al.* Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **6**, 852–863 (2024).

13. Porter, E., Velez, Y. & Wood, T. J. Factual corrections eliminate false beliefs about COVID-19 vaccines. *Public Opin. Q.* **86**, 762–773 (2022).

14. Bak-Coleman, J. B. *et al.* Combining interventions to reduce the spread of viral misinformation. *Nat. Hum. Behav.* **6**, 1372–1380 (2022).

15. 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).

16. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).

17. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Crowds can effectively identify misinformation at scale. *Perspectives on Psychol. Sci.* 17456916231190388 (2022).

18. 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).

19. Drolsbach, C. P., Solovev, K. & Pröllochs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus* **3**, pgae217 (2024).

20. Slaughter, I., Peytavin, A., Ugander, J. & Saveski, M. Community notes reduce engagement with and diffusion of false information online. *Proc. Natl. Acad. Sci.* **122**, e2503413122 (2025).

21. Solovev, K. & Pröllochs, N. References to unbiased sources increase the helpfulness of community fact-checks. *Sci. Reports* **15**, 25749 (2025).

22. Yue, Z. *et al.* Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5628–5643 (2024).

23. Zeng, F. & Gao, W. JustiLM: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions Assoc. for Comput. Linguist.* **12**, 334–354 (2024).

24. Kim, K. *et al.* Can LLMs produce faithful explanations for fact-checking? Towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401* (2024).

25. Qi, P., Yan, Z., Hsu, W. & Lee, M. L. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062 (2024).

26. Khaliq, M. A., Chang, P. Y.-C., Ma, M., Pflugfelder, B. & Miletić, F. RAGAR, Your Falsehood RADAR: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 280–296 (2024).

27. De, S., Bakker, M. A., Baxter, J. & Saveski, M. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM on Web Conference 2025*, 3751–3761 (2025).

28. Wu, J. *et al.* Beyond the crowd: LLM-augmented community notes for governing health misinformation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5152–5171 (2026).

29. Sahnan, D. *et al.* Can LLMs automate fact-checking article writing? *Transactions Assoc. for Comput. Linguist.* **14**, 489–509 (2026).

30. Xu, C. & Yan, N. Triplefact: Defending data contamination in the evaluation of llm-driven fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8808–8823 (2025).

31. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with ai. *Science* **385**, eadq1814 (2024).

32. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl. Acad. Sci.* **121**, e2322823121 (2024).

33. Lim, S. & Schmälzle, R. The effect of source disclosure on evaluation of ai-generated messages. *Comput. Hum. Behav. Artif. Humans* **2**, 100058 (2024).

34. 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).

35. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742 (PmLR, 2023).

36. Kaplan, J. More speech and fewer mistakes. *Meta* (2025).

37. 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*, 3982–3992 (2019).

38. Hofstätter, S., Lin, S.-C., Yang, J.-H., Lin, J. & Hanbury, A. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 113–122 (2021).

39. 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).

40. 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).

41. 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).

42. Yuan, W., Neubig, G. & Liu, P. BARTScore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* **34**, 27263–27277 (2021).

43. Fabbri, A. R. *et al.* SummEval: Re-evaluating summarization evaluation. *Transactions Assoc. for Comput. Linguist.* **9**, 391–409 (2021).

44. 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).

45. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Medica* **22**, 276–282 (2012).

46. 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).

47. Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11737–11762 (2023).

48. 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).

49. Ecker, U. K. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).

50. 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).

51. Walker, M. & Matsa, K. E. News consumption across social media in 2021. *Pew Res. Cent.* (2021).

52. Guay, B., Berinsky, A. J., Pennycook, G. & Rand, D. How to think about whether misinformation interventions work. *Nat. Hum. Behav.* **7**, 1231–1233 (2023).

53. Vu, T. *et al.* FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13697–13720 (2024).

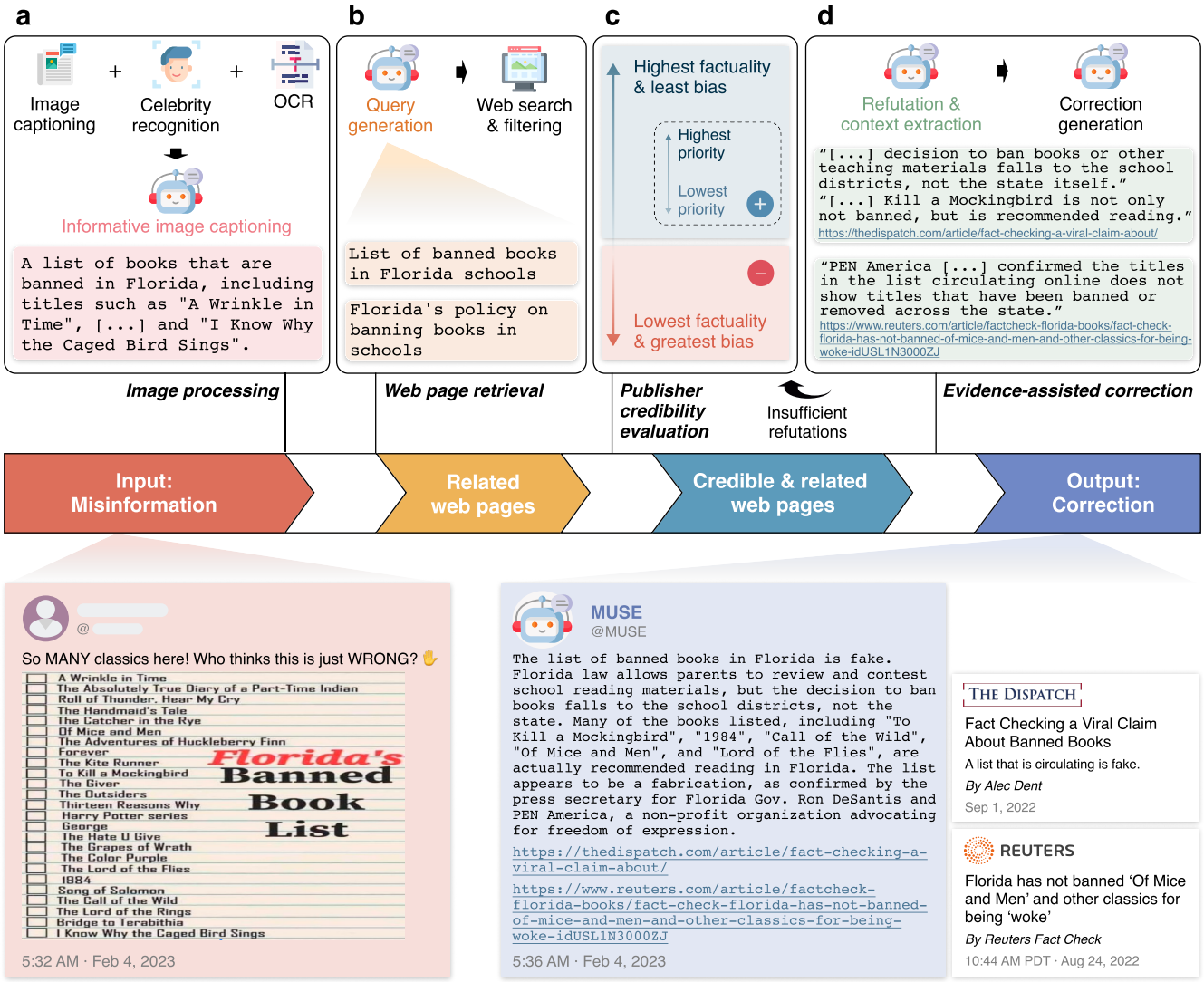


Figure 1. Overview of MUSE, an LLM augmented with trust-aware retrieval of current evidence and task-specific multimodal reasoning to identify and explain (in)accuracies and provide credible supporting references. **a**: Informative image captioning. MUSE augments image captioning models with celebrity recognition and optical character recognition to produce informative image descriptions. **b**: Web page retrieval. MUSE uses LLM-generated queries and a web search engine, then filters pages by their multimodal relevance to the content. **c**: Publisher credibility evaluation. **d**: Evidence-assisted response generation. MUSE ranks publishers by professionally assessed factuality and bias, then uses an LLM to extract refuting or contextualizing evidence, starting with the most factual and least biased sources. It proceeds down the ranking until finding sufficient refutations (at least two pages) or exhausting all credible pages.

