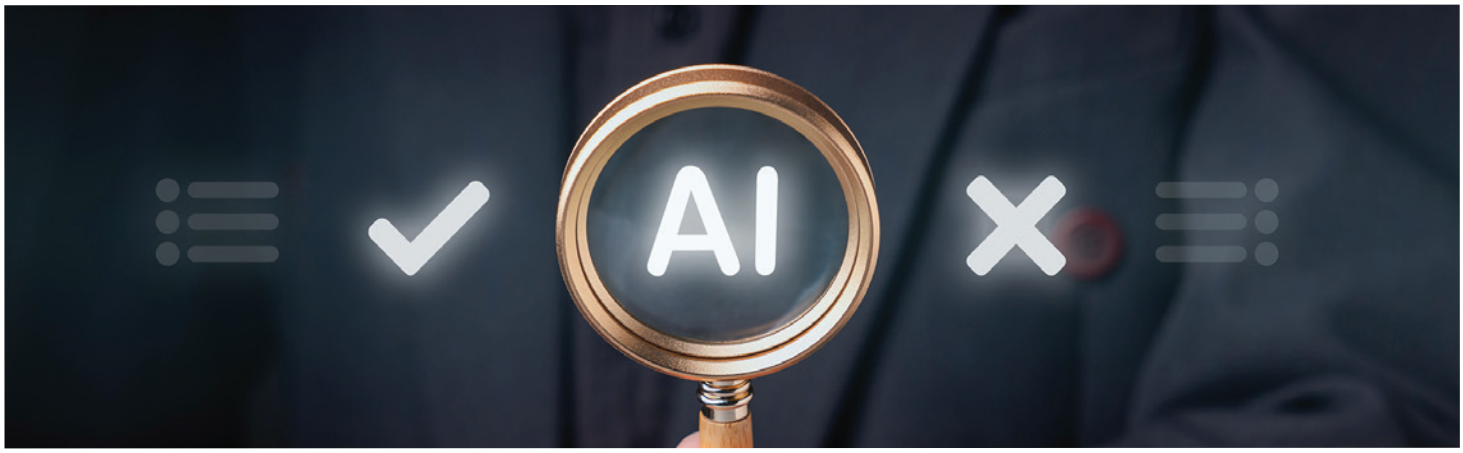
R Street
Free markets. Real solutions.

R Street Policy Study
No. 338
June 2026



AI Writers on Community Notes: An Evaluation of Seven Months of Data

By Spence Purnell

The conditions that produced Community Notes—robust First Amendment protections, federal restraint on prescriptive moderation frameworks, and Section 230 liability shielding—are themselves the most productive contribution policymakers can make to the information ecosystem.

Executive Summary

One of the central problems in content moderation policy is how best to correct misinformation at scale. Three major options for doing so include professional fact-checking institutions, platform enforcement, or crowdsourced fact-checking. X's (i.e., Twitter's) Community Notes program represents the most prominent test of the third approach: Volunteers write short, sourced corrections on posts they believe are misleading, other volunteers rate those corrections, and if enough raters with different political perspectives agree that a note is helpful, it is displayed publicly beneath the original post.

In late 2025, X introduced a new variable into this system—AI-powered accounts that scan posts and generate corrections alongside human volunteers, with those AI-written notes entering the same human-rated evaluation pipeline. This policy study provides a systematic examination of

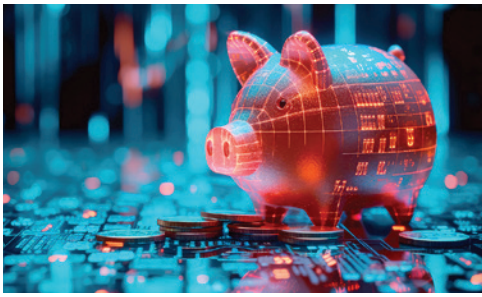
Table of Contents

The sources included in this paper were verified and active at the time of publication.

these AI note writers using seven months of publicly available Community Notes data (September 2025 through March 2026).

Our research revealed that of the 27 AI accounts now enrolled through the platform's application programming interface (API), 24 actively wrote a combined 31,464 notes, or 7.4 percent of total volume of all notes. AI notes achieved "Currently Rated Helpful" (CRH) status—the designation that causes a note to be displayed publicly beneath a post—at more than double the rate of human-written notes (18.0 percent vs. 8.9 percent, respectively). However, the vast majority of notes, AI and human alike, never receive enough ratings from volunteer reviewers to get a final determination; they remain in a "Needs More Ratings" queue indefinitely. This means that looking at CRH alone potentially understates note quality because unreviewed notes are counted as failures even though they are never evaluated. To address this issue and isolate note quality from the system's rating-capacity constraints, we introduced a new metric: the "Verdict Success Rate" (VSR), which measures how often notes succeed among those that completed the full evaluation process.

In our analysis, AI notes achieved a VSR of 88.8 percent compared to 68.5 percent for human-written notes (chi-squared $P < 0.001$). This gap between VSR and the CRH rates reveals unrealized value in the rating queue. Thus, directing rater attention toward high-VSR writers could nearly double the number of corrections that reach users. Of note, time-to-verdict did not significantly differ between AI and human notes (median 6.0 vs. 6.3 hours, Mann-Whitney $P = 0.058$). Moreover, rater feedback analysis revealed that human evaluators most commonly criticized AI notes for being unnecessary but praised their directness, sourcing, and neutral tone. Economic analysis of the rating ecosystem showed that AI notes consume rater attention roughly three times more efficiently than human notes, requiring 304 ratings per successful correction versus 908 for humans, potentially saving approximately $577,000 to $3.4 million in professional fact-checking labor.



Economic analysis of the rating ecosystem showed that AI notes consume rater attention roughly three times more efficiently than human notes, requiring 304 ratings per successful correction versus 908 for humans, potentially saving approximately $577,000 to $3.4 million in professional fact-checking labor.

1. Introduction

Community Notes, originally launched as "Birdwatch" in January 2021, is X's experiment in crowdsourced fact-checking.[1] The program invites users to write short contextual annotations on posts they believe are misleading, and a separate pool of contributors rates those annotations for helpfulness.[2] Unlike traditional fact-checking, which relies on professional newsrooms and dedicated fact-checking organizations, Community Notes distributes editorial judgment across a large body of volunteers with diverse viewpoints.[3] If enough people from different perspectives agree that a note is helpful, that consensus serves as a signal of quality and neutrality.[4]

1. Keith Koleman, "Introducing Birdwatch, a community-based approach to misinformation," X Blog, Jan. 25, 2021. https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.
2. Ibid.
3. Spence Purnell, "A Brief Review of Fact-Checking in the Digital Era," R Street Institute, March 5, 2025. https://www.rstreet.org/commentary/a-brief-review-of-fact-checking-in-the-digital-era.
4. Stefan Wojcik et al., "Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation," *arXiv preprint*, Oct. 27, 2022. https://arxiv.org/abs/2210.15723.

The program went through several formative phases. During the initial pilot (January 2021 to November 2022), X limited participation to a small group of U.S.-based users and displayed notes only within a dedicated Birdwatch site rather than on the main X feed.[5] In November 2022, X rebranded the program as Community Notes and began displaying highly rated notes directly beneath posts on the platform.[6] A phased international rollout followed throughout 2023 and 2024, eventually extending the program to users worldwide.[7]

At the heart of the system is the bridging algorithm—a matrix factorization approach (i.e., a mathematical technique for identifying hidden patterns in large datasets of user ratings) designed to surface notes that earn agreement across ideologically diverse raters.[8] Rather than relying on a simple majority vote, which could allow partisan blocs to promote favorable notes or suppress unfavorable ones, the algorithm models each rater's ideological position along a latent dimension—a hidden variable inferred from patterns in the data. It then identifies notes that receive positive ratings from raters on both sides. A note achieves "Currently Rated Helpful" (CRH) status (i.e., the designation that causes it to be displayed publicly beneath a post) only when it demonstrates this kind of cross-partisan consensus. CRH is the key outcome in the Community Notes system, as only CRH notes are visible to all users on the platform.[9]

Rather than relying on a simple majority vote, which could allow partisan blocs to promote favorable notes or suppress unfavorable ones, the algorithm models each rater's ideological position along a latent dimension—a hidden variable inferred from patterns in the data.

Notes that the algorithm has fully evaluated but that fail to achieve cross-partisan consensus receive a classification of "Currently Rated Not Helpful" (CRNH). A CRNH designation means the algorithm had enough data to render a judgment, but the community's verdict was negative, typically because the note was inaccurate, poorly sourced, or simply not needed.

Notes that have not yet accumulated enough ratings for the algorithm to render a judgment remain in "Needs More Ratings" (NMR) status.[10] Most notes end up here, as the system receives far more notes than its volunteer rater pool can evaluate. Somewhat confusingly, NMR encompasses two distinct populations: Some notes are genuinely under-reviewed, having not attracted enough raters to reach a verdict, whereas others are blocked by what the system calls **`FIRM_REJECT`**—an algorithmic gate that flags notes whose early rating patterns show sharp ideological polarization. A note flagged as **`FIRM_REJECT`** may have received dozens of ratings, and the majority may have been positive, but if that support comes predominantly from one side of the political spectrum, the algorithm prevents it from advancing to CRH. These notes remain classified as NMR rather than CRNH because the system preserves the possibility that additional raters from different ideological perspectives could still provide the cross-partisan support needed to advance the note to CRH. The distinction between NMR and CRNH matters for

5. Koleman. https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.
6. "Helpful Birdwatch Notes Are Now Visible to Everyone on Twitter in the US," X Blog, Oct. 6, 2022. https://blog.x.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us.
7. Ibid.
8. Wojcik et al. https://arxiv.org/abs/2210.15723.
9. "Community Notes," GitHub, March 2026. https://github.com/twitter/communitynotes.
10. "Evaluation," X Community Notes, last accessed April 27, 2026. https://communitynotes.x.com/guide/en/under-the-hood/guardrails.

evaluating note quality, as raw display rates conflate rating capacity constraints with genuine quality differences. This system, designed entirely around human participation and judgment, was the status quo until late 2025.[11]

In 2025, X expanded the Community Notes program to include automated participants. Developers enrolled their AI note writers through a dedicated application programming interface (API) pathway.[12] These accounts are identifiable in the public data through their enrollment state value of **`apiEarnedIn`**, distinguishing them from human contributors who earn enrollment through the standard rating-based process (i.e., **`earnedIn`**).[13] Although X has not publicly documented the criteria for API enrollment, the identities of the operators behind these accounts, or the AI models powering them, one model author has shared the details of his process online.[14]

The introduction of AI writers raises fundamental questions about the future of Community Notes.

The introduction of AI writers raises fundamental questions about the future of Community Notes. Chief among them are: (1) whether automated systems can produce notes that meet or exceed human quality standards within a system specifically designed around diverse human judgment and (2) whether AI participation reduces the timing and coverage gaps that have limited Community Notes' effectiveness or instead introduces new risks around concentration, gaming, and erosion of crowd-based legitimacy. This study addresses these questions through a comprehensive empirical examination of seven months of AI note-writing activity, drawing on publicly available data through March 2026.

