Can LLMs Contribute to Cooperative Fact-Checking? A Field Evaluation on X Community Notes

Haiwen Li [1] **Michiel A. Bakker** [1]

Abstract

Combating misinformation on social media increasingly relies on collective, user-driven fact-checking. X Community Notes exemplifies this approach: users with different viewpoints propose contextual notes to misleading content and evaluate them, and a bridging algorithm surfaces notes that achieve cross-partisan agreement. We study whether large language models (LLMs) can effectively participate in this process. We present the first field evaluation of LLM fact-check writing on a live platform, using X Community Notes' "AI writer" feature. Over a three-month period, our LLM system wrote 1,614 notes on 1,597 tweets, alongside 1,332 human-written notes on the same tweets, evaluated using 108,169 ratings from 42,521 users. At the rating level, LLM notes receive more positive evaluations than human-written notes across raters with different political leanings. Although platform constraints limit the exposure of LLM notes relative to human notes, note-level analysis that accounts for the differential exposure confirms the same advantage. Together, these findings show that LLMs can generate broadly helpful fact-checks at scale in real-world settings, and provide field evidence that LLMs can contribute to a cooperative information system where success depends on acceptance across disagreement.

1. Introduction

Combating misinformation on social media is not only a problem of identifying false claims, but of producing corrections that people with different perspectives find credible and helpful. Effective fact-checking should therefore be a collective process: corrections must be written, evaluated, and ultimately accepted by a diverse population. A fact-check that persuades one group while alienating another may be accurate, yet ineffective at improving the broader information ecosystem.

Crowdsourced systems such as X Community Notes operationalize this idea, and have been shown to increase trust and reduce misinformation engagement (Drolsbach et al., 2024; Slaughter et al., 2025). Users can propose short, source-cited notes that add context to potentially misleading posts, and other users rate their helpfulness. A note then becomes publicly visible only if it receives positive evaluations from users who have historically disagreed, via a bridging algorithm designed to reward cross-group agreement (Wojcik et al., 2022). In this way, helpfulness is defined not solely by accuracy, but by whether a note is broadly accepted across a heterogeneous set of raters.

This requirement of cross-group agreement provides an opportunity to study an important question: whether AI systems can contribute information that is acceptable across groups with different viewpoints. This question has become practically relevant because Community Notes now allows AI systems to participate in note writing through its "AI Writer" API. Like human-written notes, AI notes are evaluated by Community Notes raters and must satisfy the platform's bridging criterion before becoming publicly visible. Community Notes therefore provides a natural field setting for evaluating AI participation in a human-governed cooperative information system.

Prior work suggests that LLMs can help identify common ground. Bakker et al. (2022) and Tessler et al. (2024) show that LLM-generated consensus statements were preferred over human-written statements in deliberation about divisive political topics, and Konya et al. (2025) extends these methods to a real-world conflict setting. More broadly, research in cooperative AI studies how AI systems can facilitate agreement and consensus among agents with heterogeneous beliefs and objectives (Dafoe et al., 2020; 2021). However, in the context of misinformation and fact-checking, it remains unclear whether LLM-generated content can earn acceptance from a diverse population under the conditions of a live social media platform.

[1]Massachusetts Institute of Technology, Cambridge, US. Correspondence to: Haiwen Li <haiwenli@mit.edu>, Michiel A. Bakker <bakker@mit.edu>.

Trustworthy AI for Good (AI4GOOD) Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea.

Existing evaluations of LLM fact-checking remain largely confined to offline controlled settings. Studies have shown that LLMs can generate fact-checking notes comparable to, and under some conditions better than, human-written notes, but they evaluated performance using automated benchmarks, expert annotations, or crowdworker judgments (Zhou et al., 2024; De et al., 2025; Costabile et al., 2025; Singh et al., 2026). These evaluations also typically focus on metrics such as factual accuracy and perceived trustworthiness. A critical gap therefore remains: we lack evidence on how LLM-generated fact-checks perform on a live platform, where their effectiveness is determined not by gold-standard labels but by the aggregated judgments of a diverse population.

In this paper, we present the first field evaluation of LLM fact-check writing deployed on-platform within a cooperative human evaluation system. Using X Community Notes' "AI writer" API, we deployed a multimodal LLM-based Community Notes writer. The writer conducts web and platform-native search to collect evidence, decides whether sufficient evidence exists to warrant a note, generates contextual notes, and publishes them directly to the platform.[1] Over a three-month deployment period, from November 1, 2025 to January 31, 2026, our LLM writer published 1,614 notes on 1,597 unique tweets. We compare these notes against 1,332 human-written notes targeting the same set of tweets, drawing on 108,169 ratings from 42,521 Community Notes raters.