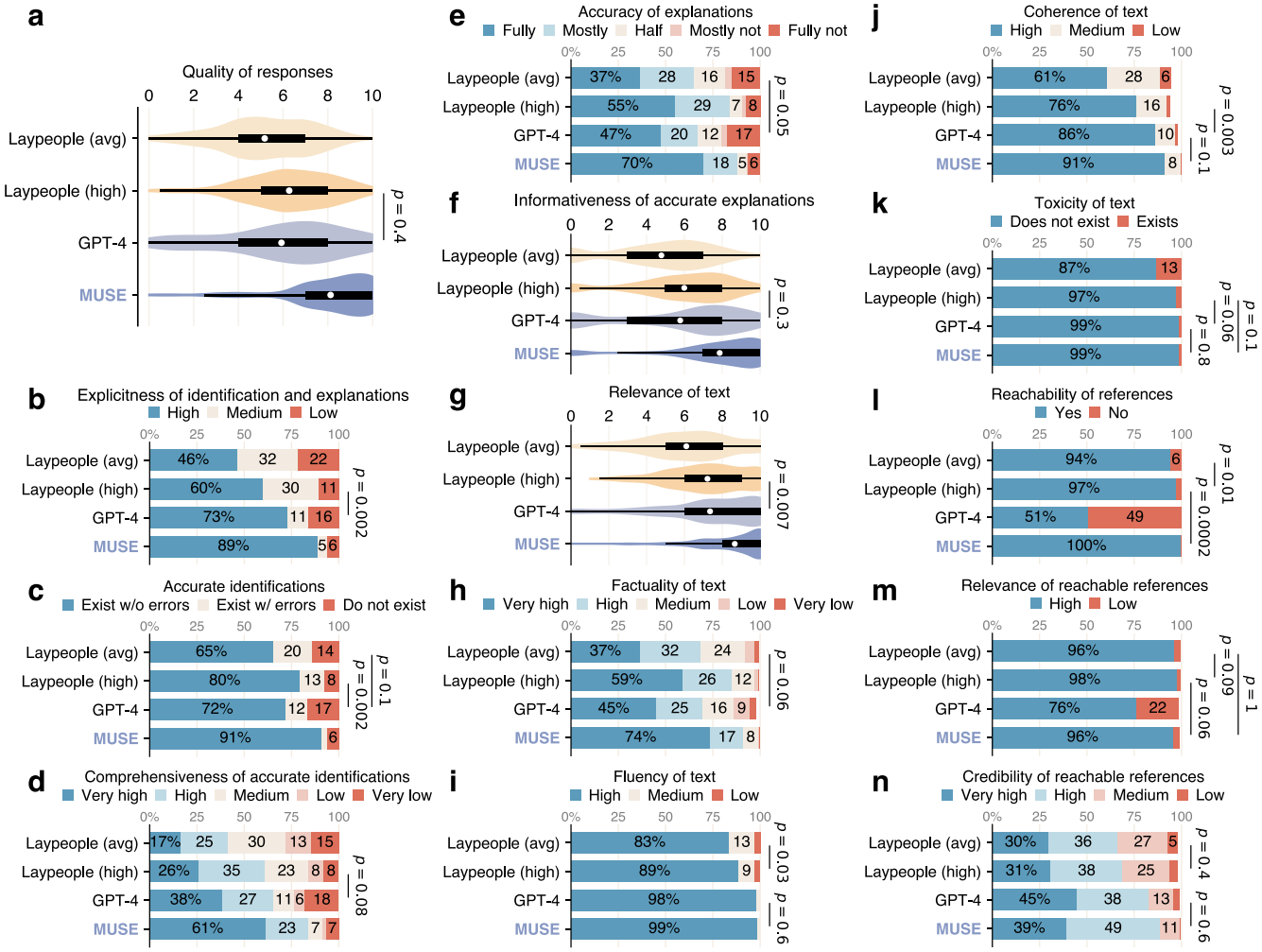


Figure 2. Results of expert evaluation (all pairwise approach comparisons in **a**–**n**: Mann–Whitney U test, $p < 2 \times 10^{-5}$; 84 experiments). **a**: Overall response quality. MUSE-generated responses (mean $\pm$ SD: 8.1 ± 2.0; $n = 232$) scored 29%, 37%, and 56% higher than laypeople's high-helpfulness responses (6.3 ± 2.0; $n = 232$), GPT-4 responses (5.9 ± 2.7; $n = 232$), and laypeople's average-helpfulness responses (5.2 ± 2.1; $n = 230$), respectively. **b**–**f**: Identification and explanation of inaccuracies. Compared with all baselines, MUSE identifies and explains inaccuracies more explicitly (**b**) and comprehensively, makes fewer accuracy misclassifications (**c**–**d**), and provides more accurate and informative explanations (**e**–**f**). **g**–**k**: Text quality. MUSE-generated text is more relevant and factual than all baselines (**g**–**h**), more fluent and coherent than laypeople's high-helpfulness responses, and less toxic than their average-helpfulness responses (**i**–**k**). **l**–**n**: Reference quality. Unlike GPT-4, MUSE rarely hallucinates references and provides more reachable, relevant links (**l**–**m**); its references are also more credible than those in laypeople's responses (**n**).

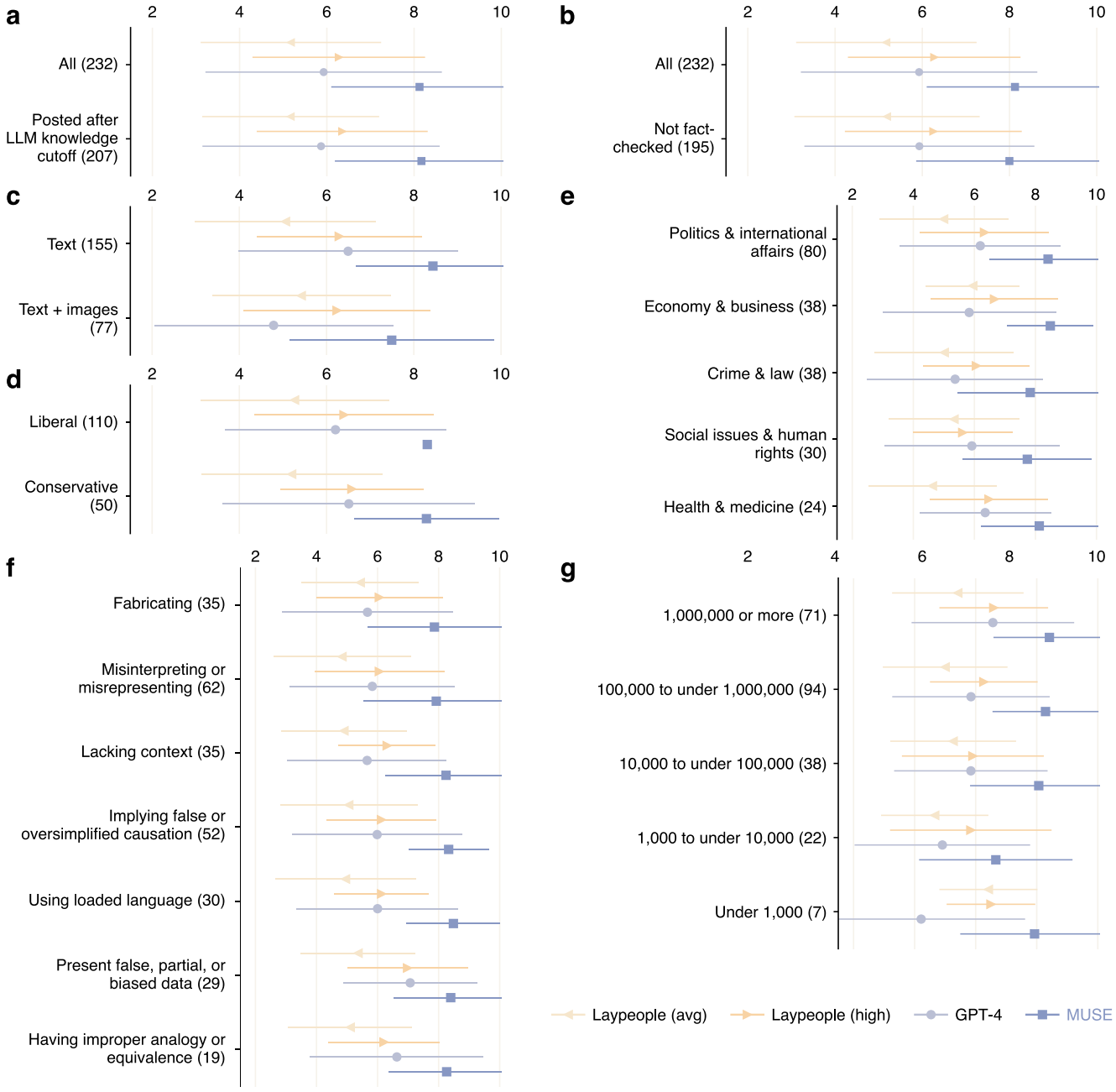


Figure 3. Robustness analyses of MUSE. **a–b**: The superior performance of MUSE does not appear to be driven by potential training-data contamination or reliance on existing fact-checks. **c–g**: Compared with all baselines, MUSE consistently generates higher-quality responses across modalities, political leanings, domains, misinformation tactics, and user popularity.