2. Data and Methods

This section describes the data sources, identification methods, and statistical tools used in our analysis. We start by explaining how we identified AI note writers within the leveraged datasets and then define the two primary outcome measures and the statistical tests used for group comparisons throughout this study.

2.1 Data Sources, AI Identification, and Study Period

All data came from the official Community Notes public download data; the last relevant download for the development of this paper was March 23, 2026.[15] Our analysis used three datasets: a merged "notes-with-status-history" dataset containing 423,915 notes with creation timestamps, author identifiers, classification labels, summary text, and scoring status; a user enrollment dataset recording 1,407,713 participants' enrollment states and histories; and a ratings dataset containing 35,521,327 individual ratings for the rater feedback analysis.[16]

All data came from the official Community Notes public download data; the last relevant download for the development of this paper was March 23, 2026.

11. Ibid.
12. Community Notes (@CommunityNotes), "Introducing AI Note Writer API AI helping humans. Humans still in charge […]," July 1, 2025, 3:35 PM. [tweet] https://x.com/CommunityNotes/status/1940132205486915917.
13. "Community Notes." https://github.com/twitter/communitynotes.
14. Nathan Young, "World's First AI Community Note," Predictive Text, Sept. 23, 2025. https://nathanpmyoung.substack.com/p/worlds-first-ai-community-note#footnote-anchor-2-174094438.
15. "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.
16. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

We identified AI writers by filtering the enrollment dataset for accounts whose **`enrollmentState`** value was **`apiEarnedIn`**, meaning the account was active and approved for AI note writing. This yielded 27 accounts, of which 24 wrote at least one note.[17] Our analysis covered September 2025 through March 2026, the seven-month window in which AI note-writing activity appeared in the data.

2.2 Statistical Methods

We used two primary outcome measures throughout this analysis. The "shown rate" captures the proportion of all notes written that achieve "Currently Rated Helpful" (CRH) status—the designation that causes a note to be displayed publicly beneath a post—(i.e., CRH divided by total number of notes), reflecting the end-to-end probability that a given note reaches users. We also introduced a new metric, the "Verdict Success Rate" (VSR), as an outcome measure. This new metric captures the proportion of notes that survived the full evaluation pipeline and received a favorable outcome. Specifically, it measures:

$$\frac{\text{CRH}}{\text{CRH} + \text{CRNH}}$$

This restricts the denominator to "verdicted" notes—that is, those that received either CRH or CRNH status. A verdict means the algorithm gathered enough rating data to make a final determination and the **`FIRM_REJECT`** polarization filter did not block the note.

We used the VSR as a quality signal because it isolates the algorithm's quality assessment from two confounding factors: the rating-capacity bottleneck (notes stuck in NMR because too few raters reviewed them) and the polarization filter (notes blocked by **`FIRM_REJECT`** because their support was ideologically one-sided rather than bridged). A note stuck in NMR may be excellent but under-reviewed, or it may be well-liked by a majority of raters but blocked because that support came from only one side of the ideological spectrum. Importantly, this does not mean that those who disagreed with the note rated it unfavorably, but rather that only raters from one end of the ideological spectrum reviewed it. The VSR measures quality only among notes that completed the entire evaluation process.

The group comparisons we conducted for our analysis employed two standard, non-parametric tests. The first was a chi-squared test of independence (a standard method for determining whether two categorical variables are related), which was used to assess whether the distribution of verdict outcomes differs significantly between AI and human writers. The second was a Mann-Whitney U test (two-sided alternative hypothesis in this case), which compares two groups without assuming their data follows a bell curve and is used for time-to-verdict distributions—an appropriate choice given the heavy right tail typical of waiting-time data. Time-to-verdict was calculated as the difference between **`timestampMillisOfFirstNonNMRStatus`** and **`createdAtMillis`**, restricted to non-negative values to exclude data anomalies.

We used two primary outcome measures throughout this analysis. The "shown rate" captures the proportion of all notes written that achieve "Currently Rated Helpful" (CRH) status. We also introduced a new metric, the "Verdict Success Rate" (VSR), as an outcome measure.

17. Ibid.

3. Results and Data

The analysis that follows begins by profiling the AI note writers themselves, including their scale, growth, content focus, and sourcing practices, before turning to a direct comparison of AI and human notes across key outcome measures. It then examines the algorithmic barriers that prevent notes from reaching users and quantifies the unrealized value sitting in the rating queue.

Before comparing AI and human note performance directly, it is useful to understand who the AI note writers are and what they produce.

3.1 AI Note Writers: Scale, Content, and Performance

Before comparing AI and human note performance directly, it is useful to understand who the AI note writers are and what they produce. This section examines the scale of AI participation, its growth trajectory over the study period, the types of content AI writers target, their sourcing practices, and variation in performance across individual accounts.

3.1.1 Scale of AI Participation

Table 1 presents a summary of AI participation in Community Notes during the study period. Of the 27 enrolled AI accounts, 24 actively contributed notes, producing a combined 31,464 notes out of 423,915 total (AI and human produced). Notably, AI-generated notes accounted for 13.9 percent of all CRH notes—nearly double their 7.4 percent share of total volume—indicating that AI notes were displayed to users at a higher rate relative to their output.

Table 1: Summary of AI Participation in Community Notes, September 2025 to March 2026

Metric	Value
Total AI accounts enrolled	27
Active AI writers	24
AI notes written	31,464
Human notes written	392,451
AI share of total notes	7.4%
AI share of CRH notes	13.9%

AI, artificial intelligence; CRH, Currently Rated Helpful.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

3.1.2 Growth Trajectory

AI note-writing activity grew rapidly over the study period, as shown in Table 2. Output began modestly in September 2025 with 93 notes and peaked at 8,109 in February 2026—a roughly 87-fold increase.[18] Over the full study period, AI's share of total monthly note volume rose from 0.2 percent (September 2025) to 12.7 percent (March 2026). Human note volume remained relatively stable, ranging from approximately 54,000 to 70,000 notes per month.[19] Taken together, these figures suggest that AI growth represented genuine additive note-writing capacity; it did not appear to be displacing human writers.

18. Ibid.
19. Ibid.

Table 2: Monthly AI Note Volume, Share, and CRH Rate, September 2025 to March 2026

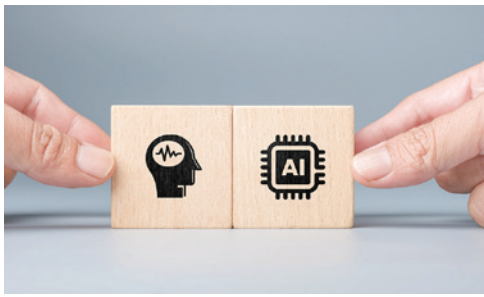
Month	No. of AI Notes	No. of Human Notes	Total Notes	AI Share (%)	AI CRH Rate (%)
Sept. 2025	93	59,880	59,973	0.2	11.8
Oct. 2025	1,656	62,373	64,029	2.6	18.5
Nov. 2025	2,376	62,378	64,754	3.7	17.5
Dec. 2025	3,688	61,996	65,684	5.6	20.2
Jan. 2026	7,641	69,669	77,310	9.9	19.5
Feb. 2026	8,109	69,401	77,510	10.5	19.4
Mar. 2026	7,901	54,433	62,334	12.7	14.4

AI, artificial intelligence; CRH, Currently Rated Helpful.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

The CRH rate held roughly in the high teens to low 20s (17.5-20.2 percent) from October 2025 through February 2026.[20] The March 2026 dip to 14.4 percent likely reflects the recency of those notes, as many had not yet accumulated enough ratings for a verdict.[21]

3.1.3 Content Profile

When a contributor writes a Community Note, they must classify the target post into one of two categories: (1) "misinformed or potentially misleading" for posts believed to contain inaccurate, incomplete, or deceptive information or (2) "not misleading" for posts that are believed to be accurate but could benefit from additional context. These classifications reflect the note writer's assessment of the post they are annotating and are distinct from the community's subsequent verdict on the note itself.[22]



Our analysis of classification labels reveals a distinctive content profile for AI writers: 100 percent of AI notes classified the target post as misleading. This is compared with 82.4 percent for human notes.

Our analysis of classification labels reveals a distinctive content profile for AI writers: 100 percent of AI notes classified the target post as misleading.[23] This is compared with 82.4 percent for human notes. This discrepancy is likely the result of an architectural constraint: AI writers can only submit a note on posts they classify as "misleading"; the API provides no pathway for AI writers to flag "not misleading" posts. However, as we show later, this does not mean that every AI note written is helpful or needed, only that AI note writers are not technically configured to contribute to posts tagged as "not misleading." AI notes also showed a notable emphasis on certain subtypes within the "misinformed or potentially misleading" category; they tagged 80.3 percent of their notes as addressing "missing important context," compared to 59.5 percent of notes tagged by human writers (Table 3).[24] AI note writers also flagged several additional subtypes at higher rates than human note writers, including "factual error" (67.9 percent vs. 58.1 percent), "unverified claim as fact" (44.1 percent vs. 34.3 percent), and "manipulated media" (39.4 percent vs. 13.3 percent). The manipulated media gap was the single largest divergence, with AI writers being nearly three times more likely than human writers to flag a post for containing altered or synthetic media.

20. Ibid.
21. Ibid.
22. "Community Notes." https://github.com/twitter/communitynotes.
23. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.
24. Ibid.

Table 3: Differences in Subtype Classification of Posts by AI and Human Note Writers

"Misleading" Subtype	AI Rate (%)	Human Rate (%)	Difference (%)
Missing important context	80.3	59.5	+20.8
Factual error	67.9	58.1	+9.8
Unverified claim as fact	44.1	34.3	+9.8
Manipulated media	39.4	13.3	+26.1
Outdated information	9.4	23.9	-14.5
Satire	2.8	8.6	-5.8
Other	0.2	14.8	-14.6

AI, artificial intelligence.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

3.1.4 Source Citations and Topic Distribution

We also analyzed the "summary" fields included in the raw data files, which contained any URLs cited as sources, as well as the full text of each AI-written note, allowing us to gather additional insights from both.[25]

With regard to the sourcing of note information, we found that AI notes cited URL sources to support assertions at a notably higher rate and density than human notes. Specifically, 100 percent of AI notes included at least one URL source, compared to 87.8 percent of human notes, and the average number of URL sources per AI note was 2.2, compared to 1.4 for human notes.[26] AI notes also tended to be longer than human notes, with a median summary length of 363 characters versus 245.[27]

The URLs offered as sources also differed substantially between AI and human writers. Table 4 shows the top-10 most-cited domains for each group as well as each domain's share of total citations (approximately 67,700 AI-provided URLs vs. 530,900 human-provided URLs).[28]

With regard to the sourcing of note information, we found that AI notes cited URL sources to support assertions at a notably higher rate and density than human notes. Specifically, 100 percent of AI notes included at least one URL source.