Our analysis highlights both the potential and the constraints of LLM participation in this collective process. Platform rules introduce systematic exposure differences: AI-generated notes are posted later and receive fewer ratings, which mechanically affects their on-platform performance. To address this, we conduct both rating-level and note-level analyses. We find that LLM-generated notes receive more positive evaluations from raters with different political leanings, and that this advantage persists in note-level analysis that accounts for exposure differences. Together, these results provide empirical evidence that an LLM writer can produce fact-checks that achieve broad agreement in a system where content is evaluated by diverse users under real-world conditions. More broadly, our findings provide a field test of a central question in cooperative AI: whether AI systems can contribute productively to human-governed systems for producing agreement across disagreement.

[1]Code for the writing pipeline and analysis is available at https://github.com/haiwen-li/ai-fact-checking-in-the-wild.

2. Community Notes as a Cooperative Evaluation Setting

2.1. Cross-Perspective Helpfulness as the Evaluation Criterion

X Community Notes uses a matrix factorization–based bridging algorithm to aggregate individual ratings and evaluate note quality.[2] Rather than relying on simple majority voting, the algorithm assigns higher helpfulness scores to notes rated helpful by raters who have historically disagreed in their past ratings. Notes whose helpfulness scores exceed a platform-defined threshold receive Currently Rated Helpful (CRH) status and become publicly visible. In addition, this algorithm estimates a numeric factor (rater factor) for each rater which aligns closely with political ideology, with negative values corresponding to left-leaning rating patterns and positive values to right-leaning ones (Wojcik et al., 2022). By operationalizing note quality as cross-perspective helpfulness, the design provides a natural setting for testing whether AI-generated factual context can be accepted across disagreement.

2.2. A Shared Environment for Human and AI Contributors

The Community Notes AI writer API enables automated systems to propose notes alongside human contributors. Notes generated by AI writers are clearly labeled as AI-created and are evaluated by the same group of human raters under the same bridging criteria as human-written notes. This creates a shared evaluation environment in which both types of contributions are assessed by the same human rating process.

While the evaluation criteria are shared, the conditions under which human and AI writers operate differ. On X Community Notes, human writers can propose notes whenever they notice a misleading post, while AI writers can only write notes on posts after sufficient users have flagged them. This rule systematically delays AI submissions. In our data, among the 814 tweets receiving both note types, 66.0% of human notes were created before our LLM notes. As a result, human notes accumulated substantially more ratings (mean $= 109.94$ vs. 59.50; median $= 51$ vs. 22; Mann–Whitney $U = 740{,}798.0$, $p < 0.001$). These exposure differences matter because Community Notes outcomes depend not only on note quality, but also on when and how often a note is seen. Earlier notes have more time to accumulate ratings and may receive more prominent placement in the interface. In addition, the Community Notes algorithm penalizes notes with fewer ratings through regularization, which mechanically lowers the helpfulness scores of less-rated notes. As

[2]https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes

a result, raw on-platform note-level outcomes may understate the quality of AI notes relative to human notes. In the analyses below, we address this challenge by combining rating-level models with note-level analyses that account for differential exposure.

3. Data and Methods

3.1. LLM Note Writing Pipeline

The LLM writer we deployed is a multi-step pipeline. Upon retrieving a flagged post via the AI writer API, we compile its complete context by concatenating the post's creation timestamp, its primary text, and any text and image descriptions from quoted or replied-to posts (if applicable). To enable multimodal note understanding, the LLM writer is directly provided with images or video thumbnails associated with the target post. Next, we use Grok-4-fast with its web search and X search capabilities to conduct information research and collect relevant evidence. (Grok was chosen for its ability to surface X posts and understand video content from X.) These research outputs are next used to inform the note-writing process. After evidence collection, the LLM writer (GPT-5-mini) decides whether a note should be written. A note may not be written if the post is unlikely to be perceived as misleading or if there is insufficient evidence. This decision step acts as a safeguard to prevent generation of spam notes on non-misleading posts. Once the decision to write a note is made, the writer LLM (GPT-5-mini) creates a note given the post content and Grok-collected evidence. The note then goes through URL validity checks, length checks, and a quality check with the Community Notes ClaimOpinion model prior to submission.

3.2. Dataset

Our dataset comprised 2,946 Community Notes written between November 1, 2025 and January 31, 2026, targeting 1,597 unique tweets, with 108,169 ratings from 42,521 unique raters. The sample included 1,614 notes written by our LLM writer and 1,332 notes written by human users on the same set of tweets. We applied two quality filters: excluding media notes[3], and filtering ratings to include only those from raters with a valid rater factor.

3.3. Analytical Strategy

Direct comparison of platform-calculated note helpfulness score and status between LLM and human notes is complicated by the exposure asymmetry described in Section 2.2. We therefore pursue two complementary strategies.