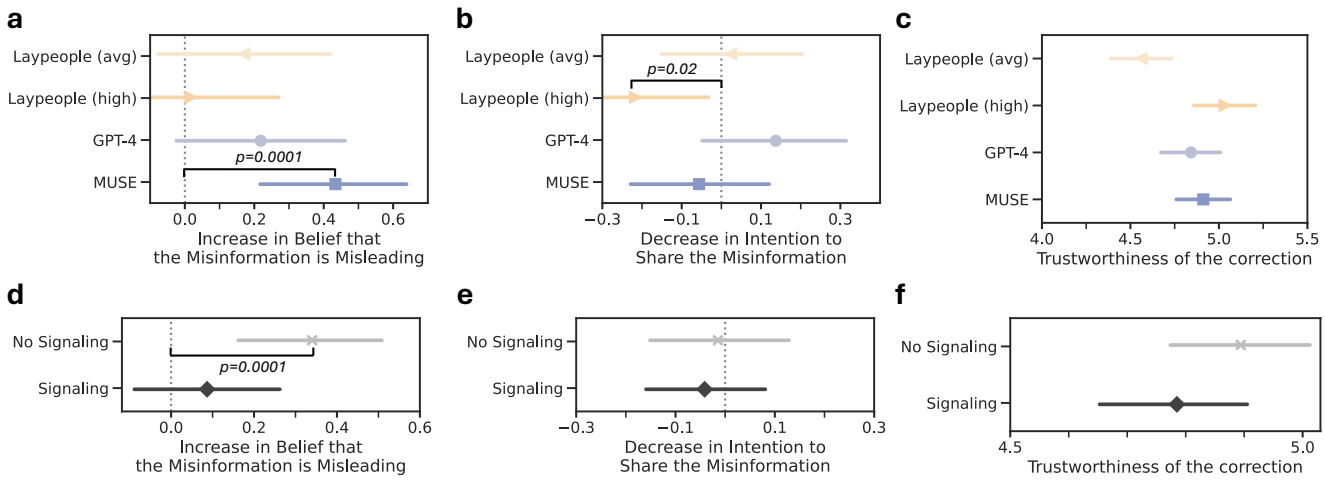


Figure 4. Prolific study with 988 participants to assess the real-world impact of MUSE and how people perceive and react to AI-generated responses (two-sided Student's t-test; 16 experiments). **a**: Only MUSE-generated corrections significantly enhanced the correct belief that the misinformation is misleading ($p = 0.0001$). **b**: None of the corrections made by laypeople, GPT-4, and MUSE significantly impacted participants' intention to share the misinformation, and participants' intentions were generally low (around 2/7). **c**: MUSE, GPT-4, and high-helpfulness responses from laypeople all received significantly higher trustworthiness ratings than average-helpfulness responses from laypeople. (d-f) While disclosing the source of the correction (human or AI) had no significant effect on trustworthiness or intention to share, not disclosing the source significantly enhanced the correct belief that the misinformation is misleading ($p = 0.0001$).

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.

Supplementary Material

Supplementary material is available for this paper.

Author Contributions

X.Z., A.X.Z., and T.A. conceptualized the work and developed the overarching research goals and aims. X.Z. led the model development. A.S. led the assessment of end-user perceptions of corrections, and X.Z. led the remainder of the evaluation. X.Z. led the manuscript writing, including visualization, and all authors revised the paper. A.X.Z. and T.A. supervised the study.

Data Availability

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

Supplementary Materials

List of supplementary materials:
Table S1
Figures S1–S27

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

Category	Definition	# of sources
Factuality:		
- 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
		Total: 5,123
Bias:		
- 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

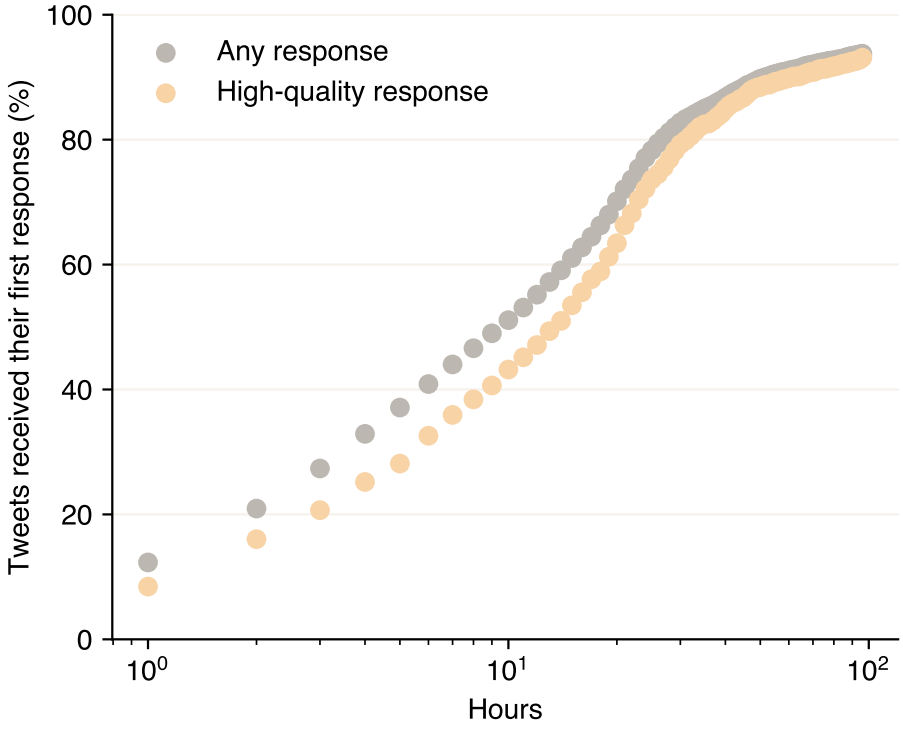


Figure S1. Distribution of potential misinformation in X Community Notes (as of February 2023) that received its first response (gray) or first high-quality response (orange) within a certain amount of time.

