

Beyond the Crowd: LLM-Augmented Community Notes for Governing Health Misinformation

Jiaying Wu^{1*}, Zihang Fu^{1*}, Haonan Wang¹, Fanxiao Li²,
Jiafeng Guo^{3,4}, Preslav Nakov⁵, Min-Yen Kan¹

¹National University of Singapore, ²Yunnan University,

³State Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences, ⁴University of Chinese Academy of Sciences,

⁵Mohamed bin Zayed University of Artificial Intelligence

jiayingwu@u.nus.edu, zihangfu@u.nus.edu, kanmy@comp.nus.edu.sg

Abstract

Community Notes, the crowd-sourced misinformation governance system on X (formerly Twitter), allows users to flag misleading posts, attach contextual notes, and rate the notes' helpfulness. However, our empirical analysis of 30.8K health-related notes reveals **substantial latency**, with a median delay of 17.6 hours before notes receive a helpfulness status. To improve responsiveness during real-world misinformation surges, we propose CROWDNOTES+, a unified LLM-based framework that augments Community Notes for faster and more reliable health misinformation governance. CROWDNOTES+ integrates two modes: (1) evidence-grounded note **augmentation** and (2) utility-guided note **automation**, supported by a hierarchical three-stage evaluation of relevance, correctness, and helpfulness. We instantiate the framework with HEALTHNOTES, a benchmark of 1.2K health notes annotated for helpfulness, and a fine-tuned helpfulness judge. Our analysis first **uncovers a key loophole** in current crowd-sourced governance: voters frequently conflate stylistic fluency with factual accuracy. Addressing this via our hierarchical evaluation, experiments across 15 representative LLMs demonstrate that CROWDNOTES+ significantly outperforms human contributors in **note correctness, helpfulness, and evidence utility**.¹

1 Introduction

Health misinformation on social media has fueled persistent “infodemics” that endanger public trust and threaten individual well-being (Islam et al., 2020; Shahbazi and Bunker, 2024). Often triggered by major real-world events (Shahi et al., 2021; Adebessin et al., 2023), such misinformation spreads at a scale and velocity that consistently exceeds the capacity of expert fact-checkers

*Equal Contribution

¹Code and data are available at: <https://github.com/jiayingwu19/CrowdNotesPlus>.

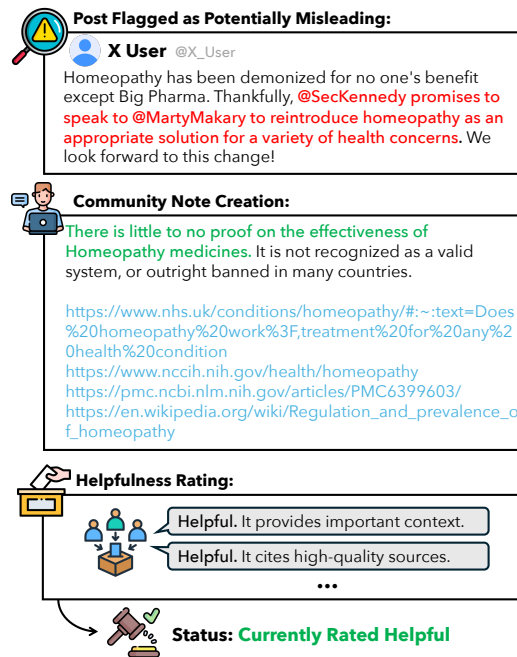


Figure 1: **Overview of Community Notes on X for crowd-sourced misinformation governance.** Users engage in three stages: (1) **flagging** potentially misleading posts, (2) **writing** notes that provide clarification or additional context, and (3) **rating** the notes' helpfulness. Based on accumulated ratings, each note receives one of three statuses: (a) *Needs More Ratings*, (b) *Currently Rated Not Helpful*, or (c) *Currently Rated Helpful*. Only notes from the last category are publicly displayed alongside the original post to inform readers.

and platform-level moderation (Godel et al., 2021; Singer, 2023). In response, **crowd-sourced fact-checking**, which leverages the collective wisdom of online contributors (Allen et al., 2021; Martel et al., 2024; Shahbazi and Bunker, 2024; Pfänder and Altay, 2025), has emerged as a scalable and timely complement to expert-driven approaches.

Community Notes (Wojcik et al., 2022), deployed on X (formerly Twitter), is the most prominent implementation of this paradigm (Figure 1). The system enables users to flag suspicious posts,

write contextual notes, and vote on their helpfulness; only notes with a status of *Currently Rated Helpful* are shown to inform the public. While prior work has demonstrated Community Notes’ potential for improving discourse quality and reducing polarization (Chuai et al., 2024a; Renault et al., 2024; Slaughter et al., 2025), our large-scale analysis of 30.8K health-related notes over four years (§3) reveals two systemic bottlenecks that limit the system’s responsiveness to fast-moving health misinformation: **(1) Delayed note generation.** Extending earlier reports of latency in Community Notes (Renault et al., 2024), we find that the first note appears a median of 10.4 hours after a misleading health post is flagged, and the first helpfulness verdict (i.e., *Helpful/Not Helpful*) arrives another 7.2 hours later—well past the period of peak public attention. **(2) Sparse helpfulness evaluation.** A striking 87.9% of health notes remain indefinitely in the *Needs More Ratings* state. As only *Helpful* notes are surfaced, this bottleneck further delays corrective information from reaching users when it is most needed.

To address these limitations, we introduce CROWDNOTES+, a unified framework that leverages large language models (LLMs) to enhance both the creation and evaluation of Community Notes for more timely and reliable misinformation governance. Given a flagged post containing a potentially misleading claim, CROWDNOTES+ extends the existing crowd-sourced pipeline through two complementary generation modes (Figure 3): **(1) Evidence-Grounded Note Augmentation**, where humans supply evidence (e.g., URLs) and LLMs synthesize it into structured notes, and **(2) Utility-Guided Note Automation**, where LLMs autonomously plan, retrieve, and select high-quality evidence before generating notes. To ensure robust and interpretable assessment, CROWDNOTES+ further incorporates a **hierarchical three-step evaluation pipeline** that progressively verifies **(1) the relevance** of the retrieved evidence, **(2) the correctness** of the evidence presented, and **(3) the overall helpfulness** of the generated note.

We instantiate CROWDNOTES+ in the health domain through the HEALTHNOTES, a benchmark of 1.2K health-related Community Notes with crowd-confirmed *Helpful* and *Not Helpful* verdicts, along with HEALTHJUDGE, a fine-tuned note helpfulness evaluator. Our extensive experiments on fifteen representative LLMs validate the framework’s reliability and practical utility. Specifically, we identify

a fundamental weakness in current crowd-sourced helpfulness assessment (§7.1), where stylistic fluency is often mistaken for factual accuracy, and show that our hierarchical evaluation substantially reduces such false positives. Across both generation modes, LLMs produce notes that are more accurate and contextually balanced than human-written notes (§7.2), and utility-guided automation consistently selects higher-quality evidence than human contributors (§7.3). These results position CROWDNOTES+ as a principled approach for improving the timeliness, factual consistency, and interpretability of crowd-sourced health misinformation governance on social media.

2 Related Work

Crowd-Sourced Fact-Checking. The scale and speed of online misinformation make it unrealistic to rely solely on professional fact-checkers (Godel et al., 2021; Singer, 2023). Crowd-sourced fact-checking (Allen et al., 2021; Martel et al., 2024; Shahbazi and Bunker, 2024; Pfänder and Altay, 2025; Xing et al., 2026), exemplified by Community Notes on X, allows users to collaboratively provide clarifications on potentially misleading content. Prior work shows that such community moderation can reduce misinformation engagement (Chuai et al., 2024b; Slaughter et al., 2025) and promote more balanced discourse (Chuai et al., 2024a; Renault et al., 2024). However, most studies assume that notes already exist and focus on voting dynamics, consensus formation, or downstream impact. The earlier stage of note creation, especially in time-sensitive contexts, remains underexplored. Initial automation attempts (De et al., 2025; Singh et al., 2025) have limited practicality because (De et al., 2025) requires multiple human-written notes for the same post, and (Singh et al., 2025) depends solely on LLM internal knowledge without web access, insufficient for emerging or unseen claims. Our work fills this gap in the health domain, where timeliness is crucial, by introducing a unified framework for systematic LLM-augmented note generation and evaluation.

Automated Governance of Textual Misinformation. Automated approaches aim to identify and counter misinformation at scale. Prior work has developed classifiers for detecting misleading posts and articles, using linguistic features (Potthast et al., 2018; Zhang et al., 2021) and network-based signals (Wu and Hooi, 2023; Wu et al., 2023). While

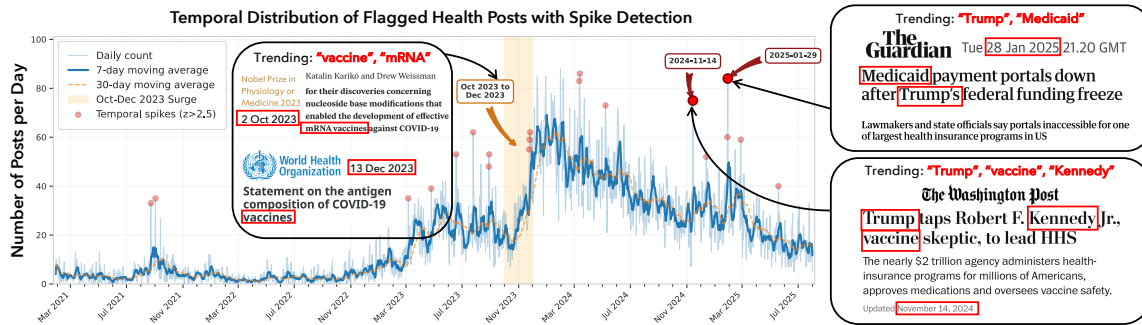


Figure 2: Spikes in flagged health misinformation posts align with major real-world health events (detailed in §3.2), including outbreak alerts, vaccine developments, and policy debates, highlighting the event-driven nature of misinformation activity on social media.

effective for flagging suspicious content, these systems rarely provide explanations that clarify why the content is misleading. Recent studies use LLMs to generate explanatory text (Hu et al., 2024; Wu et al., 2024b) and retrieve evidence from credible sources to justify predictions (Pan et al., 2023; Zhang and Gao, 2023; Zhou et al., 2024). However, these methods typically position the model as an autonomous arbiter, treating explanations merely as justifications. This overlooks the “human-in-the-loop” nature of governance systems like Community Notes. Our work bridges this gap by evaluating LLMs not as replacements, but as assistants that empower contributors with evidence-grounded drafts, preserving the human locus of control.