Table 4: Most Frequently Cited Domains by AI and Human Note Writers

Rank	AI Top Domains	No. of Citations	AI (%)	Human Top Domains	No. of Citations	Human (%)
1	x.com	5,394	8.0	x.com	12,274	2.3
2	en.wikipedia.org	3,880	5.7	en.wikipedia.org	1,912	0.4
3	reuters.com	2,431	3.6	web3antivirus.io*	1,697	0.3
4	youtube.com	1,998	3.0	en.m.wikipedia.org	1,104	0.2
5	instagram.com	1,407	2.1	youtu.be	924	0.2
6	bbc.com	1,331	2.0	help.x.com	914	0.2
7	snopes.com	1,234	1.8	youtube.com	870	0.2
8	aljazeera.com	1,020	1.5	business.x.com	862	0.2
9	nytimes.com	930	1.4	instagram.com	687	0.1
10	foxnews.com	929	1.4	t.co	686	0.1

AI, artificial intelligence.
*Web3 Antivirus is a crypto scam-detection browser extension. Its prevalence in human Community Notes reflects the program's heavy use for flagging cryptocurrency and Web3 spam posts—a category that AI writers largely do not target.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

25. Ibid.
26. Ibid.
27. Ibid.
28. Ibid.

AI citations typically offered URL sources from authoritative journalism and fact-checking outlets like *Reuters*, BBC, Snopes, AP News, AFP Fact Check, and PolitiFact.[29] This shows the interdependence between AI and human writers: functioning fact-checking depends on both. While AI writers produce more corrections at scale, they still draw upon human-written and -researched sources to ground their analysis.

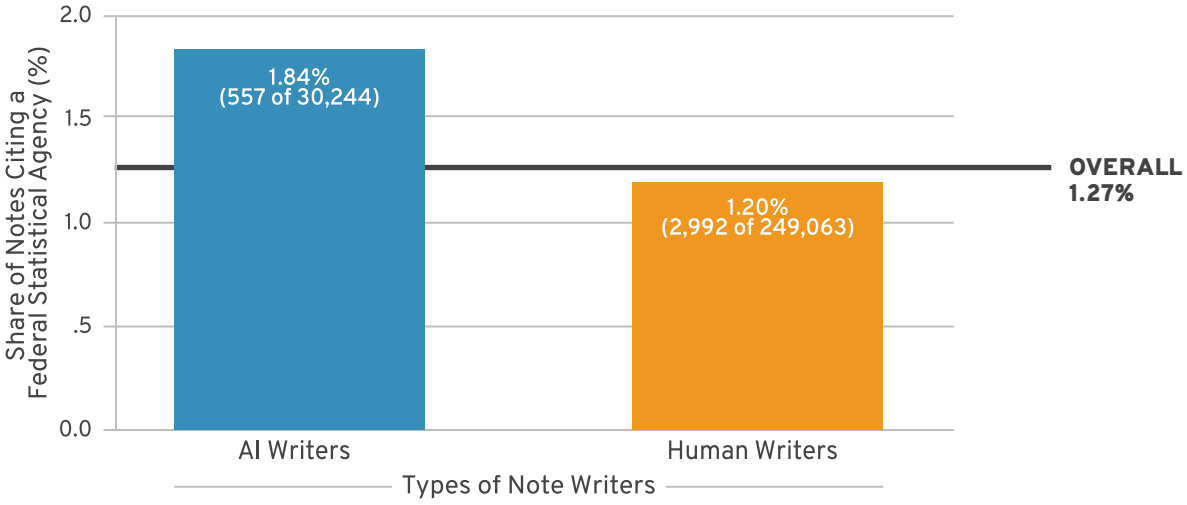
Human notes were dominated by platform self-references (x.com, help.x.com, and business.x.com together accounted for three of the top-eight human domains) as well as by web3antivirus.io, a crypto scam-detection service whose prevalence reflects Community Notes' heavy use for flagging crypto and Web3 spam.[30] AI notes also frequently cited Al Jazeera (1,020), *Times of Israel* (526), and *Hindustan Times* (448).[31]

Another important finding was that U.S. statistical agencies were prominent sources for both human and AI writers.[32] When we filtered to notes attempting to correct a factual claim, these agencies were directly cited more than 3,500 times—and at a higher rate by AI writers than by humans (Figure 1).[33]

KEY TAKEAWAY

An important finding was that U.S. statistical agencies were prominent sources for both human and AI writers.

Figure 1: U.S. Statistical Agency Citations Among Notes Contesting Factual Claims, AI vs. Human Notes Writers

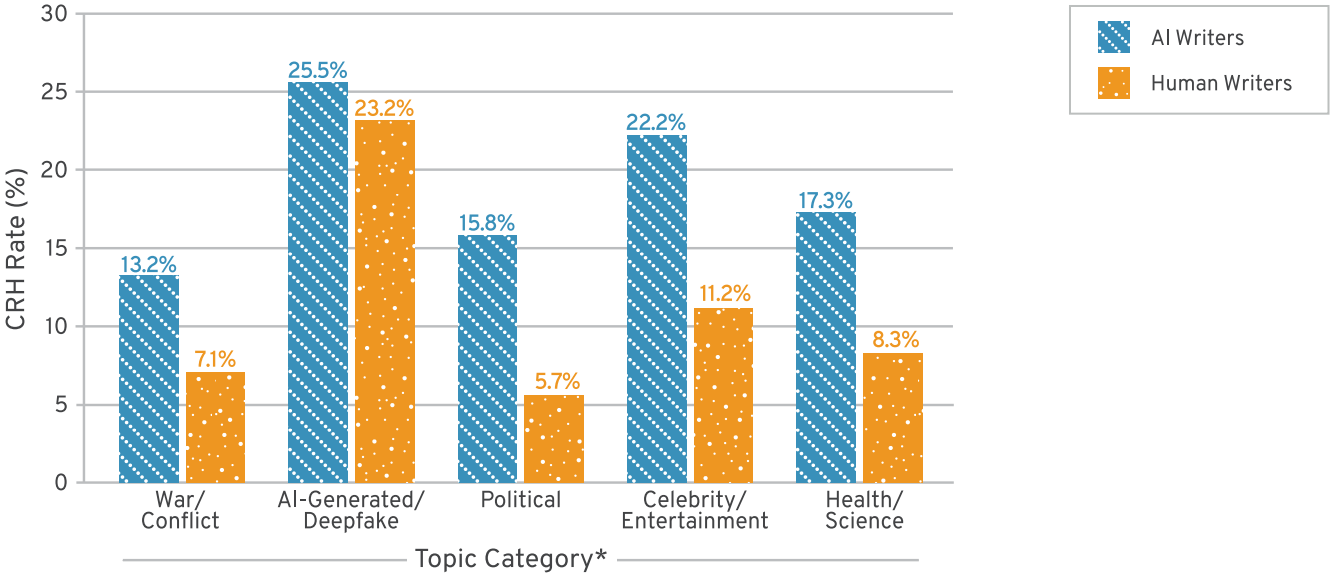


AI, artificial intelligence.
Source: "R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

We also conducted a keyword-based analysis to examine how AI and human writers allocate attention across subject areas. Our findings for key topic classification groups are shown in Figure 2.

29. Ibid.
30. Ibid.
31. Ibid.
32. Ibid.; Nicolas Bloom et al., "The Value of Reliable Statistics," National Bureau of Economic Research, April 2026. https://www.nber.org/papers/w35135.
33. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

Figure 2: Topic Distribution and CRH Rates, AI vs. Human Notes Writers

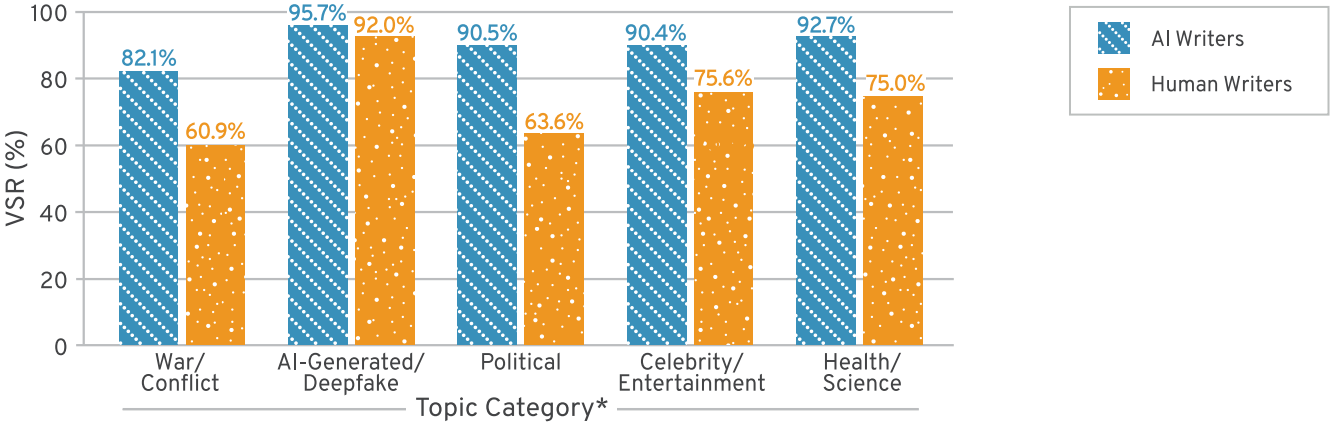


AI, artificial intelligence; CRH, Currently Rated Helpful.
*Topic classification uses regex pattern matching on the lowercased full text of each note (the summary field). Categories are non-exclusive: A single note can match multiple topics simultaneously. Because categories overlap, note counts in the topic distribution table sum to more than the total number of AI notes.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

AI notes achieved higher CRH rates than human notes in every main topic category.[34] The gap is particularly large for political notes (15.8 percent AI-written vs. 5.7 percent human-written), celebrity/entertainment notes (22.2 percent vs. 11.2 percent), and satire/joke notes (25.9 percent vs. 6.6 percent).[35]

This pattern also held when conducting topical analysis along our calculated VSRs, as AI outperformed humans across every topical category (Figure 3). The difference was most pronounced for satire/jokes, war and conflict, and political topics.