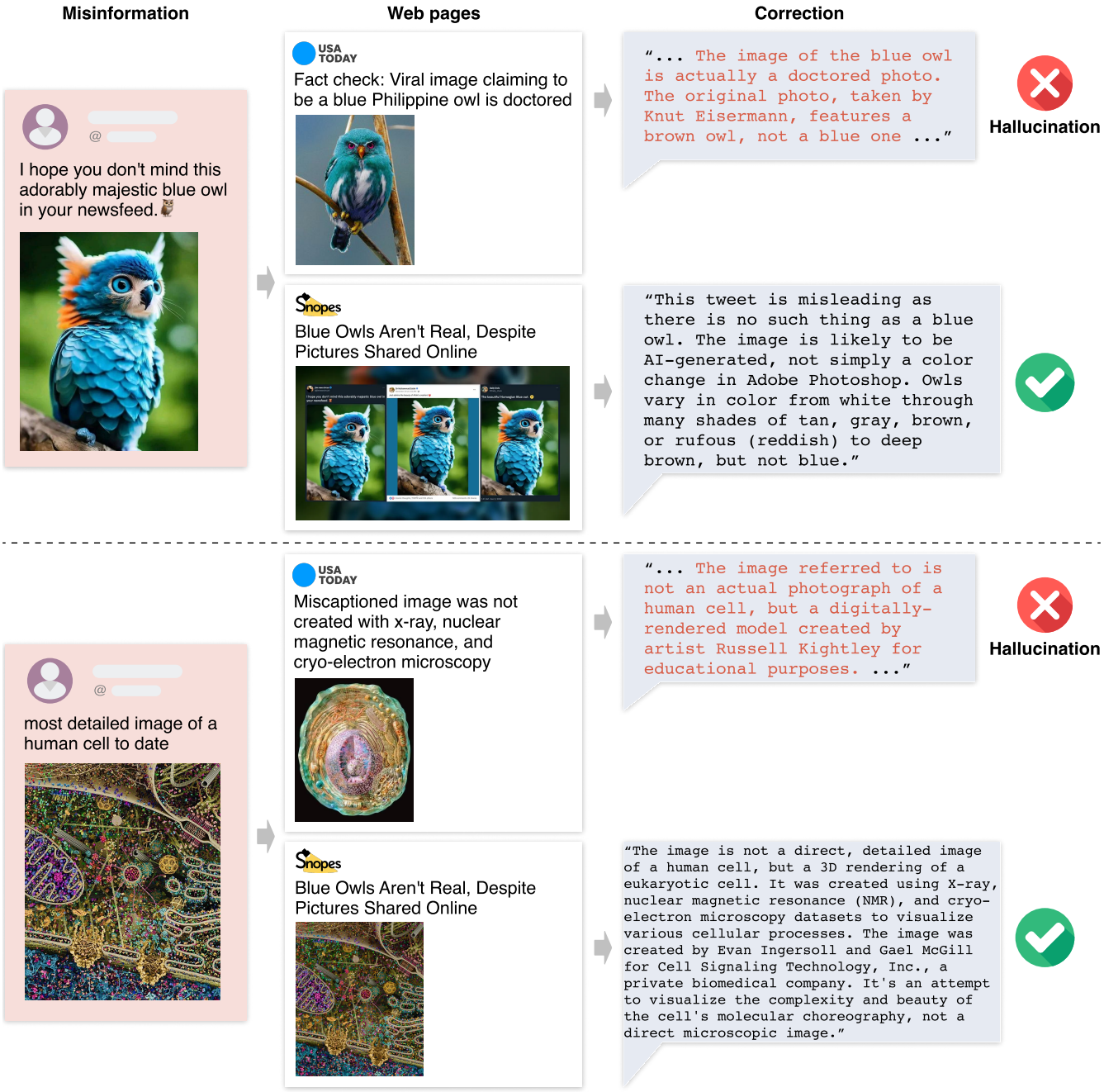


Figure S2. Examples that show how retrieved web pages with relatively low relevance to potential misinformation can promote LLM (in this case, GPT-4) hallucinations when generating responses.

Misinformation	Image captioning	Informative image captioning
	a weather map on a tv screen	A weather map displayed on a TV screen showing wind chill temperatures in various locations such as Waukegan, Woodstock, Rockford, DeKalb, O'Hare, Lakefront, Aurora, Midway, Gary, Joliet, Chicago Heights, Morris, Merrillville, and Kankakee. The map indicates extreme cold temperatures with conditions of blowing snow.
	elon musk's tweet on social media	A screenshot of a post by Elon Musk on Twitter, "Twitter has had a massive drop in revenue, due to activist groups pressuring advertisers, even though nothing has changed with content moderation and we did everything we could to appease the activists. Extremely messed up! They're trying to destroy free speech in America." The post was made at 9:28 AM on 11/4/22.
	a young person sitting on a chair	A photo of a young Aaron Swartz sitting on a chair.

Figure S3. Examples of informative image captions, which augment image captions with names of visually represented celebrities and embedded text (see Experiments for implementation details).



Figure S4. An example of a high-helpfulness response from Community Notes displayed on the corresponding tweet and visible to the public.

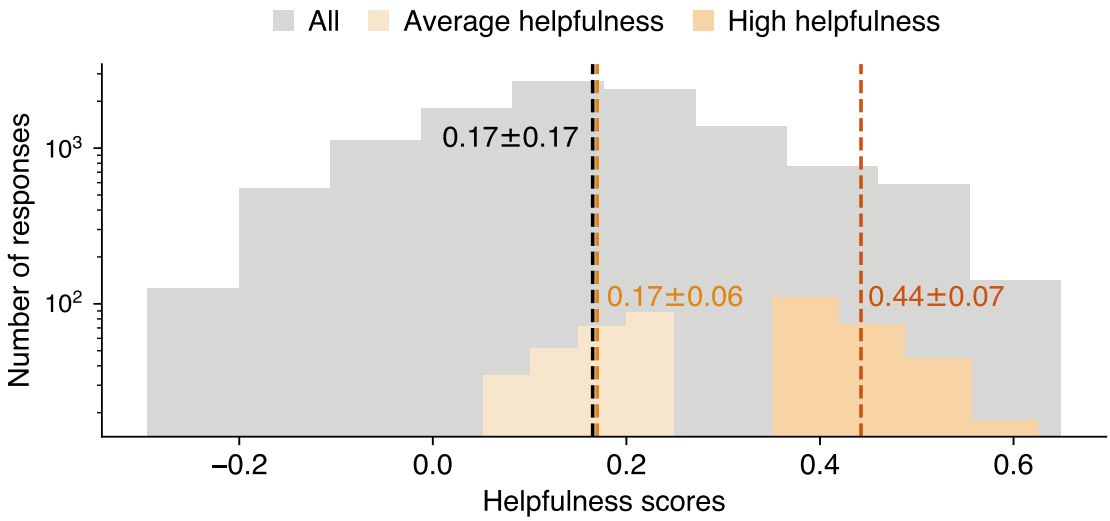


Figure S5. Distribution of helpfulness scores of laypeople's responses in X Community Notes. All: All laypeople's responses in Community Notes. Average helpfulness: Laypeople's responses in Community Notes identified with average helpfulness and used in our study. High helpfulness: Laypeople's responses 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 S6. Examples of MUSE-generated responses to accurate, partially accurate, and factually correct but misleading content on social media.

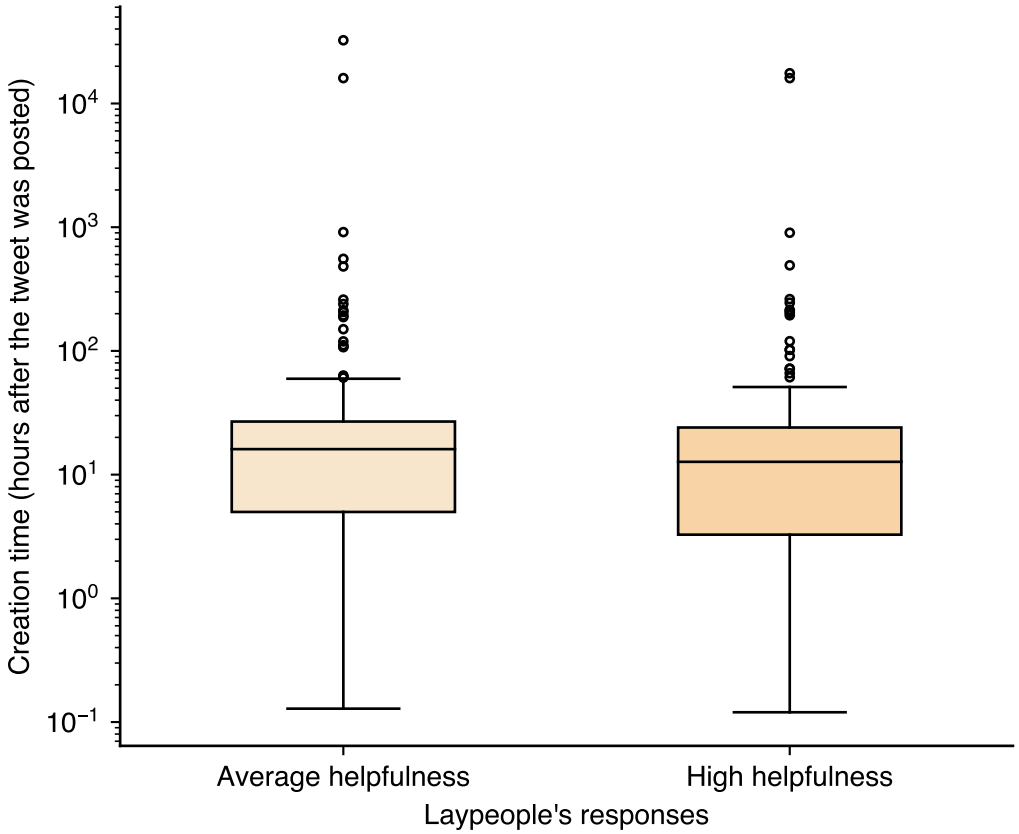


Figure S7. Distribution of creation times of laypeople's responses in X Community Notes used in our study. Median of the creation time of laypeople's average-helpfulness responses: 16 hours after the tweet was posted. Median of the creation time of laypeople's high-helpfulness responses: 13 hours after the tweet was posted.

```
Given a tweet, you are required to generate 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".
```