3 Temporal Dynamics of Health Misinformation and Community Notes

Understanding how health misinformation emerges and how community governance responds is essential for designing timely interventions. Before developing automated support, we analyze the temporal dynamics of health-related Community Notes on X to identify when misinformation surges occur and how promptly the system reacts.

3.1 Data Scope

We collected all publicly available, user-contributed Community Notes² on X up to 4 August 2025, retaining only English entries for consistency. To focus on health-related content, we define seven topical categories: (1) diseases or medical conditions, (2) drugs, vaccines, treatments, and tests, (3) public health guidance or policy, (4) wellness products, diets, and supplements, (5) healthcare professionals or systems, (6) biological

²<https://x.com/i/communitynotes/download-data>

or epidemiological concepts, and (7) health-related conspiracies or hoaxes.

We filter the collected notes using zero-shot prompting with Lingshu-32B (Li et al., 2025), a multimodal LLM with state-of-the-art performance on medical QA. To validate this filter, we cross-check its predictions against closed-source LLMs on a random sample of 1,000 notes, observing high agreement (GPT-4.1 (OpenAI, 2025a): 99.2%, Gemini-2.5-Flash (Google, 2025): 100%, Claude-4-Sonnet (Anthropic, 2025): 96.8%). Given this high reliability, we retain all notes classified as health-related by Lingshu-32B. We then retrieve the associated posts, using GPT-4.1 to keep only those with text-based health claims, while removing unavailable posts or URL-only content.

This process yields 30,791 health-related notes covering 25,484 potentially misleading posts. We base our following analysis of temporal trends and systemic bottlenecks on this data.

3.2 Event-Driven Misinformation Dynamics

We first examine the temporal distribution of the 25K health-related flagged posts to understand **how activity evolves relative to real-world events**. Daily post counts are compared against a 28-day rolling baseline, and a day is marked as a **spike** if its count exceeds the rolling mean by more than 2.5 standard deviations.

To contextualize each spike, we identify trending topics within a three-day window centered on the spike. We compute word frequencies from post text after removing stopwords, identify trending terms, and associate each surge with major health events reported by mainstream news outlets or public health authorities during the same period. Only events that are uniquely prominent within their window are retained to avoid cross-period overlap.

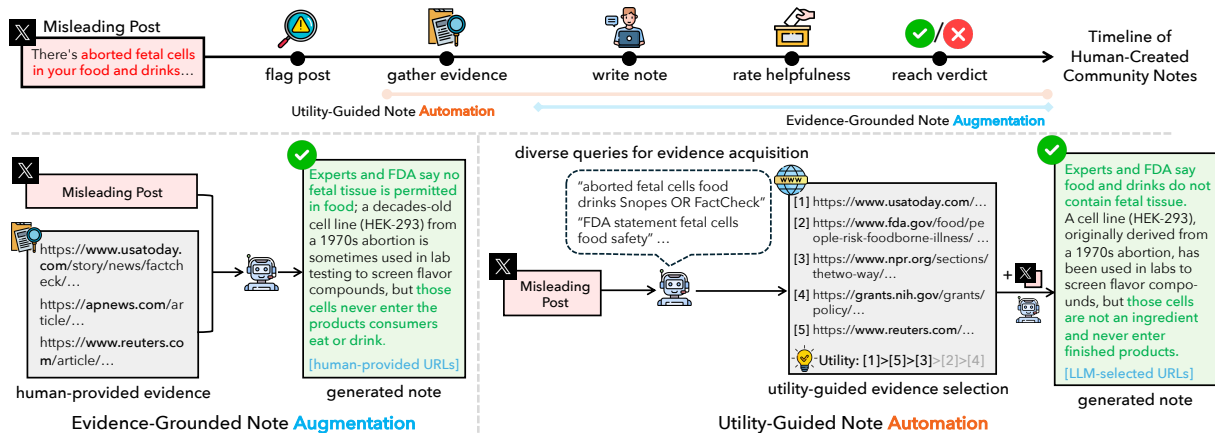


Figure 3: **Overview of the proposed CROWDNOTES+ framework for LLM-augmented Community Notes.** The upper timeline illustrates the human-created Community Notes workflow on X. The lower panels depict two note generation modes in CROWDNOTES+: (1) **evidence-grounded note augmentation**, where LLMs generate notes from human-provided evidence, and (2) **utility-guided note automation**, where LLMs autonomously retrieve and select high-utility evidence from the Web to generate notes.

Pct.	Post Published → First Note	First Note → First Verdict
25%	3.4	3.6
50%	10.4	7.2
75%	23.0	18.4
90%	49.1	76.4

Table 1: **Delays (hours) in health Community Notes**, with a median of 17.6 hours before the first note attains any helpfulness verdict (i.e., *Helpful* vs. *Not Helpful*).

As illustrated by the spikes on 14 November 2024 and 29 January 2025, and the sustained rise from October to December 2023 (Figure 2), *misinformation activity aligns closely with major health developments*, including outbreak announcements, vaccine policy updates, and high-profile public health debates. These patterns show that health misinformation is strongly event-driven and emerges rapidly in response to external developments, motivating our next analysis on how quickly Community Notes respond once such posts appear.

3.3 Delays in Note Creation and Visibility

Building on this analysis, we examine the 30K associated health-related Community Notes to assess **how quickly corrective information becomes visible**. Although Community Notes are intended to support timely, crowd-sourced fact-checking, our temporal analysis shows substantial delays. As reported in Table 1, the median time from a misleading post to the creation of the first note is 10.4 hours. The subsequent voting phase adds another 7.2 hours before the note receives a helpfulness verdict (*Helpful* or *Not Helpful*). Furthermore, 87.9

percent of notes never gather enough votes to exit *Needs More Ratings*, which prevents them from attaining any public-facing status.

Since only notes achieving *Helpful* status are surfaced to readers, *these delays significantly restrict the availability of corrective information at moments when misinformation is spreading most rapidly*. Improving responsiveness therefore requires accelerating both note creation and note evaluation while preserving factual rigor. This motivates our proposed framework, CROWDNOTES+, which leverages LLMs to enhance the timeliness and reliability of Community Notes.

4 CROWDNOTES+: Framework for LLM-Augmented Community Notes

Our analysis shows that although health misinformation closely follows real-world events, the Community Notes workflow often lags behind due to slow note creation and delayed voting. To address these, we propose CROWDNOTES+, a unified framework that uses LLMs to accelerate note creation and evaluation. CROWDNOTES+ supports two complementary modes (Figure 3): (1) evidence-grounded note augmentation and (2) utility-guided note automation, together with a hierarchical evaluation pipeline that assesses relevance, correctness, and helpfulness.

4.1 Evidence-Grounded Note Augmentation

We first examine *whether LLMs can assist contributors in the standard Community Notes setting where reliable evidence is manually provided*. In

this workflow, a user flags a potentially misleading post p and supplies a set of sources \mathcal{E}_h , where each $e \in \mathcal{E}_h$ is a URL linking to external content.

Each evidence piece e is processed through a RETRIEVE step that segments its textual content into passages. Using the post p as a query, a MATCH step selects the most relevant passage from each source, producing a set of evidence chunks \mathcal{C}_h . The model then executes a GENERATE step, conditioning on both p and \mathcal{C}_h to synthesize a concise, informativ note n_h . The evidence URLs \mathcal{E}_h are attached after n_h for transparency.

Figure 10 presents a concrete example of this mode. It preserves the factual grounding of human-curated sources while automating the synthesis of concise, well-structured notes, reducing the time and effort required for human-written explanations.

4.2 Utility-Guided Note Automation

We next examine *whether note creation can be fully automated once a post p is flagged as potentially misleading*, simulating a practical deployment scenario. Unlike the augmentation mode (§4.1), this mode requires the model to retrieve, select, and synthesize evidence without human guidance.

Motivated by findings that diverse query formulations yield complementary retrieval results (Santos et al., 2015; Wu et al., 2024a), the model generates a set of semantically diverse search queries \mathcal{Q} from p . Each query retrieves top-ranked documents through a SEARCH step, and all retrieved items are merged and de-duplicated into a candidate pool $\mathcal{P} = \text{dedup} \left(\bigcup_{q \in \mathcal{Q}} \text{TopK}(q) \right)$.

To select the most informative evidence, we introduce an LLM-based **utility judgment** module inspired by recent advances in evidence ranking (Zhang et al., 2024). Given a fixed quota τ , the model performs τ iterative selections, each time identifying and removing the evidence snippet (title and one-sentence summary) with the highest estimated utility. The resulting items form the machine-selected evidence set \mathcal{E}_m , whose corresponding URLs are appended to the generated note to ensure transparency and traceability. We then apply the same RETRIEVE and MATCH steps (§4.1) to obtain evidence chunks \mathcal{C}_m and generate an automated note n_m conditioned on p and \mathcal{C}_m .

Figure 11 illustrates the full pipeline and evidence selection behavior. This end-to-end mode enables fully automated note generation guided by evidence utility, reducing reliance on human effort while maintaining factual grounding.

4.3 Hierarchical Helpfulness Evaluation

To ensure robust and interpretable assessment of the generated notes, CROWDNOTES+ employs a three-stage evaluation pipeline that sequentially verifies (1) relevance, (2) correctness, and (3) helpfulness.

Relevance evaluates *whether the retrieved evidence offers meaningful factual context, clarification, or supporting information* that helps readers better assess the claim made in the post. It forms the foundation of retrieval-augmented generation (Saad-Falcon et al., 2024; Yu et al., 2025), ensuring that notes are grounded in appropriate information.