Figure 3: Topic Distribution and VSR, AI vs. Human Notes



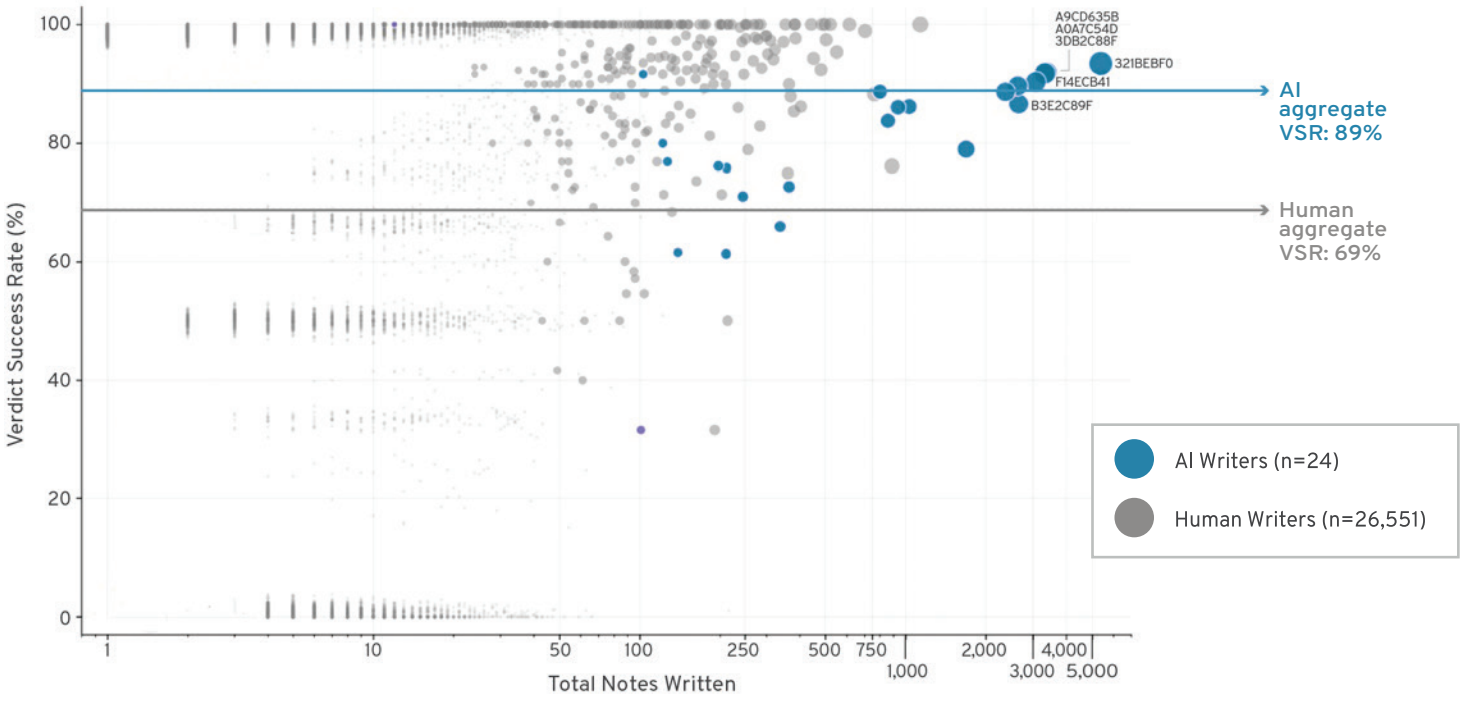
AI, artificial intelligence; VSR, verdict success rate.
*Topic classification uses regex pattern matching on the lowercased full text of each note (the summary field). Categories are non-exclusive: A single note can match multiple topics simultaneously. Because categories overlap, note counts in the topic distribution table sum to more than the total number of AI notes.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

34. Ibid.
35. Ibid.

3.1.5 Individual Writer Performance: AI vs. Human

Performance varies substantially across the 24 active AI note writers and human writers. Figure 4 shows each writer's output volume and VSR, where the size of each bubble is proportional to the total number of notes written.[36]

Figure 4: Performance of Individual AI Note Writers Writer Performance, AI vs. Human*



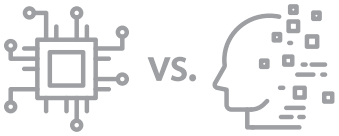
AI, artificial intelligence; VSR, verdict success rate.
*Bubble size = total number of notes written
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

Performance among the top writers was notably consistent. The seven highest-volume AI writers, each producing more than 2,000 notes, all maintained VSRs above 86 percent.[37] At the other end of the spectrum, two AI accounts—43D4A2A7 (VSR of 31.6 percent across 101 notes) and BF87E4D4 (46.2 percent across 217 notes)—fell well below the human VSR baseline of 68.5 percent.[38] These are the only AI writers whose verdict success rates underperformed the average human writer.

For human writers, achieving volume was difficult, with only one account surpassing 1,000 notes.[39] However, many of the human notes were high quality, creating a strong cluster of human writers at 100 percent VSR. The distribution also included smaller clusters near 0 percent and around 50 percent. AI distribution was more concentrated, centered on a cluster of writers who were both high-quality and high-volume. Although some individual human writers have high VSRs (indicating more consistent quality), they are not able to achieve volume at those rates in the same way as some AI writers.

36. Ibid.
37. Ibid.
38. Ibid.
39. Ibid.

3.2 AI vs. Human Note-Writing Performance



AI notes received more favorable outcomes than human notes across every status category.

Having established the profile and practices of AI note writers, we now turn to a direct comparison of outcomes between AI and human contributors.

AI notes received more favorable outcomes than human notes across every status category (Table 5).[40] The AI CRH rate of 18.0 percent is more than double the human rate of 8.9 percent, meaning AI notes are more than twice as likely to be displayed to users. As shown in Section 3.1.4, this pattern held across a variety of topic areas, including more politically divisive topics like war, misinformation, and politics, demonstrating the robustness of the finding.[41]

The community also actively rejected AI notes less often. Only 2.3 percent of AI-written notes received CRNH status compared to 4.1 percent of human notes.[42] The NMR rate was 71.5 percent for AI vs. 79.5 percent for humans. A small share of notes (8.2 percent for AI, 7.4 percent for humans) carried no status designation, reflecting recently written notes that had not yet entered the scoring pipeline.[43]

Table 5: Status Outcomes for AI and Human Notes

Writer Type	Total No. Notes	No. of CRH	No. of CRNH	No. of NMR	No. of No Status	CRH Rate (%)	CRNH Rate (%)
AI	31,464	5,673	718	22,499	2,574	18.0	2.3
Human	392,451	35,118	16,130	312,137	29,066	8.9	4.1

AI, artificial intelligence; CRH, Currently Rated Helpful; CRNH, Currently Rated Not Helpful; NMR, Needs More Ratings.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

Taken together, these results indicate that AI notes were not only more likely to be displayed to users but also less likely to be rejected by the community or left without a determination. The consistency of this advantage across every status category suggests that AI notes are systematically better aligned with the criteria the bridging algorithm uses to identify high-quality corrections.

3.2.1 Verdict Success Rates (VSRs)

As defined in Section 2.2, the VSR isolates note quality from rating-capacity constraints by measuring success only among notes that completed the full evaluation process. Notes stuck in NMR—whether due to insufficient ratings or `FIRM_REJECT`—were excluded from this analysis because they never received a final quality judgment.

Of the 6,391 AI notes that reached a verdict, 5,673 (88.8 percent) received helpful ratings (Table 6).[44] Among 51,248 verdicted human notes, 35,118 (68.5 percent) received helpful ratings.[45] The 20.3 percentage point gap is statistically significant (chi-squared = 1,124.3, $P < 0.001$).

40. Ibid.
41. Ibid.
42. Ibid.
43. Ibid.
44. Ibid.
45. Ibid.

Table 6: Verdict Success Rates for AI and Human Notes

Writer Type	No. of Verdicted Notes	No. CRH	VSR (%)
AI	6,391	5,673	88.8
Human	51,248	35,118	68.5

AI, artificial intelligence; CRH, Currently Rated Helpful; VSR, verdict success rate.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

The AI verdict success advantage persisted in every month of the study period. Seven months of data and more than 6,000 verdicted AI notes provide a substantial basis for this comparison.

The AI verdict success advantage persisted in every month of the study period. Seven months of data and more than 6,000 verdicted AI notes provide a substantial basis for this comparison.

3.2.2 Time-to-Verdict

A key question we sought to explore in this study was whether AI-written notes move through the evaluation pipeline faster than human notes. Table 7 summarizes some of the key metrics we considered for this assessment.[46]

Table 7: Time-to-Verdict for AI and Human Notes

Metric	AI Notes	Human Notes
Median time-to-verdict (hours)	6.0	6.3
Mean time-to-verdict (hours)	22.5	20.4
Interquartile range (25th to 75th percentile, hours)	2.9–15.1	2.8–14.8
90th percentile (hours)	~45	~45
No. completed in less than 2 hours	941 (12.7%)	10,569 (16.3%)
N (verdicted)	7,423	64,971

AI, artificial intelligence.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

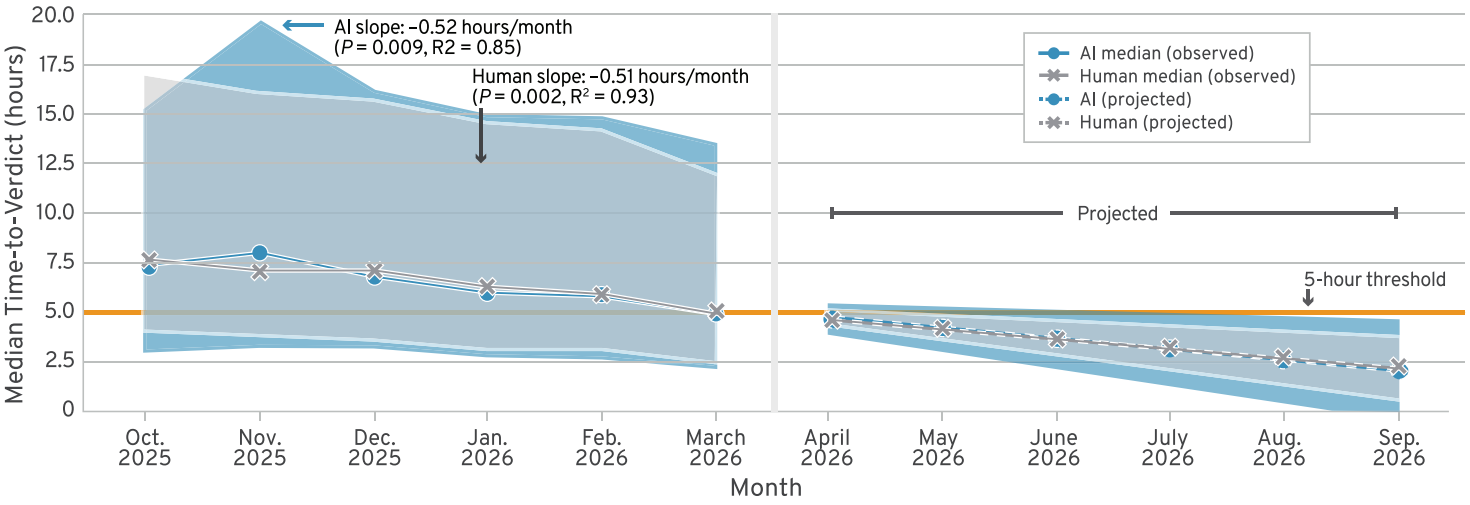
The median time from note creation to first non-NMR status was 6.0 hours for AI notes versus 6.3 hours for humans, though this difference was not statistically significant (U = 244,371,795, $P = 0.058$).[47] The large gap between the median (roughly 6 hours) and the mean (roughly 21 to 22 hours) in both groups reflects a heavily right-skewed distribution. Most notes reached a verdict relatively quickly: Approximately 75 percent received a verdict within about 15 hours.[48] But a long tail of slow notes pulled the mean upward substantially. The 90th percentile sat at approximately 45 hours, meaning 10 percent of verdicted notes took nearly two full days to receive a determination.