Figure S8. LLM prompt for query generation. The post information concatenated its textual content, informative descriptions of images (for posts with images), time, and the poster's name.

```
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.'
```

Figure S9. LLM prompt for evidence extraction. The article information included the article's content and published date. The post information concatenated its textual content, informative image captions (for posts with images), time, the poster's name, the poster's screen name, and the poster's description.

```
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.'
```

Figure S10. LLM prompt for response generation. The facts listed every piece of extracted evidence with its source link and published date. Evidence was ranked by publishers' priority and, within each priority, by relevance to the post (both from high to low), which shows to increase GPT-4's accuracy[53].

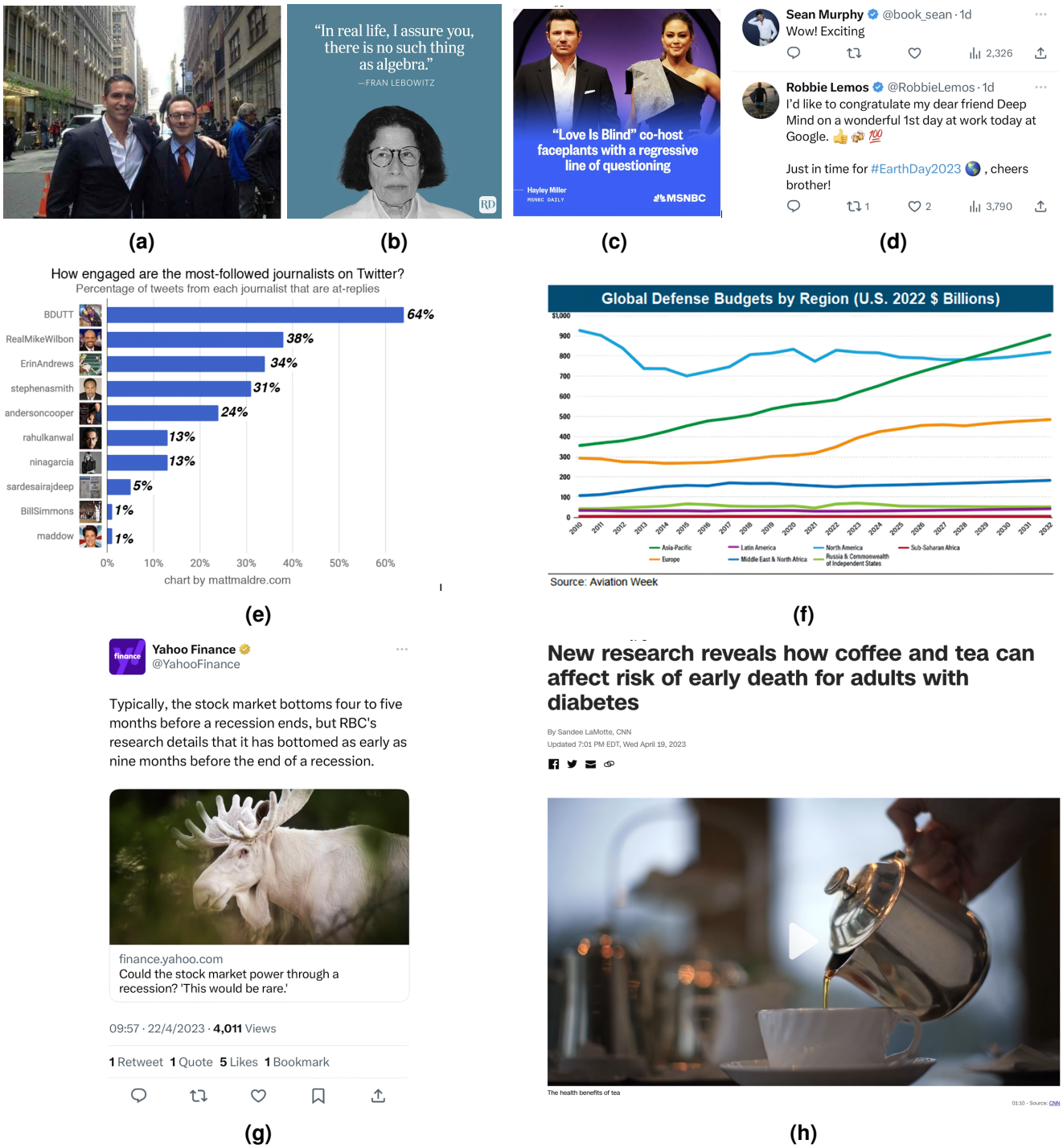


Figure S11. Images as examples used for informative image captioning with in-context learning. These example images do not appear in the dataset we used.

```
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.
short caption: {a woman with glasses and a quote that says, in real life, i assure you there is no such thing as algebra}
name of each person: {Fran Lebowitz}
raw text: {"In real life, I assure you, there is no such thing as algebra."}
image description: {A quote from Fran Lebowitz, "In real life, I assure you, there is no such thing as algebra."}
short caption: {two men in suits}
name of each person: {Jim Caviezel, Michael Emerson}
raw text: {}
image description: {A photo of Jim Caviezel and Michael Emerson in suits}
short caption: {robbie leems on twitter}
name of each person: {}
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}
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.}
short caption: {a moose}
name of each person: {}
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}
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.}
short caption: {a person pouring tea into a cup}
name of each person: {}
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}
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.}
short caption: {two people standing next to each other with the words love is blind}
name of each person: {Nick Lachey}
raw text: {\"Love Is Blind\" co-host faceplants with a regressive line of questioning Hayley Miller MSNBC DAILY MSNBC}
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.}
short caption: {a bar graph that shows how engaged are the most followed journalists on twitter}
name of each person: {Rahul Kanwal}
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% sardesairajdeep 1% BillSimmons maddow 1% 0% 10% 20% 30% 40% 50% 60% chart by mattmaldre.com}
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.}
short caption: {a graph showing the global defense budget by region}
name of each person: {}
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}
image description: {A graph showing the global defense budget by region. It is from Aviation Week.}
short caption: {[IMAGE_CAPTION]}
name of each person: {[CELEBRITIES]}
raw text: {[OCR]}
image description:
```

Figure S12. LLM prompt for informative image captioning. As examples, we selected eight diverse images and manually generated their informative descriptions.

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 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 S13. 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!

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 S14. 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 S15. 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 S16. 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 S17. 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 S18. 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 S19. Annotation instructions (page 7/7).

<< Back to Assigned Tasks Instructions Example Task

Task #: 1

Jeff Zients (@WHCOS)

See User Description

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

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



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

* indicates a required field.

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

R1 — **Time:** Dec. 18, 2022, 2:24 p.m.

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

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

R2 — **Time:** Dec. 18, 2022, 2:24 p.m.

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

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

R3 — **Time:** Dec. 18, 2022, 2:24 p.m.

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

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

Figure S20. Annotation task page.