Correctness evaluates *whether the note faithfully represents the content of the cited sources*, without factual errors, exaggeration, or selective framing. Even when evidence is relevant, its interpretation can still be distorted, a common issue in scientific and medical communication (Glockner et al., 2024; Wuehrl et al., 2024). This step ensures that the note’s claims align with the provided sources rather than relying on misinterpretation.

Helpfulness evaluates *whether the note assists readers in understanding or critically evaluating the flagged post*, following the official Community Notes criteria.³

Operationalizing the Hierarchy. We implement these criteria as sequential binary gates using LLM-based judges (implementation details in §5 and Appendix D). A note is evaluated for correctness only if it is deemed relevant, and for helpfulness only if it is correct. Formally, let R , C , and H denote binary indicators of relevance, correctness, and helpfulness. The joint probability of a note satisfying all criteria decomposes as:

$$\begin{aligned} P(R=1, C=1, H=1) &= P(H=1 \mid C=1, R=1) \\ &\quad \times P(C=1 \mid R=1) \\ &\quad \times P(R=1). \end{aligned} \quad (1)$$

This formulation enforces a strict dependency: **a note is deemed helpful only if strictly grounded in relevance and correctness**. By decomposing helpfulness into these conditional components, our design prevents the common failure mode where models rely on surface-level fluency rather than factual reasoning (Wan et al., 2025), yielding a transparent and fine-grained assessment.

³<https://communitynotes.x.com/guide/en/under-the-hood/download-data>

5 The HEALTHNOTES Benchmark

We introduce HEALTHNOTES, the first benchmark for studying LLM-augmented Community Notes in the health domain. HEALTHNOTES combines a curated dataset with a customized evaluation judge, providing a reproducible foundation for analyzing LLM augmentation and automation methods in this high-stakes setting.

Data. To capture both successful and unsuccessful corrections, we include both *Helpful* and *Not Helpful* health notes as labeled by human contributors. From the health-related Community Notes collected in §3.1, we identify 3,713 notes with crowd-confirmed helpfulness labels (*Helpful*: 2,971; *Not Helpful*: 742). Among these, 634 *Not Helpful* notes retain valid evidence URLs. To create a balanced benchmark, we sample an equal number of *Helpful* notes, resulting in 1,268 post–note pairs.

Each data instance contains a flagged post, a corresponding note text, and verified evidence URLs. Table 6 summarizes dataset statistics such as post and evidence counts. Figure 9 shows the distribution over the seven health categories defined in §3.1, confirming that HEALTHNOTES covers diverse health-related topics (See Appendix B).

Evaluation Pipeline. Our evaluation follows the hierarchical scheme in §4.3. For *relevance* and *correctness*, we use an LLM-as-a-Judge setup with GPT-4.1 (OpenAI, 2025a). For the final *helpfulness* stage, we introduce HEALTHJUDGE, a fine-tuned Lingshu-7B model (Li et al., 2025) designed for domain-specific note helpfulness assessment. We provide training details, human validation of judge reliability, and comparative performance results on helpfulness judgment in Appendix D.4.

6 Experiments

We benchmark 15 representative LLMs against a **Human Baseline** of original health community notes. The models span four categories: (1) **closed-source large reasoning models** (LRMs) such as o3 (OpenAI, 2025b), (2) **closed-source LLMs** such as GPT-4.1 (OpenAI, 2025a), (3) **open-source LLMs and LRMs** such as Qwen3 (Yang et al., 2025), and (4) **domain-specific medical LLMs** such as MedGemma (Sellergren et al., 2025). We evaluate two settings: **Augmentation** (§4.1), where models generate notes using human-provided evidence, and **Automation** (§4.2), where models retrieve their own evidence. To ensure fair comparison in the automation setting, *we restrict the retrieval*

quota and search timeframe to match the exact conditions available to the human note author. Finally, to reflect platform constraints, all generated notes are strictly truncated to the 280-character limit during helpfulness evaluation. Detailed model specifications, evidence retrieval configurations, and constraint setups are provided in **Appendix E**.

Main Results. Table 2 summarizes performance across both generation modes. We highlight six observations. (1) Models perform substantially worse on the *Not Helpful* subset, confirming its higher difficulty. (2) Human-written notes rated 100% *Helpful* by the crowd achieve only 73.19% under our framework, revealing weaknesses in current voting (see §7.1 for further analysis). (3) Models with over 14B parameters surpass humans in helpfulness, demonstrating the effectiveness of both augmentation and automation (see details in §7.2). (4) For closed-source LRMs and LLMs, automation consistently outperforms augmentation, suggesting that with well-guided retrieval, models can independently compose grounded notes. (5) The reasoning-enabled o3 model achieves highest overall scores, indicating benefits from explicit reasoning traces. (6) Domain-specific models such as MedGemma-27B outperform general-purpose models (e.g., Qwen3-32B), especially on *Not Helpful* cases, reflecting stronger medical grounding.

7 Discussion

Building on the comparative performance results in §6, we now turn to a deeper analysis of our framework’s components. We structure this discussion around three key research questions (RQs):

- **RQ1: Evaluation Reliability (§7.1):** How does CROWDNOTES+ identify and address validity gaps in crowd ratings via hierarchical evaluation?
- **RQ2: Generation Quality (§7.2):** To what extent does CROWDNOTES+ improve note correctness and helpfulness?
- **RQ3: Evidence Utility (§7.3):** How does the utility of evidence retrieved by CROWDNOTES+ compare to human-provided sources?

7.1 CROWDNOTES+ Addresses Loopholes in Crowd-Sourced Helpfulness Evaluation

Our hierarchical evaluation (§4.3) reveals a core limitation in the current Community Notes voting system: **many notes rated as *Helpful* by humans fail basic relevance or correctness.** As shown in

Setting → Model ↓	Helpful (634)					Not Helpful (634)					Overall		
	Note Aug. (R=89.27)		Note Auto.			Note Aug. (R=71.45)		Note Auto.			Aug.	Auto.	
	C	H	R	C	H	C	H	R	C	H	H	H	
Human Baseline	75.24	73.19	89.27	75.24	73.19	44.32	5.52	71.45	44.32	5.52	39.36		
G1	Gemini-2.5-pro†	88.64	85.65 ↑	95.74	<u>93.85</u>	<u>91.17</u> ↑	70.50	37.54 ↑	91.96	<u>90.22</u>	69.24 ↑	61.60 ↑	<u>80.21</u> ↑
	o3†	87.70	86.91 ↑	95.74	94.16	92.11 ↑	68.30	40.69 ↑	91.96	89.91	70.19 ↑	63.80 ↑	81.15 ↑
	Grok-4†	86.44	82.65 ↑	95.74	92.74	88.17 ↑	67.98	32.81 ↑	91.96	89.27	67.19 ↑	57.73 ↑	77.68 ↑
G2	GPT-4.1	<u>87.85</u>	<u>85.80</u> ↑	<u>94.64</u>	92.90	88.49 ↑	<u>69.56</u>	<u>40.22</u> ↑	<u>93.06</u>	90.85	<u>69.87</u> ↑	<u>63.01</u> ↑	79.18 ↑
	Claude-4-Opus	85.17	83.60 ↑	<u>94.64</u>	89.43	85.96 ↑	63.88	37.85 ↑	<u>93.06</u>	84.70	64.51 ↑	60.73 ↑	75.24 ↑
G3	Qwen3-32B	81.39	76.66 ↑	90.69	80.28	70.35 ↓	60.57	28.86 ↑	87.22	77.13	55.84 ↑	52.76 ↑	63.10 ↑
	Qwen3-14B	76.03	70.82 ↓	90.69	76.03	66.09 ↓	56.15	23.03 ↑	87.22	71.29	50.63 ↑	46.93 ↑	58.36 ↑
	Llama-3.1-8B	67.98	61.36 ↓	86.59	60.41	49.05 ↓	51.10	17.98 ↑	83.75	61.83	36.28 ↑	39.67 ↑	42.67 ↑
	Ministral-8B	56.94	51.58 ↓	86.59	53.31	44.32 ↓	43.22	14.67 ↑	83.75	51.74	27.60 ↑	33.13 ↓	35.96 ↓
	Qwen3-8B†	70.35	64.67 ↓	86.59	65.30	53.63 ↓	47.00	18.14 ↑	83.75	58.83	34.86 ↑	41.41 ↑	44.25 ↑
	Qwen3-8B	69.56	64.83 ↓	86.59	65.62	55.36 ↓	47.63	19.09 ↑	83.75	61.20	38.80 ↑	41.96 ↑	47.08 ↑
G4	Lingshu-32B	79.34	73.19 –	91.96	78.70	67.35 ↓	58.99	22.08 ↑	93.85	81.70	52.37 ↑	47.64 ↑	59.86 ↑
	MedGemma-27B	84.38	79.02 ↑	91.96	85.96	79.81 ↑	65.46	30.91 ↑	93.85	86.91	58.68 ↑	54.97 ↑	69.25 ↑
	Lingshu-7B	58.04	50.47 ↓	85.65	53.63	41.80 ↓	43.38	13.56 ↑	85.33	60.41	33.91 ↑	32.02 ↓	37.86 ↓
	MedGemma-4B	60.41	52.68 ↓	85.65	53.63	40.06 ↓	43.53	16.56 ↑	85.33	56.31	31.23 ↑	34.62 ↓	35.65 ↓

Table 2: Effectiveness (%) of 15 representative LLMs across note augmentation (§4.1) and automation (§4.2) settings on HEALTHNOTES. Human Baseline refers to original human-written Community Notes. Evaluation measures: R = relevance, C = correctness, H = helpfulness (§4.3). Model groups: G1 = closed-source LRMs, G2 = closed-source LLMs, G3 = open-source LLMs, G4 = domain-specific medical LLMs. † denotes reasoning-enabled models; Identical R scores under Note Auto. indicate shared retriever LLM for query generation and utility judgment (see §E.1 and Table 8). Best and second-best results are shown in bold and underline.

Misleading Post: The American Heart Association (AHA) has warned that 90 percent of the vaccinated population now suffers from an irreversible heart condition caused by the COVID-19 vaccines.