We also looked at monthly timing trends during our analysis. We computed median time-to-verdict separately for AI and human notes and fit ordinary least squares (OLS) trend lines—a standard statistical method for fitting a straight line to data—to both series (Figure 5). In the Community Notes program's earliest months, AI notes took longer to receive verdicts than human notes, with AI median time-to-verdict starting above the human median.[49] Over the study period, however, AI-written notes' time-to-verdict

46. Ibid.
47. Ibid.
48. Ibid.
49. Ibid.

declined steadily, while human-written notes' time-to-verdict declined more slowly. By early 2026, the two series had converged to near-parity, at nearly a 5.0-hour median, a meaningful improvement from the 7.5 hours measured at the start of the study period.[50] The OLS trend lines confirmed this pattern: The AI slope was negative and steeper than the human slope, consistent with the rating community adapting to evaluating AI-generated notes. If current trends continue, AI and human notes could converge to median processing times below 5 hours within the next six months.

Figure 5: Monthly Time-to-Verdict Trend, AI vs. Human Notes



AI, artificial intelligence.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

3.2.3 Differences in Rater Feedback

Ratings data provides a direct window into how human raters evaluate AI notes compared to human-written ones. Table 8 shows that AI notes received a modestly more favorable rating distribution than human notes across all three rating categories (i.e., helpful, somewhat helpful, not helpful).[51] However, a deeper analysis of why raters say notes are helpful or not shows why AI writers maintain some advantages over humans.

Table 8: Rating Distribution for AI and Human Notes*

Rating Category	AI Notes (%)	Human Notes (%)
Helpful	66.0	62.8
Somewhat helpful	2.3	1.9
Not helpful	31.8	35.3

AI, artificial intelligence.
*For AI notes, N = 1,724,015 ratings; for human notes, N = 31,895,441 ratings.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026.
https://x.com/i/communitynotes/download-data.

When raters marked an AI note as helpful, their top reasons were: "addresses the claim directly" (71.8 percent), "provides important context" (67.7 percent), "clearly written" (66.7 percent), "cites good sources" (59.2 percent), and "uses

50. Ibid.
51. Ibid.

unbiased language" (49.2 percent; Table 9).[52] AI notes scored modestly higher than human notes on "addresses the claim" (+6.4 percentage points) and "good sources" (+5.6 percentage points), which could be a reflection of the higher levels of sourcing and longer character counts highlighted in Section 3.1.4.[53] Helpfully rated notes show no major differences between human and AI notes; however, notes rated as not helpful show a much clearer pattern.

Table 9: Reasons for "Helpful" Ratings on AI and Human Notes

"Helpful" Reason	AI Rate (%)	Human Rate (%)	Difference (%)
Addresses the claim	71.8	65.4	+6.4
Important context	67.7	67.1	+0.6
Clearly written	66.7	67.9	-1.2
Good sources	59.2	53.6	+5.6
Unbiased language	49.2	47.6	+1.6

AI, artificial intelligence.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

When Community Notes raters mark a note as "not helpful," they also submit reasons for their designation. The most common reason for a "not helpful" designation for AI notes in the reviewed dataset was "note not needed" (50.6 percent), which exceeded the human note rate (42.4 percent) by 8.2 percentage points (Table 10).[54] Notably, AI notes were significantly less likely than human notes to be marked "not helpful" on the grounds of "opinion/speculation" and "argumentative/biased" (10.8 percent and 12.1 percent lower than human rates, respectively).[55] This could point to an overall less emotional, more fact-based grounding for AI-written notes.

Table 10: Reasons for "Not Helpful" Ratings on AI and Human Notes

"Not-Helpful" Reason	AI Rate (%)	Human Rate (%)	Difference (%)
Note not needed	50.6	42.4	+8.2
Missing key points	46.3	49.3	-3.0
Incorrect	31.9	31.6	+0.3
Opinion/speculation	28.4	39.2	-10.8
Sources missing	25.3	25.5	-0.2
Argumentative/biased	21.2	33.3	-12.1

AI, artificial intelligence.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

Taken as a whole, rater feedback suggests that a key strength of AI notes is likely precision (addressing claims directly with good sources and neutral language), whereas a key weakness appears to be judgment about whether a correction is needed in the first place. This pattern will be explored further in Section 5.2.

KEY TAKEAWAY

AI notes were significantly less likely than human notes to be marked "not helpful" on the grounds of "opinion/speculation" and "argumentative/biased" (10.8 percent and 12.1 percent lower than human rates, respectively).

52. Ibid.
53. Ibid.
54. Ibid.
55. Ibid.

3.3 Unrealized Value in the Rating Queue

The gap between VSR and shown rate quantifies the scale of unrealized value in the current system. To assess this metric, we decomposed each AI writer's NMR notes into two groups: notes blocked by the **FIRM_REJECT** polarization filter, which additional ratings could not rescue, and notes that are genuinely under-reviewed Across all 24 active AI writers, 7,040 notes fell into the genuinely under-reviewed category.[56]

Applying each writer's observed VSR to their under-reviewed notes produces a conservative projection of how many additional CRH notes the system would generate if those notes received sufficient ratings, as shown in Table 11.[57]

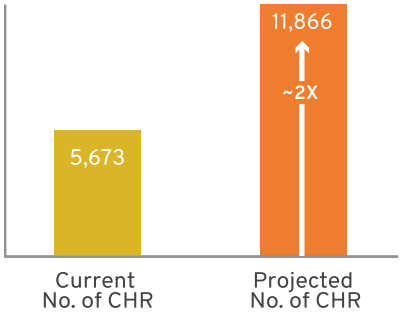
Applying each writer's observed VSR to their under-reviewed notes produces a conservative projection of how many additional CRH notes the system would generate if those notes received sufficient ratings.

Table 11: Projected CRH Yield from Under-Reviewed AI Notes

AI Writer (ID Prefix)	No. of Notes	No. Current CRH	Shown Rate (%)	VSR (%)	No. Under-Reviewed	Projected Additional CRH	Projected Shown Rate (%)
321BEBF0	5,962	1,401	23.5	93.5	1,271	1,188	43.4
3DB2C88F	3,374	697	20.7	92.0	771	709	41.7
A9CD635B	3,326	706	21.2	91.8	744	683	41.8
F14ECB41	3,075	552	18.0	90.3	708	640	38.8
B3E2C89F	2,652	461	17.4	86.7	614	532	37.4
A0A7C54D	2,633	444	16.9	89.7	628	563	38.2
D6D5E740	2,366	400	16.9	88.7	555	492	37.7
All AI Writers	**31,464**	**5,673**	**18.0**	**88.8**	**7,040**	**6,193**	**37.7**

AI, artificial intelligence; CRH, Currently Rated Helpful; VSR, verdict success rate.
Source: "R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

If raters could evaluate all 7,040 under-reviewed AI notes, the projected yield is approximately 6,193 additional CRH notes, nearly doubling the current AI CRH total from 5,673 to 11,866 and lifting the aggregate shown rate for AI notes from 18.0 percent to 37.7 percent. This projection is based on each writer's historical VSR, which already accounts for notes that the community finds unnecessary or poorly targeted.



Although this analysis focuses on AI writers, whose VSRs are individually trackable and consistently high, the same mathematical logic applies to human note writers. Any writer, human or AI, whose VSR substantially exceeds their shown rate has under-reviewed notes that would likely succeed if they received sufficient rater attention. This points to an opportunity to improve Community Notes' effectiveness by directing the system's scarce rater attention more efficiently.

4. Economic Evaluation of AI Note Writing

Previous sections established that AI notes outperform human notes on quality metrics, which indicates they may consume rater attention more efficiently. This section translates those findings into economic terms—estimating the

56. Ibid.
57. Ibid.

labor value AI writers contribute, the rater resources they consume, the cost of over-correction, and their net economic contribution to the Community Notes system.

According to data from the Editorial Freelancers Association, the average hourly rate of professional editing (fact-checking is not a discretely recorded data type) is $55.11 per hour.[58] Although there are no reliable data for estimating the average length of time it takes to fact-check a community note, studies and professional associations report that individual checks can take anywhere from hours to days, and sometimes weeks.[59] Without a clear metric to use in our analysis, we opted to set a range of estimates (from 20 minutes to 2 hours per note) at the rate of $55.11 per hour, using 1 hour as the central average, as shown in Table 12.

Table 12: Estimated Labor Value Substituted by AI Note Writers

Scenario (Time per Check)	Substituted Labor Value
Low (20 min/check)	$577,994
Mid (1 hr/check)	$1,733,981
High (2 hrs/check)	$3,467,962

Source: R Street Institute analysis, 2026.

AI notes accounted for approximately 1.72 million ratings during the study period, roughly 5.1 percent of all ratings, translating to an estimated 43,100 hours of volunteer rater time at 1.5 minutes per rating.[60] Each rating requires reading the original post, evaluating the note's claim and sources, rendering a helpfulness judgment, and tagging specific reasons for the assessment. We conservatively estimated this labor at $15/hour, roughly double the federal minimum wage of $7.25. Using these assumptions, the total estimated value of rater time spent evaluating AI notes is $646,500, a meaningful draw on the system's scarce evaluation capacity.

Importantly, our research demonstrated that AI notes use rating capacity more efficiently than human notes. Perhaps the most revealing metric of our analysis was ratings per successful correction—that is, the total ratings a group consumes divided by the number of notes that ultimately achieve CRH status. As reported in Table 13, we found that AI notes required approximately 304 ratings per successful correction, compared to 908 ratings for human notes. Thus, each rating spent on an AI note was roughly three times more likely to contribute to a correction that actually reached users.

Importantly, our research demonstrated that AI notes use rating capacity more efficiently than human notes. Perhaps the most revealing metric of our analysis was ratings per successful correction.

Table 13: Rating Demand and Efficiency for AI and Human Notes

Metric	AI Writers	Human Writers
Total ratings consumed	1,724,015	31,895,441
CRH notes produced	5,673	35,118

58. "Editorial Rates," Editorial Freelancers Association, last updated March 29, 2026. https://www.the-efa.org/rates.
59. Nils Hanson, "A Guide to Fact-Checking Investigative Stories," Global Investigative Journalism Network, Nov. 3, 2021. https://gijn.org/resource/guide-to-fact-checking-investigative-stories; Greta Warren et al., "Show Me the Work: Fact Checkers' Requirements for Explainable Automated Fact-Checking," Conference on Human Factors in Computing Systems, April 2025. https://arxiv.org/html/2502.09083v1#S1; Nicolas Micallef et al., "True or False. Studying the Work Practices of Professional Fact-Checkers," *Proceedings on the ACM on Human Computer Interaction* 6:CSCW1 (April 7, 2022), pp. 1-44. https://dl.acm.org/doi/10.1145/3512974.
60. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

Metric	AI Writers	Human Writers
Ratings per successful correction	304	908
Mean ratings per note	54.8	81.3
Ratings on CRNH notes	40,642	1,099,850
Estimated rater hours consumed	43,100	797,400
Estimated rater hours on CRNH notes	1,016	27,496
Estimated economic value of rater time ($15/hr)	$646,500	$11,961,000

AI, artificial intelligence; CRH, Currently Rated Helpful; CRNH, Currently Rated Not Helpful; VSR, verdict success rate.
Source: R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.