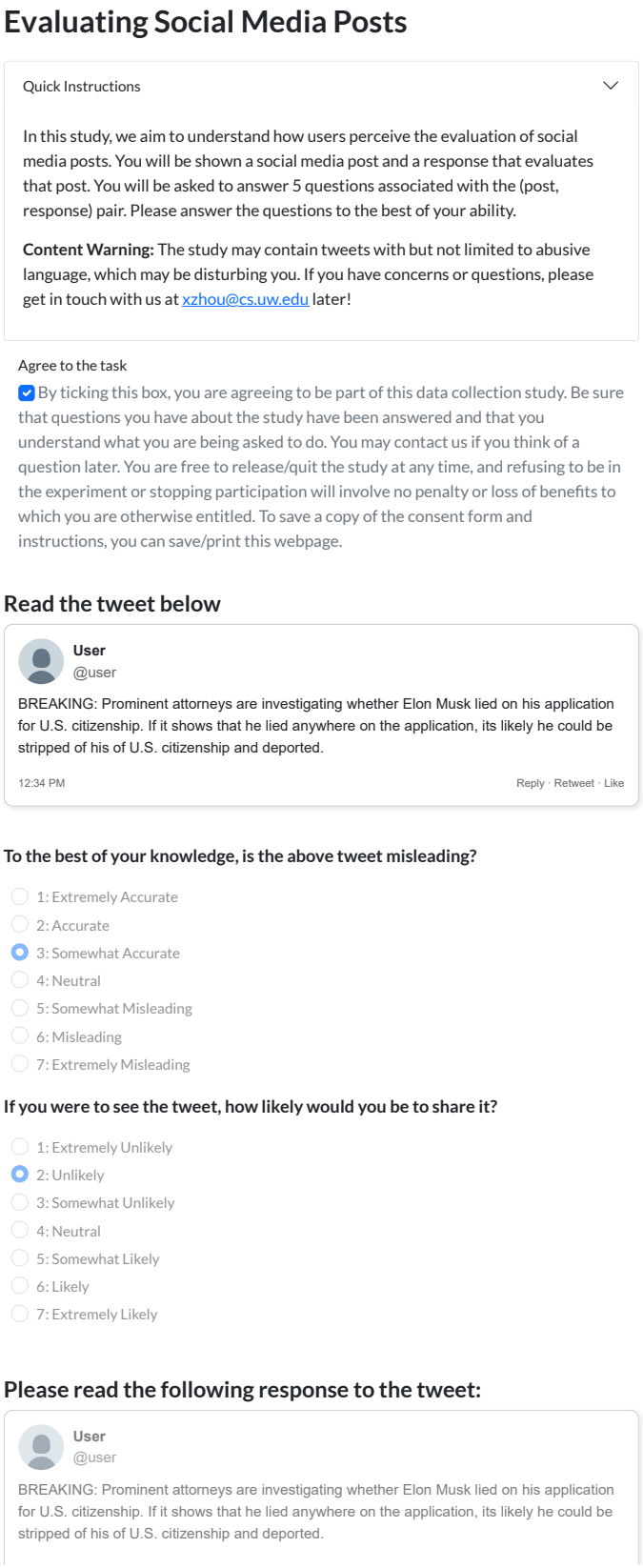
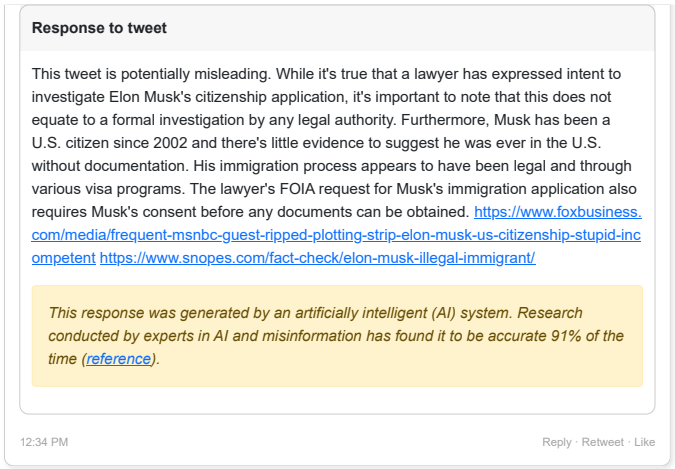


Figure S21. Study interface to measure end user perceptions of corrections (page 1/2, continued on the next page).



How trustworthy do you think the above response is?

- ○ 1: Extremely Untrustworthy
- ○ 2: Untrustworthy
- ○ 3: Somewhat Untrustworthy
- ○ 4: Neutral
- ○ 5: Somewhat Trustworthy
- ◉ 6: Trustworthy
- ○ 7: Extremely Trustworthy

After reading the response, is the tweet misleading?

- ○ 1: Extremely Accurate
- ○ 2: Accurate
- ○ 3: Somewhat Accurate
- ○ 4: Neutral
- ○ 5: Somewhat Misleading
- ◉ 6: Misleading
- ○ 7: Extremely Misleading

After reading the response, if you were to see the tweet, how likely would you be to share it?

- ◉ 1: Extremely Unlikely
- ○ 2: Unlikely
- ○ 3: Somewhat Unlikely
- ○ 4: Neutral
- ○ 5: Somewhat Likely
- ○ 6: Likely
- ○ 7: Extremely Likely

Submit

Figure S22. Study interface to measure end user perceptions of corrections (page 2/2).