Human-Provided Evidence: <https://newsroom.heart.org/news/heart-disease-risk-prevention-and-management-redefined>
Content: Interactions among obesity, Type 2 diabetes, chronic kidney disease and cardiovascular disease drive the new approach, says new American Heart Association presidential advisory...

✗ The URL only provided general information about heart disease risks and prevention methods, but did not mention COVID-19 vaccines or related effects.

Figure 4: Example of a human-written note mislabeled as *Helpful* by human voters but correctly identified as *Not Helpful* by CROWDNOTES+ due to citing irrelevant evidence.

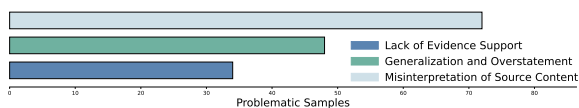


Figure 5: Error distribution of 89 human-written notes that misrepresented evidence, grouped by three primary causes.

Table 2, our framework aligns closely with human judgments on the *Not Helpful* subset (only 5.5% divergence) but shows significant drops on *Helpful*: 11.7% for relevance and 14.0% for correctness.

To investigate these inconsistencies, we examine two failure types in notes mislabeled by humans as “*Helpful*.” First, some notes show no meaningful connection between their claims and cited evidence (Figure 4). Second, we conduct a focused, qualita-

tive analysis of 89 notes that our framework rates as relevant but incorrect, yet were judged helpful by humans. Two human experts reviewed these cases and reached consensus on error attribution. As shown in Figure 5, three recurring causes emerge: (1) *Lack of Evidence Support*, where claims are not substantiated by the cited sources; (2) *Misinterpretation of Source Content*, where factual details are distorted or misrepresented; and (3) *Overgeneralization*, where the note draws conclusions not warranted by the evidence.

These findings suggest human voters often reward stylistic fluency over factual rigor in helpfulness judgment. By enforcing staged checks for relevance and correctness, CROWDNOTES+ substantially reduces false positives and provides a more reliable basis for helpfulness evaluation.

7.2 CROWDNOTES+ Produces Better Notes

We next evaluate CROWDNOTES+ in two settings: (1) augmentation (§4.1), where models write notes from human-provided evidence, and (2) automation (§4.2), where models retrieve evidence and generate notes end to end.

Better Use of the Same Evidence. Under augmentation, CROWDNOTES+ produces notes with higher correctness than human-written notes given the same evidence (Table 2). This suggests stronger

[Misleading Post: Dr. Ryan Cole claims COVID vaccines caused a spike in cancer.]

<p>Human-Provided Evidence: https://pubmed.ncbi.nlm.nih.gov/34919267/ Content: The coronavirus disease 2019 (COVID-19) pandemic has <i>impacted health care delivery worldwide</i>. Cancer is a leading cause of death, and <i>the impact of the pandemic on cancer diagnoses is an important public health concern</i>... Beginning March 4, 2020 ... weekly cancer cases declined precipitously ... followed by a moderate recovery ... Thereafter, weekly cancer cases trended slowly back toward pre-COVID-19 baseline levels. ... Following the pandemic onset, there was a cumulative year-over-year decline in cancer cases overall of 7.3%... The data in this study demonstrate a substantial reduction in cancer diagnoses following the onset of COVID-19...</p>
<p>Human-Written Note: Dr. Ryan Cole claims that vaccination has caused an increase in the cancer rate... There has actually been a decline in cancer rates since covid-19 began. [PubMed] ❌ Misinterprets PubMed study → implies actual drop in cancer incidence</p>
<p>CrowdNotes+ (GPT-4.1): There is no evidence that COVID-19 vaccines have caused an increase in cancer rates; studies show a temporary decline in cancer diagnoses during the pandemic, likely due to reduced healthcare access, not vaccination. [PubMed] ✅ Provides full PubMed context; explains what might have caused the decline</p>
<p>CrowdNotes+ (MedGemma-27B): A study analyzing electronic medical records found a substantial reduction in cancer diagnoses following the onset of COVID-19, which returned to pre-pandemic levels about 12 months later. [PubMed] ✅ Provides full PubMed context; describes decline then subsequent increase</p>

Figure 6: **Effectiveness of CROWDNOTES+ augmentation:** Given the same evidence, the note generated by CROWDNOTES+ supplies complete contextual information that the human-written note omits.

Model	Helpful	Not Helpful	Overall
CROWDNOTES+ (o3)	92.11	70.19	81.15
- Query Diversity	79.50	69.09	74.30
- Utility Judgment	79.02	64.83	71.93
CROWDNOTES+ (MedGemma-27B)	79.81	58.68	69.25
- Query Diversity	74.76	54.73	64.75
- Utility Judgment	66.25	50.47	58.36

Table 3: **Effectiveness of CROWDNOTES+ automation:** Ablation performance in note helpfulness (%) of utility-guided note automation in CROWDNOTES+.

faithfulness to the cited sources and clearer use of supporting context. Figure 6 shows that CROWDNOTES+ often recovers important evidence details omitted in human-written notes, improving completeness and interpretability.

Query Diversity and Utility Judgment Both Matter. In automation, both retrieval components are important. As shown in Table 3, removing either diverse query generation or utility judgment substantially reduces overall helpfulness. Query diversity expands the candidate evidence pool, while utility judgment prioritizes the most informative and reliable sources.

Humans Prefer CROWDNOTES+ Notes. We also conduct a human evaluation in automation mode (§4.2), comparing human-written notes with notes generated by CROWDNOTES+. Three annotators performed pairwise comparisons on 100 randomized and anonymized note pairs from HEALTHNOTES, including 50 *Helpful* and 50 *Not Helpful* cases. For each pair, they selected the better note based on accuracy, relevance, specificity, neutrality,

Model	Win Rate (%)
CROWDNOTES+ (o3)	87.0
CROWDNOTES+ (GPT-4.1)	77.0
CROWDNOTES+ (MedGemma-27B)	62.0

Table 4: **Human preference for CROWDNOTES+ notes on 100 note pairs.** Win rate denotes the percentage of pairwise comparisons in which annotators preferred CROWDNOTES+ over the human-written note.

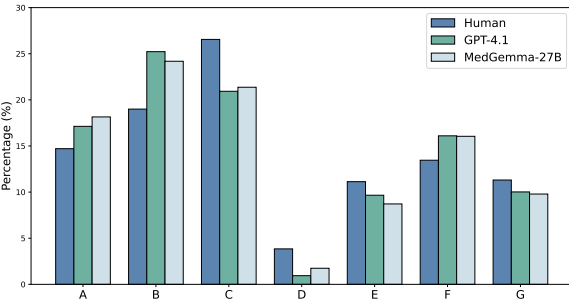


Figure 7: **Comparison of human-selected and CROWDNOTES+ evidence sources.** A: Health Authorities; B: Research Literature; C: News Media; D: Social Media; E: Health Portals; F: Commercial / Advocacy / NGO Sites; G: Others.

and helpfulness. Table 4 reports win rates computed from the aggregated annotator votes. The results align with our automatic evaluation: annotators consistently prefer notes generated by CROWDNOTES+, indicating higher-quality explanatory context than human-written notes.

7.3 CROWDNOTES+ Selects Better Evidence

To understand whether CROWDNOTES+ retrieves better supporting evidence than average human contributors, we compare human-selected and CROWDNOTES+-selected evidence along two dimensions: (1) source distribution and (2) practical utility. We first examine where the evidence comes from, then assess how useful it is for producing helpful, well-grounded notes.

CROWDNOTES+ Locates More Authoritative Sources. We examine the distribution of human-selected and CROWDNOTES+-selected evidence across seven health-related categories: (1) Health Authorities, (2) Research Literature, (3) News Media, (4) Social Media, (5) Health Portals, (6) Commercial / Advocacy / NGO Sites, and (7) Others. A web-search-enabled GPT-4.1 model assigns each source to its primary category. As shown in Figure 7, humans rely more on news media, social media, and general health portals, whereas LLMs

Model (vs. Human)	Win	Lose	Tie
CROWDNOTES+ (o3)	65.85	22.48	11.67
CROWDNOTES+ (MedGemma-27B)	57.57	33.20	9.23

Table 5: **Overall, CROWDNOTES+ selects higher-utility evidence than humans**, demonstrated through pairwise comparisons (%) of evidence utility between human-provided and LLM-selected sources (Figure 3).

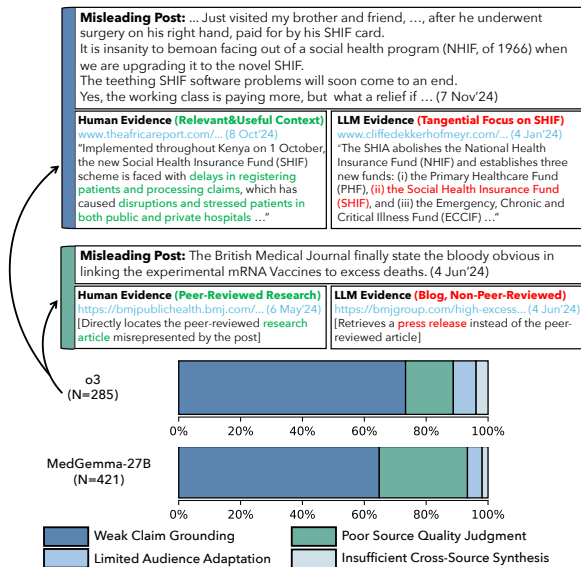


Figure 8: **Why human-provided evidence is sometimes preferred over evidence selected by CROWDNOTES+.** Distribution of failure types among cases where human-provided sources are judged more useful.

more often select institutional and agency sources. This shift toward more authoritative evidence helps explain why the automation mode consistently outperforms the augmentation mode in Table 2.

CROWDNOTES+ Often Outperforms Human Evidence Selection. To quantify utility, we compare human evidence \mathcal{E}_h and machine-selected evidence \mathcal{E}_m on all 1,268 samples in HEALTHNOTES. For each pair, a web-search-enabled GPT-4.1 judge evaluates which evidence set is more useful for producing a helpful note, with CROWDNOTES+ instantiated using o3 and MedGemma-27B. Table 5 shows win rates above 50%, indicating that CROWDNOTES+ frequently matches or surpasses human evidence selection.