This efficiency advantage stems from two factors. First, AI notes reach CRH at a higher rate (18.0 percent vs. 8.9 percent), so a larger share of the ratings they attract contribute to a successful outcome. Second, AI notes receive fewer wasted ratings on notes that are ultimately rejected: an estimated 1,016 rater-hours on AI CRNH notes versus 27,496 rater-hours on human CRNH notes.

The efficiency picture is not uniformly positive, however. Our rater feedback analysis identified "note not needed" as the most common reason an AI note would be flagged as "not helpful," and this rating data allowed us to quantify its cost. Across all "not helpful" ratings on AI notes, 277,282 individual ratings flagged the note as unnecessary, meaning raters spent time reading a post, evaluating an AI correction, and concluding that the note was not needed.[61] At 1.5 minutes per rating, that represents approximately 6,900 hours of volunteer effort spent on notes that raters considered unwarranted, valued at roughly $103,500 at $15 per hour.

fx

The only genuine cost to subtract is the over-correction waste: approximately $103,500 in rater time (6,900 hours at $15 per hour) spent on notes that evaluators considered unnecessary.

Given these findings on the value and costs associated with AI note ratings, we can calculate the net economic contribution AI-written notes might offer (Table 14). AI note writers contributed the equivalent of anywhere from $577,000 to roughly $3.4 million in professional fact-checking labor over seven months. Rater time spent evaluating AI notes that ultimately achieve CRH status is not a cost, but a productive investment, as those ratings directly enabled corrections that reached users. The only genuine cost to subtract is the over-correction waste: approximately $103,500 in rater time (6,900 hours at $15 per hour) spent on notes that evaluators considered unnecessary. The net economic contribution can therefore be expressed as:

$$(\text{WRITING LABOR SUBSTITUTED}) - (\text{OVER-CORRECTION COST}) = \text{NET LABOR VALUE}$$

Table 14: Net Economic Contribution of AI Note Writers

Scenario	Net Value Calculation (Labor – Over-Correction)
Low end	$577,000 – $103,500 = $473,500
Mid-range	$1,733,981 – $103,500 = $1,630,481
High end	$3,467,962 – $103,500 = $3,364,462

Source: R Street Institute analysis, 2026.

61. Ibid.

It is important to note that these estimates reflect current conditions, in which only 18.0 percent of AI notes achieve CRH status.[62] The VSR-CRH gap analysis in Section 3.3 demonstrated that 7,040 AI notes were genuinely under-reviewed and remained in the rating queue not because they were low quality but because raters had not yet evaluated them. If the Community Notes rating queue were to implement priority routing and direct rater attention toward notes from high-VSR writers, the economic equation could shift in two meaningful ways. First, the writing labor value realized through displayed corrections would roughly double, given that more of the fact-checking labor would actually reach users. Second, the over-correction cost would decline because routing raters toward proven writers would mean fewer ratings spent on notes the community would likely consider unwarranted. This means that both writer and rater time would be used more efficiently and that more corrections would reach users.

5. Discussion

The empirical findings presented in Sections 3 and 4 paint a consistent picture: AI notes outperform human notes on quality, efficiency, and cost-effectiveness. But these aggregate findings carry important nuances about the durability of the quality advantage, the costs of over-correction, the risks of growing concentration, and the structural constraints that limit how many corrections reach users. This section examines each.

5.1 The Quality Advantage

Our analysis demonstrated that AI notes achieve substantially higher CRH rates and VSRs than human notes, and this advantage is large in magnitude, durable over time, and consistent across topic areas. An 88.8 percent vs. 68.5 percent VSR represents a statistically significant 20.3 percentage point gap, which was sustained across seven months, 24 active accounts, and more than 6,000 verdicted notes.

AI writers appear specifically optimized—whether by design or by emergent capability—for the bridging algorithm's criteria by producing notes that are well sourced, clearly written, and nonpartisan.

AI writers appear specifically optimized—whether by design or by emergent capability—for the bridging algorithm's criteria by producing notes that are well sourced, clearly written, and nonpartisan. Section 3.2.3 and the rater feedback data identified the mechanism more precisely. The AI notes from our dataset scored higher on "addresses the claim" and "good sources"—the two qualities most directly associated with the retrieval-augmented generation (a technique where a language model retrieves relevant documents from an external knowledge base before generating a response) pipeline that likely powers these systems. They scored lower on "argumentative or biased" and "opinion/speculation," which are precisely the qualities the bridging algorithm penalizes. These findings are not surprising, given that AI systems draw from source material rather than personal viewpoints, making them less prone to the partisan framing that the bridging algorithm is designed to filter out.

62. Ibid.

Notably, AI writers achieved substantially higher CRH rates even on divisive topics like war and conflict (13.2 percent vs. 7.1 percent for humans), political content (15.8 percent vs. 5.7 percent), and explicit misinformation (19.9 percent vs. 8.6 percent).[63] One frequent criticism of Community Notes is that the algorithmic requirement for cross-partisan agreement limits the display of notes on divisive topics. Our data show that AI note writers craft notes that pass this threshold far more often than humans, suggesting that AI participation can extend the reach of corrections on contested topics and thereby improve the information ecosystem by adding helpful corrective material that a human-only system struggled to produce.[64]

The economic implication is that AI notes are not merely "cheap" substitutes for human effort; they produce corrections at a measurably higher success rate per unit of rater attention consumed. At 304 ratings per successful correction versus 908 for human writers, AI notes are roughly three times more resource-efficient, making them a better use of the system's scarce rating capacity. They also contributed the equivalent of $577,000 to $3.4 million in professional fact-checking labor.

The content profile data also revealed areas where AI and human writers have distinct advantages. AI's emphasis on "manipulated media" (39.4 percent vs. 13.3 percent for humans) likely reflects AI's natural advantage at detecting synthetic or altered images and videos through retrieval-augmented generation.[65] Conversely, AI writers underweight categories that require cultural literacy (satire detection) or temporal reasoning (outdated information), both areas where contextual human judgment retains an advantage.[66]

The per-writer data reinforces the durability of the quality signal. Writers that produced 1,000 to 6,000 notes sustained VSRs above 85 percent, suggesting that high volume does not compromise quality. At the other extreme, the lowest-performing AI writer (43D4A2A7) achieved only a 31.6 percent VSR across 101 notes, and another (BF87E4D4) managed only 46.2 percent across 217 notes—the only accounts below the human baseline of 68.5 percent.

5.2 The Over-Correction Problem

The most important finding from the rater feedback analysis is the "note not needed" signal. Our research revealed that the most common reason an AI note received a "not helpful" flag was that raters considered it unnecessary (50.6 percent of "not helpful" ratings), which exceeded the human rate by 8.2 percentage points and points to a systematic weakness in AI note-writing strategy.[67]

The most common reason an AI note received a "not helpful" flag was that raters considered it unnecessary (50.6 percent of "not helpful" ratings), which exceeded the human rate by 8.2 percentage points and points to a systematic weakness in AI note-writing strategy.

63. Ibid.
64. Tom Stafford, "Do Community Notes work?," The London School of Economics and Political Science, Jan. 14, 2025. https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/14/do-community-notes-work.
65. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.
66. Ibid.
67. Ibid.

The CRNH note samples illustrate the pattern, as AI notes that receive "not helpful" verdicts are often factually accurate but contextually misguided. For example:

- Correctly identifying a meme character's origin but missing that the post was making a pop-culture joke, not a factual claim.
- Accurately stating that a public figure was at a security conference in Germany but missing that the post's point was about the conference discussion, not the person's physical location.
- Providing a precise source for a viral clip but fact-checking what was obviously an entertainment post not intended as a factual assertion.

These failures reflect a potential limitation of current AI systems: They excel at retrieving and verifying discrete factual claims but struggle when a post's meaning is carried by irony, rhetoric, or implied argument rather than by a literal assertion. The 100 percent misleading-classification rate underscores this pattern: Operators configure AI writers to look for misinformation, and they find it even where it does not exist. The author of "World's First AI Community Note," explained that the first iteration of his note writer "was showing far too many mediocre/bad notes."[68] Although the precise reasons are unclear, he implemented a stronger rhetorical and satire filter that improved the bot's performance.[69] This suggests that the bot was targeting too many posts that clearly did not need a note, and the system struggled to detect satire and comedic-style rhetorical questions.

This over-correction tendency carries measurable costs. The 277,282 individual "note not needed" flags on AI notes represent roughly 6,900 hours of rater effort, valued at approximately $103,500, directed at evaluating notes the community considered unwarranted. Priority routing that directs raters toward notes from writers with demonstrated high VSRs would help address this problem by concentrating evaluation capacity where the expected return is highest, reducing wasted effort while increasing the volume of corrections that reach users.

While the Notes algorithm partly protects against bias through its bridging requirement, there is a risk that a small number of high-volume AI operators could game the system or inject a consistent bias.

5.3 Concentration Risk and the Growth Question

AI participation in Community Notes is scaling rapidly. Monthly AI output grew from 93 notes in September 2025 to 8,109 in February 2026 (the last complete month of data in our dataset).[70] At the time we downloaded the dataset on March 23, 2026, AI output accounted for 12.7 percent of the month's notes.[71] If this trajectory continues, AI could produce 10 percent to 20 percent of all community notes by mid-2026, with further growth possible as additional operators gain API access.

The fundamental structural risk is that a small number of operators, whose identities and models remain unknown, could control an increasingly

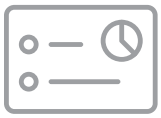
68. Young. https://nathanpmyoung.substack.com/p/worlds-first-ai-community-note#footnote-anchor-2-174094438.
69. Ibid.
70. R Street Institute analysis of "Download data," X, last accessed March 23, 2026. https://x.com/i/communitynotes/download-data.
71. Ibid.

significant share of the Community Notes output that users see. At 13.9 percent of CRH notes, AI accounts already produce a larger share of shown notes than their share of total output, meaning that any disruption to these accounts would disproportionately affect the corrections users actually encounter.

While the Notes algorithm partly protects against bias through its bridging requirement, there is a risk that a small number of high-volume AI operators could game the system or inject a consistent bias. This could introduce a consistent slant into the system, which could negatively affect the overall ecosystem. Disruption or loss of productive AI writers would mean that users would see fewer notes overall as well. X could consider creating transparency rewards or requirements for AI writers so that if some are lost or disrupted, they could be more easily replaced.

5.4 The Timing Convergence

The convergence in time-to-verdict to an insignificant difference (6.0 hours for AI-written notes vs. 6.3 hours for human-written notes, $P = 0.058$) deserves attention.[72] This suggests that the rating community has adapted to AI notes. Whatever novelty effect initially slowed evaluations has dissipated, and the rating pipeline now processes AI and human notes at comparable speeds.



AI time-to-verdict started higher in the program's early months and gradually declined toward parity with human notes. The OLS trend lines suggest this convergence should persist, with AI and human notes continuing to receive comparable processing speeds over the coming months.