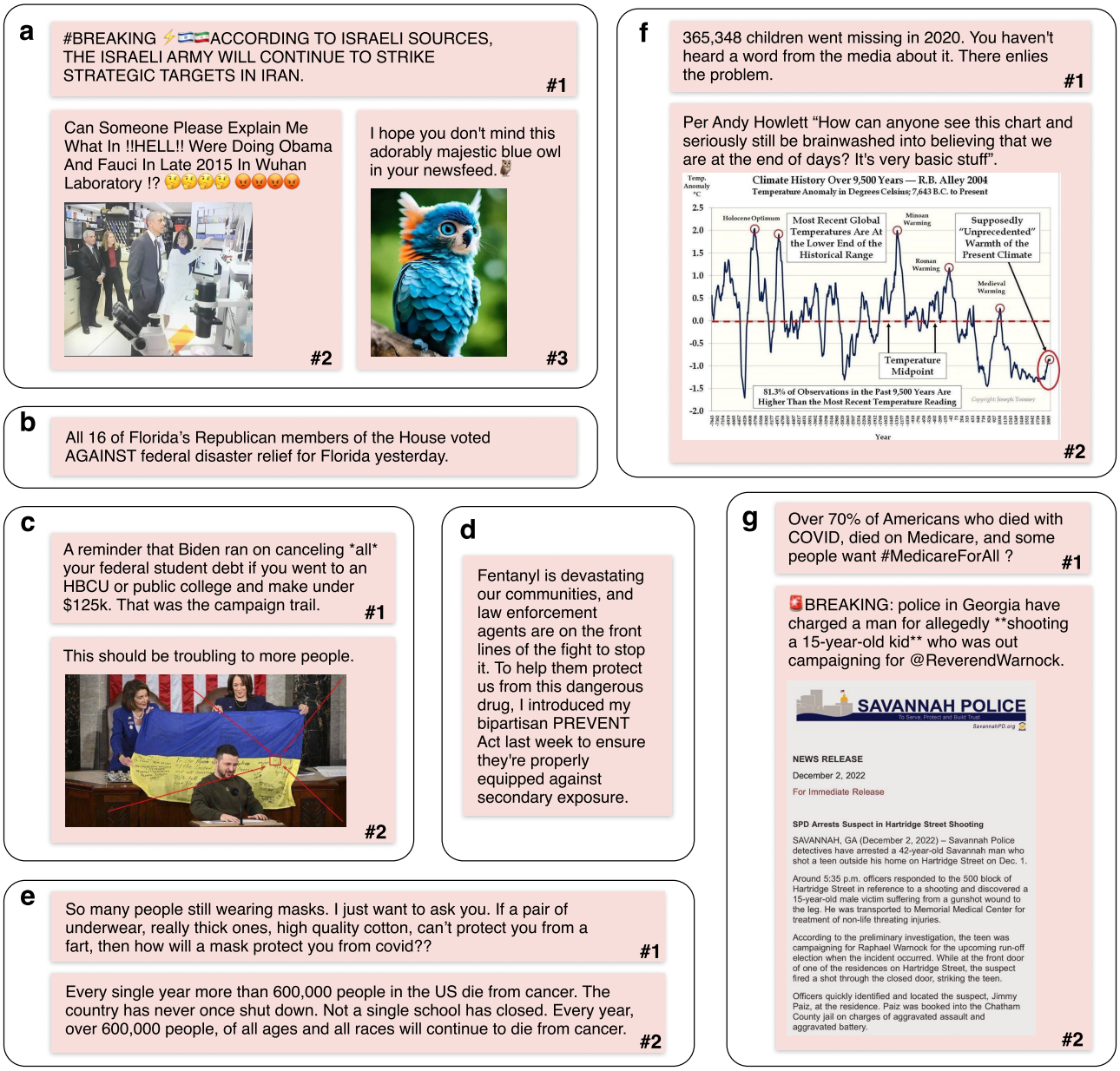
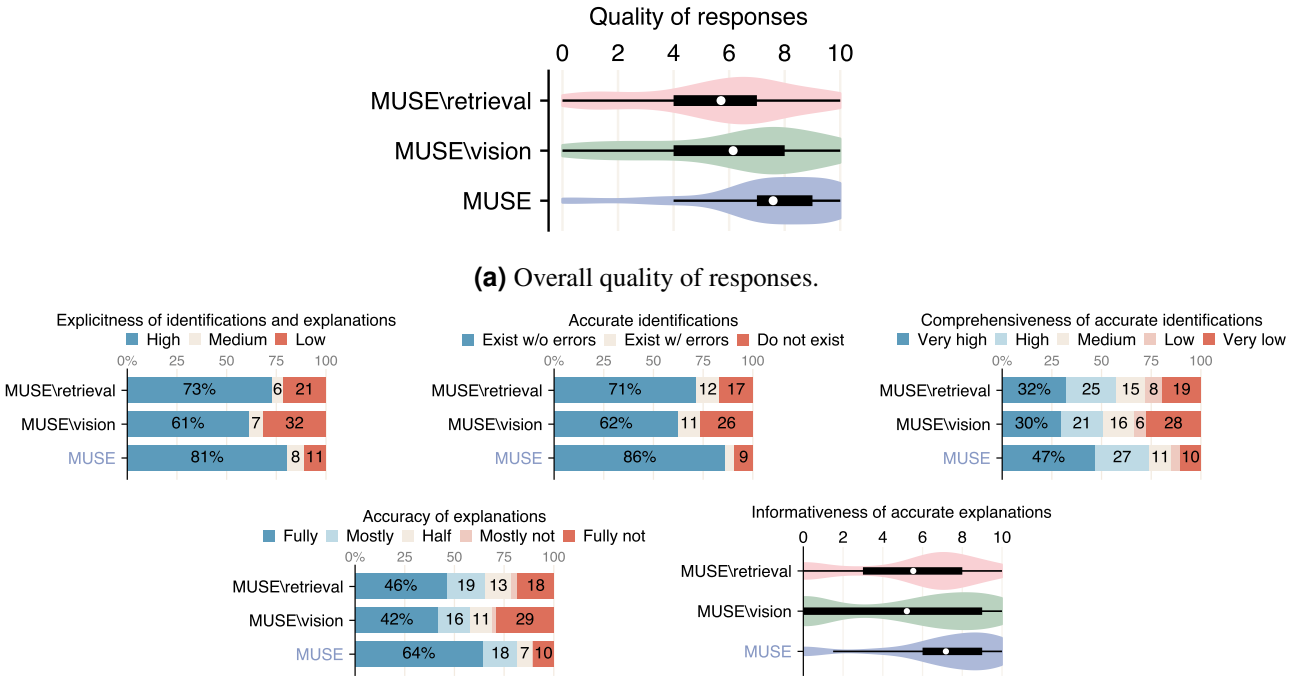
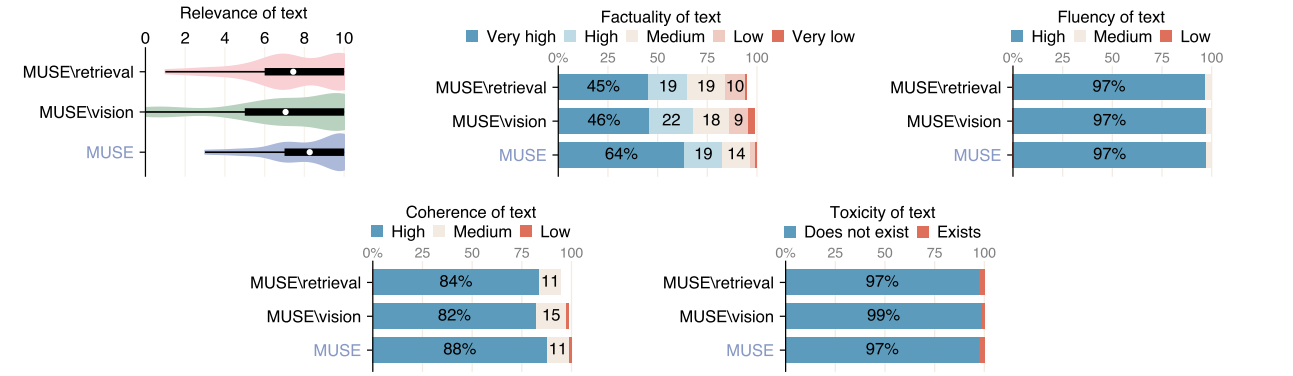


Figure S23. Examples of social media content that uses various tactics to make it or part of it false or misleading. These tactics include **a**: *Fabricating* a news event (#1), a story with an authentic image (#2, also related to conspiracy theory), and an image with AI (#3). **b**: *Lacking context* regarding the motivation behind behavior. Here, it is the consideration that the bill contained no funding for Florida, without which could mislead readers. **c**: *Misinterpreting or misrepresenting* a plan, where Biden did not promise to cancel all the debt (#1), or symbol, which is not Nazi SS runes but a shorthand for the 46th Separate Airmobile Brigade (#2). **d**: *Using loaded language*. **e**: *Improperly analogizing or equating* masks blocking respiratory droplets and underwear blocking gas molecules (#1), and COVID-19 and cancer (#2). **f**: *Presenting false, partial, or biased data.* 365348 includes multiple reports for the same child and runaways who may not be considered missing because of their guardians know their whereabouts (#1). The data ends in 1885 and is from a specific high elevation site in Greenland, not global temperatures (#2). **g**: *Implying false or oversimplified causation*. Many COVID-19 deaths occurred among Medicare beneficiaries because Medicare primarily serves the groups who are at greater risk of adverse outcomes from COVID-19 (#1). The boy was shot when—but not because—he was campaigning for Raphael Warnock (#2). Note that one content may apply more than one tactic intentionally or accidentally.

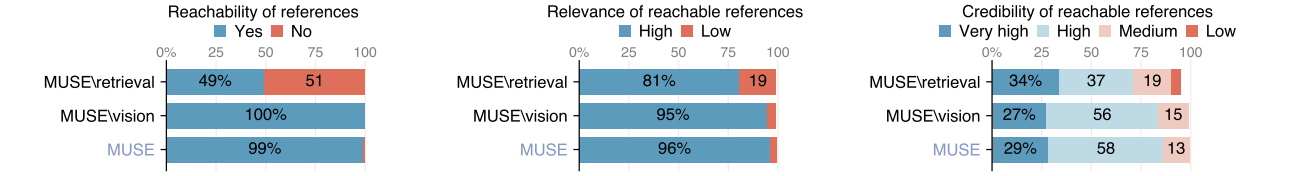


(a) Overall quality of responses.

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



(c) Quality of responses in generated text.



(d) Quality of responses in references.