We next examine cases where human evidence is preferred. Two human experts first collaboratively reviewed 100 cases where human evidence was preferred and identified four recurring causes: (1) *Weak Claim Grounding*, where the LLM misses the core claim or retrieves only loosely relevant evidence; (2) *Poor Source Quality Judgment*, where

the model fails to distinguish stronger sources from weaker ones; (3) *Limited Audience Adaptation*, where retrieved sources are overly technical or inaccessible; and (4) *Insufficient Cross-Source Synthesis*, where the model does not combine multiple sources into a coherent conclusion. The remaining cases were labeled using GPT-4.1.

Figure 8 highlights the first two causes and provides representative examples. In the **Weak Claim Grounding** case, a post praises Kenya’s transition from the National Health Insurance Fund (NHIF) to the Social Health Insurance Fund (SHIF) based on anecdotal experience. The human-selected evidence directly addresses this claim by citing reputable reporting on delays in SHIF registration and claim processing that disrupted hospital services. In contrast, o3 retrieves a high-level overview of the NHIF-to-SHIF transition that is only tangentially relevant. In the **Poor Source Quality Judgment** case, the post misrepresents a research article as linking mRNA vaccines to excess deaths. Humans retrieve the original peer-reviewed article from the British Medical Journal, whereas the LLM retrieves a secondary press release, reflecting weaker source credibility judgment.

Overall, while LLMs often select high-utility evidence (as shown in Table 5), their remaining failures point to shallow retrieval or insufficient integration across evidence sources, suggesting room for improvement in query formulation, multi-hop reasoning, and credibility-aware search.

8 Conclusion and Future Work

We identify a substantial latency gap in crowd-sourced health Community Notes, where a median delay of 17.6 hours causes corrective interventions to trail misinformation spread. To address this, we propose CROWDNOTES+, a unified framework that augments note generation and evaluation through evidence grounding, utility-guided automation, and hierarchical assessment. Experiments on HEALTHNOTES show that CROWDNOTES+ can produce more accurate and helpful notes than human contributors, while also exposing a key weakness in current crowd evaluation, where note fluency is often mistaken for factual accuracy. These findings support a shift toward human–AI collaboration (see **Appendix A**), in which LLMs act as evidence-grounded assistants that improve the speed and reliability of community-based moderation, with clear paths for extension across domains, languages, and integrated detection pipelines.

Limitations

Our work offers an important first step toward LLM-augmented Community Notes in the health domain, hinting several extensions that can broaden its scope and impact. *First, our investigation focuses on health content in English language.* While health misinformation presents a high-stakes and well-defined setting, applying CROWDNOTES+ to more subjective topics (e.g., political and socio-cultural discourse) or low-resource languages may introduce new challenges regarding subtle boundaries and consensus that are not captured in this study. *Second, although CROWDNOTES+ improves evidence utility over human contributors, it remains constrained by the reasoning capabilities of current LLMs in evidence retrieval.* As observed in §7.3, LLMs sometimes rely on surface-level lexical overlap rather than deeper semantic reasoning when selecting evidence, indicating that advances in retrieval backbones are important for handling complex, multi-hop claims. *Finally, we evaluate CROWDNOTES+ as a standalone module for advancing note creation and helpfulness assessment.* We do not model upstream detection or prioritization of misleading posts, which would be required to support fully end-to-end, real-time intervention.

Ethical Considerations

Potential Harms and Safety. Although CROWDNOTES+ is designed to mitigate health misinformation, deploying generative models in medical contexts carries inherent risks. A central concern is **hallucination**, where a model may produce fluent but inaccurate notes. If surfaced without oversight, such errors could lead to real-world harm. To mitigate this risk, *we position CROWDNOTES+ strictly as a human-augmenting system rather than a fully autonomous decision-maker.* We explicitly discourage end-to-end automation in health misinformation governance and treat human verification of retrieved evidence as a required safety layer.

Automation Bias. While our study identifies the “fluency trap” in human voting, introducing AI assistance introduces the complementary risk of **automation bias**, where moderators may over-trust model outputs due to their authoritative tone. Rapid generation may also incentivize speed over careful scrutiny. To counteract this risk, *future interfaces built on CROWDNOTES+ should promote active human engagement*, for example by requiring mod-

erators to inspect or validate specific evidence snippets rather than simply approving generated notes.

Dual Use and Fairness. Automated fact-checking technologies have inherent dual-use potential. The same retrieval and generation mechanisms could be misused to produce persuasive, citation-backed disinformation or to selectively suppress legitimate scientific debate through biased evidence selection. In addition, reliance on indexed English-language sources may introduce **western-centric bias**, potentially under-representing non-English or local health authorities. *Ongoing auditing of retrieval sources and deliberate inclusion of diverse perspectives are therefore essential.*

Compliance with Platform Policies. All data collection and usage in this work comply with platform policies and public data guidelines. X posts and web evidence were obtained through authorized APIs and exclude private or personally identifiable information. To balance reproducibility with user privacy, we will release HEALTHNOTES under controlled, research-only access.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (T1 251RES2508) and MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). We thank Sahajpreet Singh (National University of Singapore) for early conversations related to the Community Notes concept.

References

- Funmi Adebesein, Hanlie Smuts, Tendani Mawela, George Maramba, Marie Hattingh, and 1 others. 2023. The role of social media in health misinformation and disinformation during the COVID-19 pandemic: bibliometric analysis. *JMIR infodemiology*, 3(1):e48620.
- Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393.
- Anthropic. 2025. Introducing claude 4. <https://www.anthropic.com/news/claude-4>.
- Yuwei Chuai, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini, and Nicolas Pröllochs. 2024a. Community-based fact-checking reduces the spread of misleading posts on social media. *arXiv preprint arXiv:2409.08781*.

- Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024b. Did the roll-out of community notes reduce engagement with misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).
- Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. 2025. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM on Web Conference 2025*, page 3751–3761.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024. Missci: Reconstructing fallacies in misrepresented science. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4372–4405, Bangkok, Thailand.
- William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, 1(1).
- Google. 2025. Gemini 2.5 pro. <https://deepmind.google/technologies/gemini/pro/>.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 22105–22113.
- Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, and 1 others. 2020. COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4):1621.
- Haiwen Li, Soham De, Manon Revel, Andreas Haupt, Brad Miller, Keith Coleman, Jay Baxter, Martin Saveski, and Michiel A Bakker. 2025. Scaling human judgment in community notes with LLMs. *arXiv preprint arXiv:2506.24118*.
- Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G Rand. 2024. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, 19(2):477–488.
- Mistral AI Team. 2024. Un ministral, des ministraux. <https://mistral.ai/news/ministraux>.
- OpenAI. 2025a. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>.
- OpenAI. 2025b. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.
- Jan Pfänder and Sacha Altay. 2025. Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. *Nature human behaviour*, pages 1–12.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia.
- Thomas Renault, David Restrepo Amariles, and Aurore Troussel. 2024. Collaboratively adding context to social media posts reduces the sharing of false news. *arXiv preprint arXiv:2404.02803*.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.
- Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search result diversification. *Found. Trends Inf. Retr.*, 9(1):1–90.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Maryam Shahbazi and Deborah Bunker. 2024. Social media trust: Fighting misinformation in the time of crisis. *International Journal of Information Management*, 77:102780.
- Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media*, 22:100104.
- Jane B Singer. 2023. Closing the barn door? Fact-checkers as retroactive gatekeepers of the COVID-19 “infodemic”. *Journalism & Mass Communication Quarterly*, 100(2):332–353.
- Sahajpreet Singh, Jiaying Wu, Svetlana Churina, and Kokil Jaidka. 2025. On the limitations of LLM-synthesized social media misinformation moderation. In *Proceedings of the ICLR 2025 Workshop ICBINB*.

- Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. 2025. Community notes reduce engagement with and diffusion of false information online. *Proceedings of the National Academy of Sciences*, 122(38):e2503413122.
- Herun Wan, Jiaying Wu, Minnan Luo, Zhi Zeng, and Zhixiong Su. 2025. Truth over tricks: Measuring and mitigating shortcut learning in misinformation detection. *arXiv preprint arXiv:2506.02350*.
- Yuyan Wang, Cheenar Banerjee, Samer Chucuri, Fabio Soldo, Sriraj Badam, Ed H. Chi, and Minmin Chen. 2025. Beyond item dissimilarities: Diversifying by intent in recommender systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 2672–2681.
- Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. 2022. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*.
- Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. 2024a. Result Diversification in Search and Recommendation: A Survey. *IEEE Transactions on Knowledge & Data Engineering*, 36(10):5354–5373.
- Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024b. Fake news in sheep’s clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3367–3378.
- Jiaying Wu and Bryan Hooi. 2023. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 2582–2593.
- Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-align: Prompt-based social alignment for few-shot fake news detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 2726–2736.
- Amelie Wuehrl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding fine-grained distortions in reports of scientific findings. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6175–6191, Bangkok, Thailand.
- xAI. 2025. Grok 4. <https://x.ai/news/grok-4>.
- Rui Xing, Preslav Nakov, Timothy Baldwin, and Jey Han Lau. 2026. COMMUNITYNOTES: A dataset for exploring the helpfulness of fact-checking explanations. In *Findings of the Association for Computational Linguistics: EACL 2026*, Rabat, Morocco.
- Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of retrieval-augmented generation: A survey. In *Big Data*, pages 102–120.
- Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1941–1951.
- Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011.
- Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, page 3465–3476.
- Xinyi Zhou, Ashish Sharma, Amy X Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model. *arXiv preprint arXiv:2403.11169*.