The monthly trend analysis in Section 3.2.2 reinforces this finding. AI time-to-verdict started higher in the program's early months and gradually declined toward parity with human notes. The OLS trend lines suggest this convergence should persist, with AI and human notes continuing to receive comparable processing speeds over the coming months.

This is a positive signal for scalability. It suggests that increasing AI note volume would not inherently strain the rating pipeline's processing speed, at least not at current levels of note output. Whether this would hold true if AI output were to approach 20 percent or more of total volume remains an open question. However, our OLS analysis suggests that AI notes can achieve a median rating time of under 5 hours within the next six to 12 months, and possibly even faster in the future. Given the rapid decay of social media engagement, faster time-to-verdict could increase the impact of corrections by reaching users while posts are still circulating.

Misinformation will always be easier to spread because it can be posted without review, fact-checking, or research diligence.[73] By its nature, fact-checking will always take more time because it needs to engage in review processes and conform to standards. While there was a small advantage for humans in writing notes in less than 2 hours, it may be unreasonable to expect that all misinformation can be corrected so rapidly. However, it does seem that AI may become more capable of scanning posts, identifying misinformation, and creating well-sourced, unbiased corrections to that information than

72. Ibid.
73. Peter Dizikes, "Study: On Twitter, false news travels faster than true stories," MIT News, March 8, 2018. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.

humans. Our economic analysis also shows the high labor costs (potentially in the millions) of having humans perform that work. Because AI note writers can perform this task for a fraction of the cost, the potential improvements in the speed and efficiency of fact-checking are noteworthy.

5.5 Priority Routing: Closing the VSR-CRH Gap

The VSR-CRH gap analysis in Section 3.3 identifies 7,040 under-reviewed AI notes that would yield an estimated 6,193 additional corrections if raters could evaluate them. The Community Notes rating queue currently surfaces notes to raters without regard for the writer's track record. A reputation-weighted priority queue—that is, one that routes raters toward notes from writers with demonstrated high VSRs—could concentrate the system's scarce evaluation capacity where the expected return per rating is highest. The data to build such a system already exists. Every writer's VSR is computable from the public data, and the gap between that rate and their shown rate quantifies exactly how much value is being left on the table.

Priority routing would improve the economic equation in two ways. It would increase the volume of displayed corrections, realizing more of the writing labor value, and it would reduce over-correction costs by directing raters away from low-performing writers.

As discussed above, priority routing would improve the economic equation in two ways. It would increase the volume of displayed corrections, realizing more of the writing labor value, and it would reduce over-correction costs by directing raters away from low-performing writers. This approach complements both the AI rater proposal and the human rater incentive framework we will discuss in Section 6. AI raters could handle initial triage across the full queue, while human raters, whose attention is the binding constraint, could be directed first to notes from high-VSR writers where their evaluations are most likely to produce a correction that reaches users.

5.6 Limitations

This analysis is subject to several important limitations. First, the public data on Community Notes do not include engagement metrics for the posts that receive notes, so we cannot assess whether AI notes actually reduce misinformation exposure more effectively than human notes. Second, X has not publicly documented the criteria for **`apiEarnedIn`** enrollment, and we cannot determine who operates these accounts or what AI models power them. Third, notes currently in NMR status may eventually receive verdicts that shift the VSR calculations; the March CRH rate dip likely reflects this. Fourth, the keyword-based topic classification is approximate and may have misclassified notes on a topic the keyword lists do not capture. Finally, we cannot observe rater behavior at the individual level, so we cannot determine whether raters evaluate notes differently when they recognize them as AI-generated.

6. Implications and General Recommendations

The findings from this study point to specific actions that AI note-writing operators, X, and policymakers can take to strengthen the Community Notes system. This section outlines recommendations for improving AI content targeting, expanding rating capacity through AI raters and human incentive programs, and directing future research.

6.1 Improving AI Content Targeting and Model Utilization

Our rater feedback analysis highlights the importance of implementing better targeting logic to reduce the "note not needed" issue. A pre-screening step, evaluating whether a post's virality, factual claims, and potential for harm warrant a note before generating one, could substantially reduce the 50.6 percent "not needed" complaint rate and improve both rater efficiency and AI VSRs.

A pre-screening step, evaluating whether a post's virality, factual claims, and potential for harm warrant a note before generating one, could substantially reduce the 50.6 percent "not needed" complaint rate and improve both rater efficiency and AI VSRs.

The **FIRM_REJECT** analysis reveals a structural ceiling on note display rates that cannot be overcome by improving note quality alone. The algorithm still suppresses AI notes that are well-sourced, clearly written, and endorsed by a majority of raters when the minority opposition is large enough to indicate polarization. This is the algorithm working as designed, but it means the most socially valuable AI contributions—corrections on conflict misinformation and politically charged claims—are systematically the least likely to reach users. X should evaluate whether alternative scoring approaches can be expanded to recover high-quality notes on politically sensitive topics without undermining the bridging algorithm's core purpose.

The variance in AI writer performance also warrants attention. Two AI writers (43D4A2A7 and BF87E4D4) fell well below the human-verdict success baseline, demonstrating that API enrollment alone does not guarantee quality. X should consider establishing minimum performance thresholds for maintaining API access, with periodic review of per-account VSRs and automatic suspension for accounts that consistently underperform.

6.2 AI Note Raters

Note volume continues to grow as AI writers add capacity, and the rating economy analysis underscores why this matters. Rater attention is the binding economic constraint in the Community Notes system, and AI note growth places increasing demand on a resource that is already insufficient to process the existing queue. These dynamics point to a natural extension of the AI note-writing experiment: AI note raters that could augment human rating capacity. Just as AI note writers are addressing the supply side of the fact-checking equation, AI raters could address the demand side.

A natural extension of the AI note-writing experiment would be to have AI note raters augment human rating capacity, providing initial assessments that would help surface notes for human review more quickly.

Importantly, AI note raters would not replace human raters—they would supplement them, providing initial assessments that help surface notes for human review more quickly. Trained on the corpus of existing human ratings across the political spectrum, multiple AI rater models could represent different ideological perspectives, mirroring the bridging algorithm's requirement for cross-partisan agreement. Several safeguards would be necessary. AI ratings should carry lower weight than human ratings, be transparently labeled in the data, and be subject to periodic quality review to ensure they do not introduce systematic biases.

6.3 Incentivizing the Rater Workforce

The existing incentive structure offers only one meaningful reward: participants who rate enough notes with sufficient accuracy can eventually

earn the privilege of writing their own notes. However, for the large majority of participants who enroll as raters without aspirations to write, this incentive is invisible. Thus, Community Notes faces a classic volunteer-retention problem: It has successfully recruited a large workforce but lacks the mechanisms to keep that workforce engaged.

Community Notes faces a classic volunteer-retention problem: It has successfully recruited a large workforce but lacks the mechanisms to keep that workforce engaged. A tiered incentive framework, rather than any single mechanism, would best address the rater-retention problem while managing the risks inherent in each approach.

Other platforms that depend on volunteer contributions have confronted and partially solved this challenge. Reddit introduced its moderator rewards program in 2023, offering monetary compensation and premium features to high-performing moderators, complementing a long-standing karma system that provides visible, cumulative reputation signals.[74] Stack Overflow's reputation system gates access to platform features.[75] Users unlock editing privileges, review queues, and moderation tools as their peer-validated contributions accumulate. Although each of these could be considered for Community Notes, each also has downsides.

A tiered incentive framework, rather than any single mechanism, would best address the rater-retention problem while managing the risks inherent in each approach. At the first tier, low-cost recognition tools such as profile badges for accuracy milestones, anonymized contributor leaderboards, and public acknowledgment of high-quality raters would reward sustained accuracy rather than sheer volume. At the second tier, access-based progression could grant high-accuracy raters expanded analytics dashboards, early access to new Community Notes features, and contributor-specific performance reports, much as Stack Overflow rewards sustained contributors with expanded capabilities. At the third tier, material rewards offer the highest activation potential but also the highest gaming risk. Free X Premium subscriptions for sustained top performers, priority placement in any future creator revenue-sharing programs, or algorithmic visibility boosts for their posts could be options.

Section 5 of this paper established that rater attention is the binding constraint in the Community Notes system, and growing AI note volume is placing increasing demand on a finite pool of evaluators. Section 6.2 proposed AI raters as a technological solution to this bottleneck. Incentivizing the human rater workforce is the complementary institutional solution. Neither approach alone is sufficient. AI raters without human oversight risk introducing systematic biases that compound over time, and human-incentive programs without capacity augmentation simply shift the bottleneck rather than resolving it. The most robust path forward combines both: AI raters to handle initial triage and expand throughput, paired with a human-incentive system that activates the large dormant pool of enrolled participants to provide the diverse, cross-partisan oversight the bridging algorithm requires.

6.4 Future Research Directions

Several important questions remain for future work. Assessing the extent to which AI notes duplicate or complement human-written notes on the

74. Morgan Sung, "Reddit launches moderator rewards program amid site-wide discontent," TechCrunch, Aug. 24, 2023. https://techcrunch.com/2023/08/24/reddit-mod-helper-program-update-moderation-protest.
75. "Privileges," Stack Overflow, last accessed April 11, 2026. https://stackoverflow.com/help/privileges.

same posts would require post-level linkage analysis. Rater behavior studies investigating whether raters evaluate AI-generated notes differently from human notes would benefit from experimental designs not possible with observational data alone. In addition, the question of whether AI operators are learning from rater feedback and adapting their models to reduce the "note not needed" problem over time could be answered by longitudinal analysis of topic selection and CRNH rates by writer.

6.5 The Engagement Question

Perhaps the most consequential open question is whether these notes—written by humans or AI—actually change user behavior. We do not know whether users share, like, or engage with a post differently after a note appears, nor do we know whether the presence of a note reduces the subsequent virality of misinformation; both are fundamental to assessing whether the system delivers on its core promise.

Perhaps the most consequential open question is whether these notes—written by humans or AI—actually change user behavior.

Answering them rigorously would require access to post-level engagement data (likes, reposts, replies, and quote posts) before and after a note is displayed. The Community Notes public data includes "tweetId" for every note, which could serve as a linkage key to engagement metrics. A rigorous study would employ a causal inference design, synthetic control, or difference-in-differences methodology, comparing engagement trajectories of noted posts against matched posts that did not receive notes. This would require, at minimum, six to 12 months of continuous API access, a matched-pair sampling framework, and careful attention to selection effects (posts that receive notes may differ systematically from those that do not). Alternative approaches include partnering with X's research team or academic data access programs, which have periodically offered subsidized or free access for approved studies.

Perhaps one misunderstood use of engagement metrics is the implication that notes actually should reduce exposure to the original source material, which most studies show that they do to some degree.[76] However, if notes are in fact high quality and educational, in some cases more exposure to them would benefit users. Future research could consider this potential benefit to the information ecosystem.