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

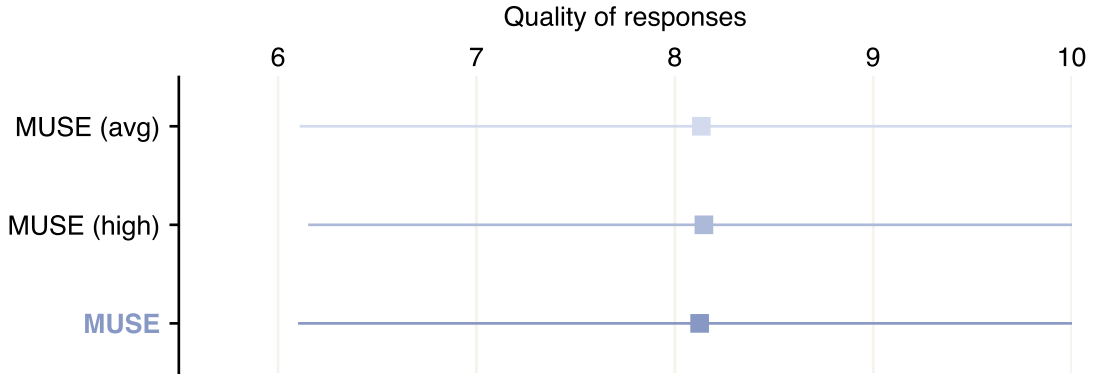
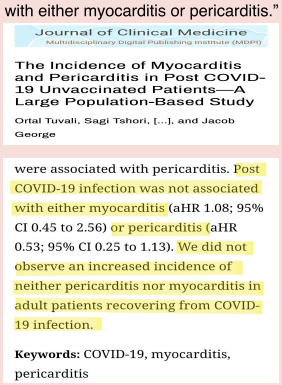


Figure S25. Impact of starting times of responding to tweets on MUSE's performance. The simulated starting time for MUSE (avg): Thirty minutes before the corresponding laypeople's average-helpfulness responses was created (median: 16 hours after the corresponding tweet was posted; Supplementary Fig. S7). 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. S7). The simulated starting time for MUSE: The post time of the corresponding tweet (i.e., 0 hours after the corresponding tweet was posted).

Social media post	MUSE-generated response
FBI Director Christopher Wray paid $2,500 a night for prostitutes in Davos.	`There's no credible source or evidence provided to support the claim about FBI Director Christopher Wray. It's important to verify such serious allegations from reliable sources before sharing.`
Mitch McConnell said he would resign if he doesn't become the Senate Majority Leader in November. I think Democrats should run this message in every state. Voters despise Moscow Mitch all across the country. If you want to end McConnell's political career, vote for the Democrats!	`There's no public record of Mitch McConnell stating he would resign if he doesn't become the Senate Majority Leader. It's important to verify such claims from reliable sources before sharing. Misinformation can lead to uninformed decisions.`
1200 dead of COVID yesterday in Florida. Not the total for a week or a month, but ONE SINGLE DAY.	`According to the Florida Department of Health, the highest single-day death toll was 276 on August 11, 2021. The number mentioned` **`seems`** `to be a cumulative total over several days, not a single day count. Always refer to official health department data for accurate information.`
The future republicans are working towards 	`This tweet is potentially misleading as it` **`seems`** `to be based on hypothetical scenarios. The headlines in the attached image are from a future date and appear to be speculative, not factual. It's important to base political discussions on current, verifiable information to avoid misinformation.`
⚡Russian Chief of the General Staff, Valery Gerasimov, reportedly fired today. Gerasimov was very highly regarded, the most important military leader of the past generation, & the architect of today's Russian Armed Forces. He is served as the head of the military since 2012. 	`This tweet is potentially misleading without further context or confirmation. As of now, there are no official reports or statements from the Russian government confirming that Valery Gerasimov has been fired. It's crucial to verify such significant information from reliable sources before sharing.`

Figure S26. Examples of MUSE-generated responses to social media posts that lack evidence to verify.

Social media post	Response by GPT4-based MUSE	Response by Llama3-based MUSE
So Covid vaccines may cause myocarditis, but Covid itself causes myocarditis too, right? WRONG! A large-scale Israeli study of 196,992 unvaccinated adults after Covid infection showed that “Covid infection was not associated with either myocarditis or pericarditis.”  Journal of Clinical Medicine Multidisciplinary Digital Publishing Institute (MDPI) **The Incidence of Myocarditis and Pericarditis in Post COVID-19 Unvaccinated Patients—A Large Population-Based Study** Ortal Tuvali, Sagi Tshori, [...], and Jacob George were associated with pericarditis. Post COVID-19 infection was not associated with either myocarditis (aHR 1.08; 95% CI 0.45 to 2.56) or pericarditis (aHR 0.53; 95% CI 0.25 to 1.13). We did not observe an increased incidence of neither pericarditis nor myocarditis in adult patients recovering from COVID-19 infection. **Keywords:** COVID-19, myocarditis, pericarditis	While the Israeli study cited does indeed find no increased incidence of myocarditis or pericarditis in post-COVID-19 patients, other studies have found a higher risk of myocarditis in COVID-19 patients compared to those without COVID-19. For instance, a CDC report found that patients with COVID-19 had, on average, 15.7 times the risk for myocarditis compared to patients without COVID-19. However, the absolute risk difference was small due to the low risk of myocarditis in both groups. It's important to consider all available research and understand that results can vary based on different study designs and populations. https://pubmed.ncbi.nlm.nih.gov/35456309/ https://www.cdc.gov/mmwr/volumes/70/wr/mm7035e5.htm	While the study cited in the tweet is correct that post-COVID-19 infection was not associated with either myocarditis or pericarditis in the specific population studied, this finding contradicts other recent studies that found a significant association between COVID-19 infection and myocarditis. For example, a CDC study published in 2021 found that patients with COVID-19 had nearly 16 times the risk for myocarditis compared with patients who did not have COVID-19, and risk varied by sex and age. https://www.cdc.gov/mmwr/volumes/70/wr/mm7035e5.htm https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9025013/
Louis DeJoy plans to slash 50,000 USPS jobs. Why in the world does Trump's postmaster general still have his job? And how is this not a massive national scandal?	While it's true that Louis DeJoy has plans to reduce USPS positions, it's not a direct job cut but rather a reduction through voluntary early retirements and expected attrition over the next few years. The aim is to reach a break-even point for USPS, which has been facing financial difficulties. The number of positions to be reduced represents only a quarter of the attrition the agency expects to see over a similar period. DeJoy has also stated that USPS will remain in a hiring posture in the coming years. **As for why DeJoy still holds his position, the Postmaster General is not a political appointee and cannot be removed by a president. The power to remove the Postmaster General lies with the Postal Service's governing board, whose members were all appointed by President Trump.** https://www.govexec.com/workforce/2022/07/usps-plans-slash-50000-positions-coming-years-reach-break-even-point/375096/ https://abcnews.go.com/Politics/controversial-postal-chief-poised-remain-biden-administration/story?id=74469137 https://www.reuters.com/business/us-postal-service-will-offer-early-retirements-consolidate-postal-districts-2021-03-03/	The claim that Louis DeJoy plans to slash 50,000 USPS jobs is misleading. According to DeJoy's statement at an American Enterprise Institute event, the net reductions of 50,000 positions represent only one-quarter of the attrition the agency expects to see over a similar period, and USPS will remain in a hiring posture in the coming years. Additionally, DeJoy stated that the agency will be able to manage any headcount reduction through normal attrition, which results from employees retiring or leaving the agency. https://www.govexec.com/workforce/2022/07/usps-plans-slash-50000-positions-coming-years-reach-break-even-point/375096/ https://federalnewsnetwork.com/workforce/2022/07/usps-may-need-50k-fewer-employees-under-break-even-plan-dejoy-says/

Social media post	Response by GPT-4	Response by Llama-3
While we’re distracted by pronouns, UFOs, & Woke Jesus, it’s important to remember that over 100,000 gallons of burning vinyl chloride is destroying the air, water, soil, & living creatures of Palestine, Ohio. This is one of the biggest ecological disasters in American history. 	There's no record of an ecological disaster involving 100,000 gallons of burning vinyl chloride in Palestine, Ohio. Vinyl chloride is indeed a hazardous substance, but there's no evidence of such an incident in the mentioned location. Please ensure to verify information before sharing.	The location of the ecological disaster is actually East Palestine, Ohio, not Palestine, Ohio. Palestine is not a town in Ohio. The rest of the tweet regarding the scale of the disaster appears to be accurate, highlighting the importance of attention to this environmental crisis.

Figure S27. Comparison between responses generated by GPT-4, Llama-3, and their based MUSE. We observed that GPT4-based MUSE and Llama3-based MUSE frequently generate similar responses to social media content (see the top example). GPT4-based MUSE occasionally generates more comprehensive corrections than Llama3-based MUSE (see the middle example). Note that the knowledge cutoff for Llama3 is December 2023, whereas that of GPT-4 is September 2021. In other words, Llama-3’s training data include more recent events and knowledge than GPT-4’s (see the bottom example, where the post was posted in February 2023, and Llama-3 more accurately identifies its falsehood than GPT-4).