A Discussion: Implications for Human–AI Collaborative Misinformation Governance

LLMs as End-to-End Assistants in the Note Creation Pipeline. Our findings in §7.1 and §7.2 suggest that integrating LLMs into Community Notes for (1) evidence selection, (2) note generation, and (3) hierarchical evaluation can *substantially improve the relevance, correctness, and helpfulness of crowd-sourced misinformation mitigation.*

LLM Support for Evidence Selection and Note Generation. As discussed in §7.2, the quality and appropriateness of evidence play a central role in shaping note accuracy. When LLMs are given the same human-selected sources (Figure 6), they are able to organize and synthesize this evidence more effectively during note generation. Building on this foundation, the strong performance of **utility-guided automation** (§7.3, Table 2) shows that LLMs can also enhance the evidence selection process itself by retrieving more authoritative and contextually relevant sources. These improvements in evidence availability and quality naturally lead to notes with stronger factual grounding. Future refinements such as intent-aware search (Wang et al., 2025) and query diversification (Wu et al., 2024a) may further strengthen this evidence foundation and support even more reliable note generation.

LLM Support for More Reliable Evaluation. Recent commentary from the X Community Notes Team (Li et al., 2025) envisions a hybrid workflow that continues to rely on human voting for helpfulness assessment. In contrast, our analysis in §7.1 shows that human voting often rewards stylistic fluency even when factual support is weak. CROWDNOTES+ addresses this issue through a **hierarchical evaluation pipeline** (§4.3) that assesses relevance, correctness, and helpfulness in sequence using reliable judges (Appendix D.4), producing more reliable and interpretable note assessments.

Toward Hybrid Human–AI Governance. Taken together, these findings point to a hybrid human–AI misinformation governance model in which LLMs provide factual rigor, high-quality evidence selection, and consistent first-pass evaluation, while human contributors contribute oversight, social context, and pluralistic perspectives. Such a division of responsibilities offers a path toward more scalable, timely, and trustworthy misinformation governance.

	#. of Notes	#. of Posts	#. of URLs
Helpful	634	608	1,330
Not Helpful	634	622	907

Table 6: Dataset statistics for HEALTHNOTES. Notes span May 2022–Aug 2025, and their corresponding posts span Jun 2020–Jul 2025.

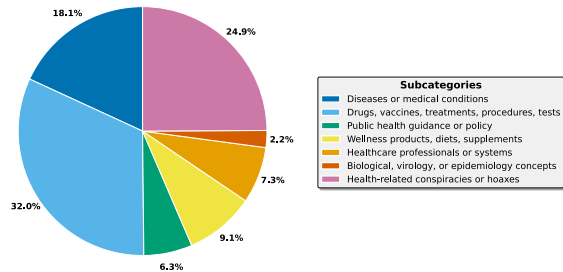


Figure 9: Topic distribution of notes in HEALTHNOTES.

B The HEALTHNOTES Benchmark

Using the 1,268 human-written health-related notes described in §5, we leverage their corresponding post IDs from the public Community Notes dataset⁴ to retrieve the associated flagged posts via the X API.

Table 6 summarizes core statistics of HEALTHNOTES. To examine topical coverage, we group notes by the primary category assigned during the filtering step (following the seven major health-related categories defined in §3.1). As shown in Figure 9, HEALTHNOTES spans a broad range of medical and public health issues. Three categories—diseases or medical conditions, public health guidance and policy, and health-related conspiracies or hoaxes—are particularly prominent, reflecting the types of claims that frequently generate community attention and require timely clarification on social media.

C Details of Note Generation in CROWDNOTES+

This section provides additional details on how CROWDNOTES+ constructs notes in both augmentation and automation modes. We describe (1) how evidence is curated through utility-guided selection (Appendix C.1), (2) how retrieved webpages are processed into evidence chunks (Appendix C.2), and (3) how LLMs synthesize these chunks into contextual notes (Appendix C.3).

⁴<https://x.com/i/communitynotes/download-data>

C.1 Utility-Guided Evidence Curation

In the automation mode (§4.2), evidence is sourced from the Web through a utility-guided selection process rather than human-provided URLs as in the augmentation mode (§4.1). Given a candidate pool \mathcal{P} of evidence snippets (each containing a webpage title and summary from Google Custom Search API⁵), an LLM estimates the *utility* of each snippet for supporting or contextualizing the flagged post. The prompt template used for utility judgment is shown below:

```
SYSTEM_PROMPT = "You are a careful selector.
Output exactly ONE integer as instructed."
```

```
You are selecting one source (healthcare-related Com-
munity Note utility).
This is selection round #{round_no}. Choose exactly
ONE result that has the highest utility.
```

```
## Utility should reflect whether the search result is:
- Relevant to the tweet's topic.
- Likely to add meaningful background or clarifica-
tion.
- Reliable enough to be worth retrieving.
```

```
## OUTPUT FORMAT (critical):
- Output EXACTLY one integer, the index of your
chosen item (1..{len(items_remaining)}).
- No extra words. No numbering other than the single
integer. No explanations.
```

```
## Tweet: {tweet}
```

```
## Search Results (candidates):
[{{idx}}] Title: {title} Snippet: {snippet} URL: {url}
```

Across τ iterative rounds, the highest-utility snippet is selected and removed from \mathcal{P} , yielding a final quota of τ evidence items. The URLs associated with these items form the machine-selected evidence set \mathcal{E}_m , which is subsequently used for retrieval and note generation. The distributional differences between human- and LLM-selected evidence are shown in Figure 7.

C.2 Evidence Retrieval and Processing

For each evidence set, whether human-provided (\mathcal{E}_h) or LLM-selected (\mathcal{E}_m), we retrieve the corresponding webpages using the Jina API⁶. Retrieved pages are cleaned to remove non-essential elements such as headers, footers, navigation bars, and reference sections. The remaining body text is segmented into overlapping passages of 512 tokens with a 128-token overlap.

⁵<https://developers.google.com/custom-search/>

⁶<https://jina.ai/>

Each 512-token passage is embedded using sentence-transformers/all-mpnet-base-v2, and within each evidence piece, the passage with the highest semantic similarity to the flagged post p is selected. These selected passages form the evidence chunk sets \mathcal{C}_h (human-selected sources) or \mathcal{C}_m (LLM-selected sources), which are then used for note generation.

C.3 Note Generation

Given the evidence chunks, either human-provided (\mathcal{C}_h) or LLM-retrieved (\mathcal{C}_m), CROWDNOTES+ generates contextual notes for flagged posts identified as potentially misleading. Both the augmentation and automation settings (§4.1 and §4.2) employ the same prompt template for note generation:

```
SYSTEM_PROMPT = "Community notes is
a collaborative way to add helpful context
to posts and keep people better informed. Now
you are a highly experienced community note writer."
```

```
Task: Write a community note based ONLY on the
source snippets below.
```

```
Hard constraints:
```

- The note MUST be in English.
- DO NOT include any URLs in the note.
- The note MUST be a single line (no line breaks, no bullets).
- Note length MUST be \leq {budget_chars} characters. Do not exceed this budget.
- Be specific, objective, and verifiable.

```
Tweet: {tweet}
```

```
Source snippets:
```

```
[S{index}] url (chunk {chunk_id}) {text}
Output only the note content. Remember: length  $\leq$ 
{budget_chars}, no URLs.
```

The model conditions on the flagged post p and the selected evidence chunks to produce a concise, fact-grounded explanation. The generated note text is paired with its corresponding evidence URLs in the final output, ensuring transparency and traceability in line with Community Notes conventions.

D Details of Hierarchical Evaluation in CROWDNOTES+

As introduced in §4.3, CROWDNOTES+ employs a three-step hierarchical evaluation scheme in which a note advances to the next stage only after passing the current one. This appendix provides implementation details for the three progressive evaluation stages: (1) evidence relevance (Appendix D.1), (2) evidence representation correctness (Appendix D.2), and (3) note helpfulness (Appendix

D.3). We additionally report **human and automated assessments validating the reliability of the judge models** used at each stage (Appendix D.4).

D.1 Evidence Relevance

Setup. The relevance stage assesses whether the retrieved evidence provides meaningful factual context or clarification that helps readers evaluate the claim made in the post. We use GPT-4.1 to perform this assessment via the following prompt:

```
SYSTEM_PROMPT = "You are a very meticulous inspector."
You are given a Tweet and one or more Source snippets: Tweet: {tweet}

Source snippets:
[S{index}] {url} (chunk {chunk_id}) {text}

Task: Determine whether any of the Source snippets adds meaningful factual background, clarification, or supporting information that helps better understand or evaluate the claim made in the Tweet.
1. Check each snippet independently.
2. If at least one snippet meets the requirements, output "Final decision: yes"; otherwise output "Final decision: no".
```

D.2 Evidence Representation Correctness

Conditioned on passing relevance, we next evaluate whether the note accurately represents the cited sources, avoiding factual errors, exaggeration, and misleading framing. This step also uses GPT-4.1 with the prompt shown below:

```
SYSTEM_PROMPT = "You are a very meticulous inspector"

You are given a Community note and one or more Source snippets:
Community Note:
{note}
Source snippets:
[S{index}] {url} (chunk {chunk_id}) {text}

Task: Decide whether the Community note distorts the information in any of the provided Source snippets.
1. Check each snippet independently.
2. If at least one distortion is found, output "Final decision: yes"; otherwise output "Final decision: no".
```

D.3 Note Helpfulness

Conditioned on passing correctness, the final stage assesses whether a note provides useful context that helps readers understand or critically evaluate the flagged post, following the official Community

Notes guidelines. We use HEALTHJUDGE (a fine-tuned Lingshu-7B model (Li et al., 2025)) with temperature 0 for deterministic and domain-adapted scoring. To mirror platform constraints, **this is the only stage where the 280-character limit used by Community Notes is applied**: if the combined length of the note text and URLs exceeds 280 characters (URLs count as one each), the note text is truncated before evaluation.

```
SYSTEM_PROMPT = "You are a precise text classifier."
```

```
You are given a Tweet and its corresponding Note:
Tweet: {tweet}
Note: {note (only note text, w/o evidence URLs)}
```

The purpose of note is to add helpful context to tweet and keep people better informed. Your task is to evaluate whether the Note is Helpful or Not Helpful based on the following criteria:

****Helpful Criteria:****

- Clear and/or well-written
- Cites high-quality sources
- Directly addresses the Tweet's claim
- Provides important context
- Neutral or unbiased language
- Other (any additional positive reason)

****Not Helpful Criteria:****

- Incorrect information
- Sources missing or unreliable
- Misses key points or irrelevant
- Hard to understand
- Argumentative or biased language
- Spam, harassment, or abuse
- Sources do not support note
- Opinion or speculation
- Note not needed on this Tweet
- Other (any additional negative reason)

Instructions:

1. Carefully read the Tweet and the Note.
2. Analyze the Note using the Helpful and Not Helpful criteria above.
3. Respond with "Final decision: yes" (if Helpful) or "Final decision: no" (if Not Helpful).