7. Public Policy Recommendations

The findings of this study carry implications that extend beyond the Community Notes program itself. The seven months of analyzed data reveal a pattern of productive interdependency between human and automated contributors. Human volunteers supply the ideological diversity, cultural literacy, and rating judgments that the bridging algorithm requires to separate genuine quality from partisan noise. AI writers draw on the corpus of human-sourced journalism, reference material, and institutional fact-checking to produce corrections at a volume and sourcing density that human contributors alone cannot match. This collaborative structure, which is not directed by

76. Isaac Slaughter et al., "Community notes reduce engagement with and diffusion of false information online," *Proceedings of the National Academy of Sciences* 122:38 (Sept. 18, 2025). https://www.pnas.org/doi/10.1073/pnas.2503413122.

government or mediated by professional fact-checkers, bears directly on four current policy debates: (1) proposals to import European-style content moderation regimes to the United States, (2) the rapid proliferation of state-level AI regulation, (3) continuing efforts to narrow Section 230, and (4) the federal role in funding the statistical research on which fact-checking depends.

7.1 Avoid Adopting European-Style Content Moderation Frameworks

The constitutional problem with European-style, government-directed fact-checking is that the arrangement creates a permanent channel for official pressure on private speech.

The most prominent alternative to the Community Notes approach to platform-level misinformation correction, embodied in the European Union's (EU's) Digital Services Act (DSA) and the European Commission's coordinated fact-checking apparatus, treats content moderation as a regulated compliance function.[77] Very large online platforms in the EU must assess and mitigate "systemic risks" related to disinformation, submit to regulatory audits, and coordinate with approved fact-checking organizations funded through the European Digital Media Observatory.[78] The approach assumes that accurate information is best produced by government-designated arbiters operating within a prescribed institutional framework.

The Community Notes data discussed in this paper suggests that this assumption is wrong on both counts. The signal that determines whether a note reaches users is the cross-partisan agreement of ordinary raters, not the professional judgment of credentialed fact-checkers. And the most rapidly improving contributors to that signal are privately operated AI systems developed without regulatory direction. The system works because it is permissionless, distributed, and responsive to the design incentives of a private platform rather than administrative rules.

The United States has no equivalent to the DSA, and policymakers should ensure that it remains that way. The constitutional problem with government-directed fact-checking is not that the fact-checkers are wrong. It is that the arrangement creates a permanent channel for official pressure on private speech—the kind of coercive entanglement the Supreme Court recently addressed in *NRA v. Vullo*, which held that government officials violate the First Amendment when they use regulatory leverage to coerce private intermediaries into suppressing disfavored speech.[79] Even where such pressure falls short of an outright command, it distorts the moderation ecosystem: Platforms that align with government preferences are rewarded, and platforms that experiment outside of the approved framework are penalized. These distortions also travel beyond EU borders. Recent congressional findings describe how EU regulatory requirements—including the fact-checking arrangements they prescribe—already reach American platforms indirectly, as multinational companies apply the more stringent EU compliance standards globally rather than maintaining

77. "DSA: Very large online platforms and search engines," European Commission, March 10, 2026. https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops.
78. "DIGITAL success stories – European Digital Media Observatory," European Commission, Dec. 8, 2025. https://digital-strategy.ec.europa.eu/en/factpages/digital-success-stories-european-digital-media-observatory.
79. *National Rifle Association of America v. Vullo*, Supreme Court of the United States, May 30, 2024. https://www.supremecourt.gov/opinions/23pdf/22-842_6kg7.pdf.

separate systems.[80] Importing that framework domestically would reproduce these distortions rather than avoid them. In contrast, every additional Community Note correction that reaches users is a correction that did not require a subpoena, a take-down order, or a regulatory referral.

7.2 Preempt State-Level AI Regulation That Would Foreclose Decentralized Moderation

A 50-state patchwork of rules would slow the development of AI-moderation tools, whereas a federal framework focused on transparency and performance would support them.

State-level AI regulation poses a similar, but more immediate, threat—in this case to the developer profile that has produced the highest-quality AI notes in the data examined for this study. The most productive AI note writers in our dataset were deployed by small-scale operators (often individuals) without institutional backing or dedicated compliance staff. Several states have proposed bills that would establish new AI-related compliance obligations (e.g., pre-deployment risk assessment, algorithmic audit, disclosure, civil liability rules) that these operators would struggle to meet.[81] Some of these proposals go even further, attaching special liability or labeling rules to AI-generated content on politically sensitive topics like those seen in Community Notes. A 50-state patchwork of such rules would not merely slow the development of AI-driven moderation tools—it would structurally disadvantage the small and experimental developers whose notes the Community Notes rating system has already proven capable of evaluating without external oversight.

The most constructive policy response to this issue is federal preemption of state-level AI content regulation in narrowly defined areas, including AI-generated content used for moderation and fact-checking. A federal framework focused on transparency and performance, rather than on prescriptive design requirements, would establish a single set of national expectations under which an individual operator could build and deploy a note writer.

The stakes are significant: Community Notes is one of the most consequential experiments in applied AI, now operating to improve fact-checking at scale. A patchwork of state mandates risks foreclosing the iteration that makes the program work by prescribing how models are trained and what they are capable of doing. Transparency requirements, by contrast, would give researchers access to the operational data needed to evaluate performance, identify failure modes, and refine the system over time. Improvement in this space will come from observing how these systems perform, not from prescribing in advance how they should be built.

7.3 Preserve Section 230 in Its Current Form

Section 230 allows platforms to host millions of user-written and user-rated notes without being treated as the publisher of each individual correction.

The decentralized moderation model described above depends on a specific legal foundation. Section 230 of the Communications Decency Act is the statute that makes Community Notes and systems like it possible.[82] By shielding platforms from liability for user-generated content and protecting their good-

80. Spence Purnell, "How the 'Brussels effect' harms the U.S. tech sector," R Street Institute, Feb. 6, 2026. https://www.rstreet.org/commentary/how-the-brussels-effect-harms-the-u-s-tech-sector.
81. Adam Thierer and Kevin Frazier, "Congress Should Lead On AI Policy, Not The States," R Street Institute, Feb. 4, 2026. https://www.rstreet.org/commentary/congress-should-lead-on-ai-policy-not-the-states.
82. 47 U.S.C. § 230. https://www.law.cornell.edu/uscode/text/47/230.

faith moderation decisions, Section 230 allows platforms to host millions of user-written and user-rated notes without being treated as the publisher of each individual correction. Without that protection, the legal exposure created by displaying a note that later proves inaccurate, or by declining to display a note that a plaintiff believes should have been shown, would quickly foreclose the kind of open, iterative moderation the program relies on. Proposals to narrow or repeal Section 230 frequently overlook this issue. Importantly, the statute does not immunize bad actors from legal consequences for their own unlawful conduct; it immunizes the platform infrastructure that allows lawful user speech and lawful user moderation to coexist at scale. Community Notes is one of the most successful demonstrations of what that infrastructure makes possible, and the program's experience should inform how policymakers weigh proposals to alter Section 230.

7.4 Sustain Federal Funding for the Research Corpus on Which Fact-Checking Depends

One important contribution federal policy can make is sustained investment in the federal statistical agencies that produce the baseline data on which nearly all downstream research depends.

The recommendations above identify policy levers the federal government should refrain from pulling. The case for restraint, however, should not be confused with a case for federal disengagement. Government has a continuing and constructive role to play in supporting the information ecosystem, but that role lies upstream of fact-checking itself: Human and AI note writers will continue to need statistics and research on which to base their facts. Government agencies are often a key source for those statistics.[83]

Thus, one important contribution federal policy can make is sustained investment in the federal statistical agencies that produce the baseline data on which nearly all downstream research depends. The Bureau of Labor Statistics, Bureau of Economic Analysis, Census Bureau, Energy Information Administration, and Centers for Disease Control and Prevention's National Center for Health Statistics generate employment figures, output measures, demographic counts, energy data, and public health indicators that both private and public researchers treat as authoritative starting points.

When a Community Notes contributor or an AI note writer corrects a misleading claim about inflation, unemployment, population trends, or disease prevalence, the correction almost invariably traces back, directly or through an intermediate source, to one of these agencies. Our research reflected this, with more than 3,500 direct federal statistical citations, and possibly more in secondary sources. The fact that AI cited statistical agencies at a higher rate than humans also suggests that AI may perform better at settling pure fact-based disputes with authoritative figures rather than speculation or opinion.

In short, U.S. statistical agencies are the load-bearing layer beneath the entire information-correction apparatus. Congress should treat appropriations for the federal statistical system as core information-ecosystem infrastructure rather than as discretionary spending subject to routine cuts.

83. "Federal Research and Development (R&D) Funding: FY2026," Congress.gov, April 17, 2026. https://www.congress.gov/crs-product/R48694.

7.5 Policy Takeaway: Preserve the Conditions That Produced the System

Preserve the conditions that produced Community Notes rather than substituting regulatory frameworks built for a different model of moderation.

Taken together, these recommendations point in a single direction: preserve the conditions that produced Community Notes rather than substituting regulatory frameworks built for a different model of moderation. The private sector has not solved the problem of online misinformation, and it may not be a fully solvable problem, as misinformation will always be easier to create and distribute than a correction system like Community Notes can address. The seven months of data examined here, however, suggest that the system is making measurable progress. That progress is the best argument against government displacement of the systems producing it.

8. Conclusion

AI note writers have become a significant force in Community Notes. In seven months, they grew from zero to 12.7 percent of monthly note volume and 13.9 percent of all shown notes, while sustaining VSRs nearly 20 percentage points above the human average. Their content profile is distinctive and specialized, with a singular focus on misinformation correction, heavy reliance on institutional sources, and frequent flagging of manipulated media. On the economic side, they contributed the equivalent of approximately $577,000 to $3.4 million in professional fact-checking labor and consumed the system's scarce rater resource roughly three times more efficiently than human writers, requiring only 304 ratings per successful correction versus 908 for human-written notes.

The path forward requires action on multiple fronts. AI operators should refine their targeting to reduce unnecessary notes—a change that would both improve VSRs and release thousands of hours of rater capacity for the existing evaluation backlog. Priority routing (i.e., directing raters toward notes from writers with demonstrated high VSRs) could nearly double the number of AI corrections that reach users while reducing wasted rater effort. The rating pipeline should be augmented to handle growing volume by using AI raters to expand throughput and incentive programs to activate the large pool of enrolled-but-dormant human raters.

Whether the system continues to improve depends on the choices platforms, operators, and policymakers make from here. The findings of this study suggest those choices are worth making carefully.

On the policy front, the conditions that produced Community Notes—robust First Amendment protections, federal restraint on prescriptive moderation frameworks, and Section 230 liability shielding—are themselves the most productive contribution policymakers can make to the information ecosystem. Community Notes is not a finished system, and the seven months of data examined here capture only its earliest experiment with AI participation. But the possibility is promising: a permissionless, cross-partisan, human–AI hybrid producing measurable improvements in the accuracy and reach of online corrections, at a fraction of the cost of professional fact-checking. Whether the system continues to improve depends on the choices platforms, operators, and policymakers make from here. The findings of this study suggest those choices are worth making carefully.

About the Author

Spence Purnell is a senior fellow for the R Street Institute's Technology and Innovation team. His work focuses on the impact of misinformation on society and public policy, including creating a new framework for information governance.