HEALTHJUDGE Training Setup

HEALTHJUDGE is trained on human-labeled health-related post–note pairs, using only note text (without appended URLs) to ensure that helpfulness judgments reflect explanatory quality rather than evidence relevance or correctness. The dataset contains 2,971 Helpful and 742 Not Helpful post–note pairs, with 1,000 pairs (800 Helpful, 200 Not Helpful) reserved for evaluation.

Each instance is formatted as a chat prompt following the helpfulness evaluation template, with loss applied only to the final decision tokens ("Final

Model	Macro-F1 (%)	Macro-Accuracy (%)
GPT-4.1	74.28	74.19
Gemini-2.5-flash	68.36	65.13
Claude-Sonnet-4	78.14	76.44
Lingshu-32B	64.71	62.25
Lingshu-7B	51.66	51.63
HEALTHJUDGE	81.03	81.44

Table 7: **Effectiveness of HEALTHJUDGE for note helpfulness assessment**, validated by its superior performance on 1,000 unseen post–note pairs.

decision: yes/no”) and left padding used for causal alignment. Training uses full-parameter fine-tuning for 2 epochs with AdamW (learning rate 1×10^{-5}), gradient accumulation of 16 steps, and bfloat16 precision.

The resulting model produces deterministic, parseable outputs suitable for automatic evaluation. *Although some posts in HEALTHNOTES overlap with those present in HEALTHJUDGE’s training data, all associated notes in HEALTHNOTES are distinct, ensuring that no helpfulness labels or note content leak into evaluation.*

D.4 Judge Reliability Assessment

This section evaluates the reliability of the judge models used at each stage of the hierarchical evaluation in CROWDNOTES+. For relevance and correctness, we assess LLM-as-a-Judge decisions through human evaluation. For helpfulness, we measure HEALTHJUDGE’s alignment with human-labeled ground truth.

D.4.1 Reliability of Relevance Judgments

To assess the reliability of LLM-based evidence relevance judgments (Appendix D.1), we conduct a human evaluation that inspects 100 sampled relevance judgments made by the model: 50 notes derived from the *Helpful* subset of HEALTHNOTES and 50 notes from the *Not Helpful* subset (see §5).

Three graduate student annotators independently labeled each instance following standardized instructions, detailed as follows.

Human Evaluation Objective. The goal of this evaluation is to assess whether the reasoning produced by the LLM judge provides a reasonable and sufficient justification for its final predicted **relevance label**.

Each data instance contains the following fields:

- `id`: data identifier.
- `tweet`: the text of the flagged post.
- `evidence_snippets`: retrieved evidence snippets (each with a URL and its associated text chunk)
- `relevance_label`: the LLM’s predicted relevance label. “Yes” indicates at least one evidence snippet is relevant to the tweet, and “No” indicates all evidence snippets are irrelevant.
- `reasoning`: the LLM’s explanation supporting its prediction.

For reference, the exact prompt used for LLM judgment is reproduced below:

{prompt for evaluating evidence relevance, presented in Appendix D.1}

Annotation Guidelines. For each instance, you will need to assign one of two labels:

- **0 (Reliable):** (a) The reasoning is coherent and clearly articulated. (b) The relevance label is consistent with the reasoning. (c) The final decision is acceptable to a human annotator.
- **1 (Unreliable):** (a) The reasoning is inconsistent with the evidence or the tweet. (b) The reasoning does not justify the final relevance label. (c) There are clear logical errors, misinterpretations, or unsupported conclusions in the LLM reasoning trace.

We report the **agreement rate** between the LLM judge and majority-voted human annotations. **The LLM prediction matches the aggregated human judgment in all 100 cases.** Inter-annotator disagreement occurs in only one instance with a majority *Reliable* label. As this is a verification task where high agreement is expected, this result serves as a sanity check confirming the LLM judge’s consistency with human assessments.

D.4.2 Reliability of Correctness Judgments

To evaluate the reliability of LLM-based correctness judgments (Appendix D.2), we follow a similar procedure as Appendix D.4.1. We sample 100 correctness judgments made by the model: 50 notes derived from posts in the *Helpful* subset and 50 notes from the *Not Helpful* subset.

The same three annotators from Appendix D.4.1 independently assessed whether the LLM’s justification and decision accurately reflected the pro-

vided sources, using the following instructions.

Human Evaluation Objective. The goal of this evaluation is to determine whether the reasoning produced by the LLM provides a reasonable and sufficient justification for its final predicted **distortion label**.

Each data instance contains the following fields:

- id: data identifier.
- note: the Community Note text.
- evidence_snippets: retrieved evidence snippets (each with a URL and its associated text chunk)
- distortion_label: the LLM’s prediction of whether the note distorts the evidence. “Yes” indicates that the note contains at least one instance of misrepresenting the evidence, and “No” indicates that the note does not distort any provided evidence.
- reasoning: the LLM’s explanation supporting its prediction.

For reference, the exact prompt used for LLM judgment is reproduced below:

{prompt for evaluating evidence representation correctness, presented in Appendix D.2}

Annotation Guidelines. For each instance, you will need to assign one of two labels:

- **0 (Reliable):** (a) The reasoning is coherent and clearly articulated. (b) The relevance label is consistent with the reasoning. (c) The final decision is acceptable to a human annotator.
- **1 (Unreliable):** (a) The reasoning conflicts with the content of the note or the evidence. (b) The reasoning does not justify the final distortion label. (c) There are clear logical errors, misinterpretations, or unsupported conclusions in the LLM reasoning trace.

As with the relevance evaluation, we report the **agreement rate** between the LLM judge and majority-voted human annotations. **The LLM prediction matches the aggregated human judgment in 97 out of 100 cases.** Inter-annotator disagreement occurs in only 3 cases, comprising 2 instances with a majority *Reliable* label and 1 instance with a majority *Unreliable* label. As this is also a verification task where high agreement is expected, this result serves as a sanity check confirming the LLM judge’s consistency with human assessments.

D.4.3 Reliability of Helpfulness Judgments

For the final stage, we evaluate HEALTHJUDGE by comparing its *Helpful/Not Helpful* predictions with human-contributed labels on the 1,000 test samples described in Appendix D.3. As shown in Table 7,

HEALTHJUDGE achieves higher alignment with human judgments than GPT-4.1, Claude-4-Sonnet (Anthropic, 2025), and Gemini 2.5 Flash (Google, 2025). These results demonstrate strong reliability for domain-specific helpfulness evaluation.

E Experimental Setup

This section details the setup for evidence acquisition (Appendix E.1), note generation (Appendix E.2), and evaluation constraints (Appendix E.3) used in CROWDNOTES+ experiments.

E.1 Evidence Acquisition Setup

For the **Automation** setting described in §4.2, we select six representative LLMs to perform utility-guided evidence retrieval: o3, GPT-4.1, Qwen3 (32B and 8B), and MedGemma (27B and 4B). Correlations between Retriever LLMs and Generator LLMs are summarized in Table 8.

To ensure fair comparison with human-provided evidence, we apply the following controls:

- **Quota Matching:** The evidence quota τ for each sample equals the number of URLs in the human evidence set ($|\mathcal{E}_h|$).
- **Temporal Restrictions:** *Web search results are constrained to content available up to the timestamp of the human-written note*, preventing access to future information.
- **Passage Extraction:** For each retrieved webpage, we extract the highest-ranked 512-token passage to serve as the evidence snippet for synthesizing notes.

E.2 Note Generation Setup

Under both **Augmentation** (§4.1) and **Automation** (§4.2) settings, we evaluate **15 representative LLMs** grouped into four categories ([G1] to [G4]):

- **[G1] Closed-Source Large Reasoning Models (LRMs):** Models trained with chain-of-thought or extensive reasoning capabilities, including o3 (OpenAI, 2025b), Gemini-2.5 (Google, 2025), and Grok-4 (xAI, 2025).
- **[G2] Closed-Source LLMs:** Standard state-of-the-art proprietary models, specifically GPT-4.1 (OpenAI, 2025a) and Claude-4 (Anthropic, 2025).

Evidence Retriever	Note Generator
o3 †	Gemini-2.5-pro † o3 † Grok-4 †
GPT-4.1	GPT-4.1 Claude-4-Opus
Qwen3-32B	Qwen3-32B Qwen3-14B
Qwen3-8B	Llama-3.1-8B Minstral-8B Qwen3-8B † Qwen3-8B
MedGemma-27B	Lingshu-32B MedGemma-27B
MedGemma-4B	Lingshu-7B MedGemma-4B

Table 8: **Correlation between retriever LLMs (used for query generation and utility judgment) and generator LLMs (used for note generation) in the Automation setting.** This mapping explains identical relevance scores observed across certain generator models (see Table 2). † denotes reasoning-enabled models.

- **[G3] Open-Source LLMs and LRMs:** High-performing open weights models, including Qwen3 (Yang et al., 2025), Llama-3.1 (Dubey et al., 2024), and Minstral (Mistral AI Team, 2024).
- **[G4] Domain-Specific Medical LLMs:** Models fine-tuned for biomedical contexts, such as Lingshu (Xu et al., 2025) and MedGemma (Sellergren et al., 2025).

Unless otherwise specified, we use non-reasoning variants of open-source models at temperature 0 for reproducibility, and run all experiments once. Detailed model specifications are listed in Table 9.

E.3 Note Length Constraints

Community Notes imposes a strict character limit of 280. We mirror this in our evaluation:

- **Constraint Application:** If the combined length of an LLM-generated note and its appended URLs exceeds 280 characters, we truncate the text content. Following X’s policy, URLs count as a single character⁷.
- **Evaluation Scope:** This truncation applies only to the *Helpfulness* evaluation. We do not truncate

⁷<https://docs.x.com/x-api/community-notes/quickstart>

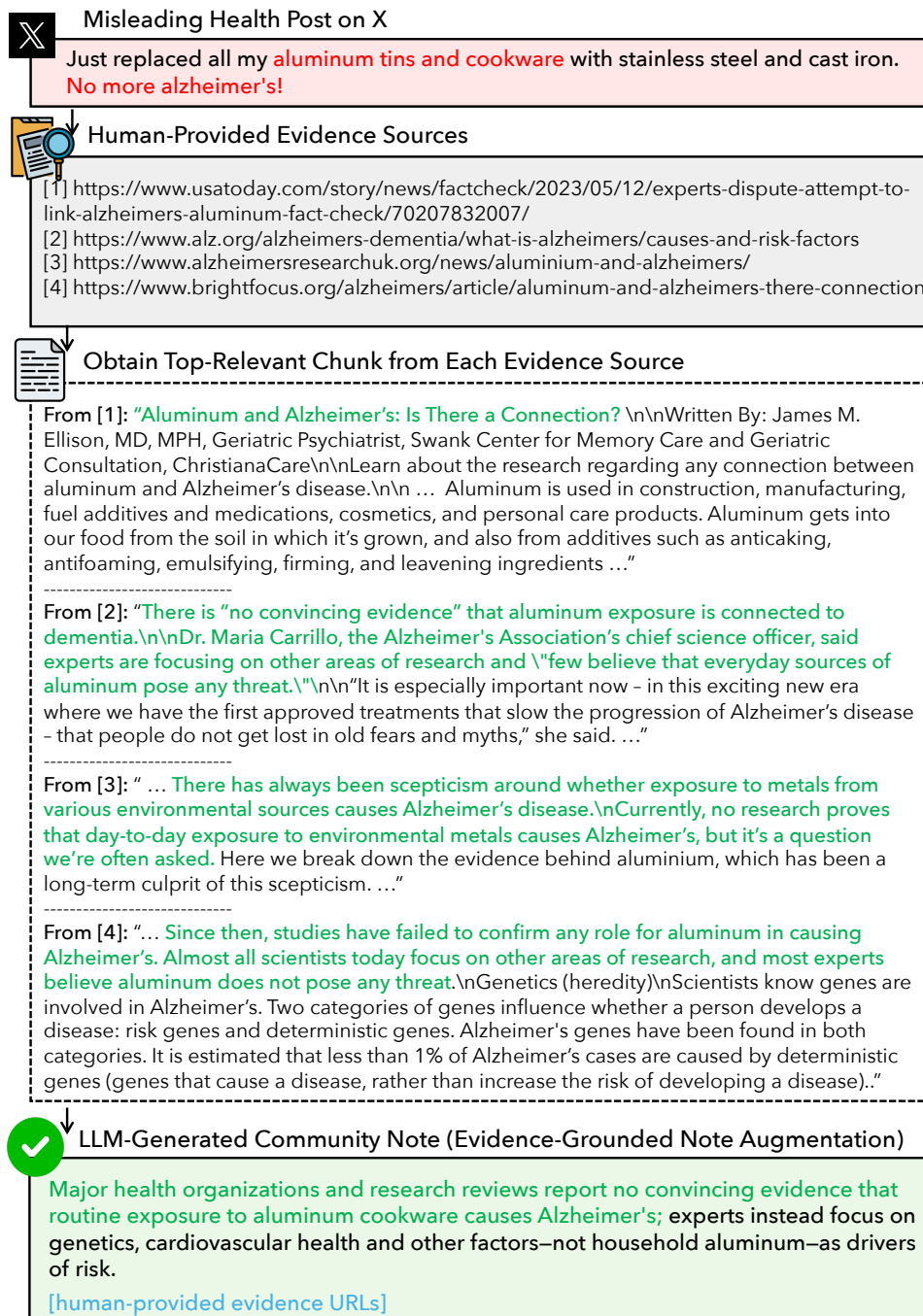
Model	Model Card
Gemini-2.5-Pro o3 Grok-4	gemini-2.5-pro-preview-03-25 o3-2025-04-16 x-ai/grok-4
GPT-4.1 Claude-Opus-4	gpt-4.1-2025-04-14 claude-opus-4-20250514
Qwen3-32B Qwen3-14B Llama-3.1-8B Minstral-8B Qwen3-8B	Qwen/Qwen3-32B Qwen/Qwen3-14B meta-llama/Llama-3.1-8B-Instruct mistralai/Minstral-8B-Instruct-2410 Qwen/Qwen3-8B
Lingshu-32B MedGemma-27B Lingshu-7B MedGemma-4B	lingshu-medical-mlm/Lingshu-32B google/medgemma-27b-text-it lingshu-medical-mlm/Lingshu-7B google/medgemma-4b-it

Table 9: The model versions of LLMs used in CROWDNOTES+ for note generation.

notes for *Relevance* or *Correctness* evaluations, as these metrics assess the logical validity of the generated content rather than its final presentation format.

F Demonstrations of CROWDNOTES+ Workflow

We present two end-to-end examples that illustrate how CROWDNOTES+ performs evidence acquisition, note generation, and hierarchical evaluation. Figure 10 illustrates a case from our evidence-grounded note augmentation setting (§4.1), and Figure 11 illustrates a case from our utility-guided note automation setting (§4.2).



Hierarchical Evaluation:
 Relevance: **yes** Correctness: **yes** Helpfulness: **yes**

Figure 10: **Illustration of CROWDNOTES+ under the evidence-grounded augmentation setting (§4.1).** Using evidence chunks retrieved from human-provided sources, the o3 model synthesizes the information to generate a helpful note, which addresses the post's misleading claim that aluminum exposure causes Alzheimer's disease.

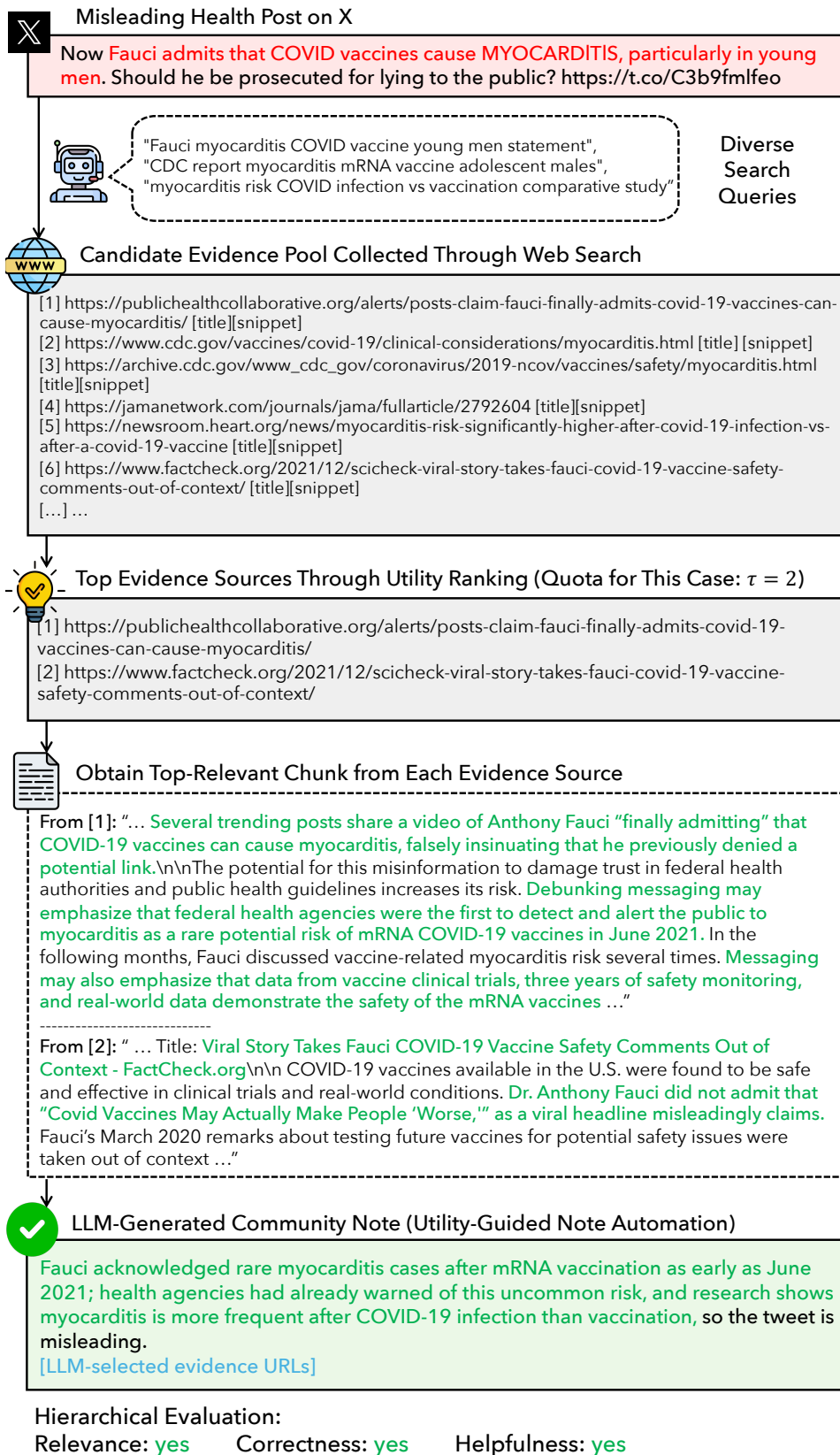


Figure 11: Illustration of CROWDNOTES+ under the utility-guided automation setting (§4.2). Using evidence chunks retrieved from LLM-selected sources, the o3 model synthesizes the information to generate a helpful note addressing the misleading claim that Fauci “admitted” COVID vaccines cause myocarditis. **For a fair comparison with human-written notes, the evidence quota for this case is set to $\tau = 2$ to match the number of human-provided sources.**