[3]https://communitynotes.x.com/guide/en/contributing/notes-on-media

Rating-level analysis. The rating-level analysis models individual rater evaluation. Community Notes allows users to rate each note as Helpful, Somewhat Helpful, or Not Helpful; we code these as 1.0, 0.5, and 0.0 following platform convention. We estimate a linear mixed effects model predicting rating scores from LLM authorship, rater ideology (`rater_factor`), and their interactions:

$$\begin{aligned}\texttt{rating\_score} \sim\ & \texttt{AI} \times \texttt{rater\_factor} \\ & + \texttt{AI} \times (\texttt{rater\_factor})^2 \\ & + (1|\texttt{noteId}) + (1|\texttt{raterId})\end{aligned}$$

This specification mirrors the bridging criterion of the Community Notes algorithm by modeling how the LLM advantage varies as a function of rater ideology. Specifically, the model includes both the linear and quadratic forms of the ideology factor, each interacted with LLM authorship, to allow the LLM effect vary asymmetrically across the political spectrum. Note and rater random intercepts account for within-note correlation and between-rater heterogeneity. We also conduct exploratory subgroup analyses by tweet modality and topic by re-estimating this specification within subsets.

Note-level analysis. We complement the rating-level analysis with an equal-exposure note-level analysis, by comparing LLM and human notes on three platform-defined note-level outcomes: note helpfulness scores, Currently Rated Helpful status, and Currently Rated Not Helpful status. To equalize exposure, for each tweet, we retain only ratings from raters who saw and rated every note on that tweet, ensuring that all notes on the same tweet received the same set of ratings. We then recompute note helpfulness scores by applying the Community Notes scoring algorithm to these filtered ratings alone, and estimate differences using a linear mixed effects model with tweet random intercepts to account for within-tweet dependence among notes targeting the same post. We apply the Benjamini–Hochberg correction across the three outcome tests to control the false discovery rate.

4. Results

4.1. LLM Notes Receive More Positive Ratings Across the Ideological Spectrum

We first examine how individual raters evaluate LLM versus human notes across the ideological spectrum. We stratify raters into three ideology groups based on their rater factor: left (factor < -0.15), neutral ($-0.15 \leq$ factor ≤ 0.15), and right (factor > 0.15). For each note and rater group, we computed the percentage of ratings marked helpful and not helpful, then aggregated across notes to obtain mean

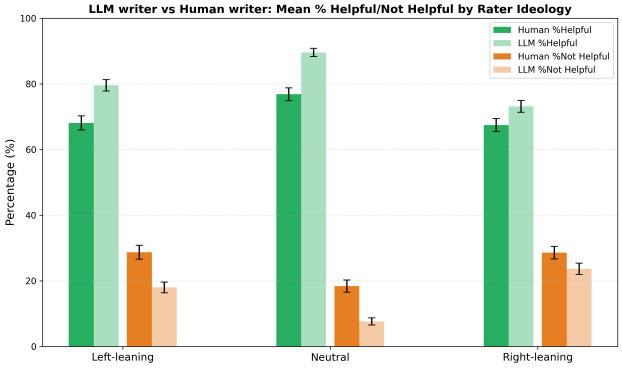

Figure 1. Mean % helpful and % unhelpful ratings per note, stratified by rater ideology group. Error bars show 95% confidence intervals across notes.

Table 1. Rating-level linear mixed effects model results. Standard errors are listed in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Note + Rater RE	Tweet + Rater RE	OLS Clustered	Note + Rater RE (Group)
(Intercept)	0.785***	0.842***	0.826***	0.785***
	(0.006)	(0.005)	(0.003)	(0.007)
AI	0.104***	0.066***	0.082***	0.098***
	(0.009)	(0.004)	(0.004)	(0.010)
coreRaterFactor1	-0.015***	-0.007	-0.005	
	(0.004)	(0.004)	(0.003)	
coreRaterFactor1^2	-0.175***	-0.195***	-0.291***	
	(0.010)	(0.010)	(0.009)	
AI:coreRaterFactor1	-0.087***	-0.112***	-0.120***	
	(0.005)	(0.005)	(0.005)	
AI:coreRaterFactor1^2	-0.190***	-0.169***	-0.181***	
	(0.013)	(0.013)	(0.014)	
left-leaning rater				-0.031***
				(0.005)
right-leaning rater				-0.059***
				(0.005)
AI:left-leaning rater				-0.003
				(0.007)
AI:right-leaning rater				-0.086***
				(0.007)
SD (Intercept rater)	0.160	0.160		0.164
SD (Observations)	0.304	0.335		0.304
SD (Intercept note)	0.196			0.197
SD (Intercept tweet)		0.142		
Num.Obs.	108169	108169	108169	108169

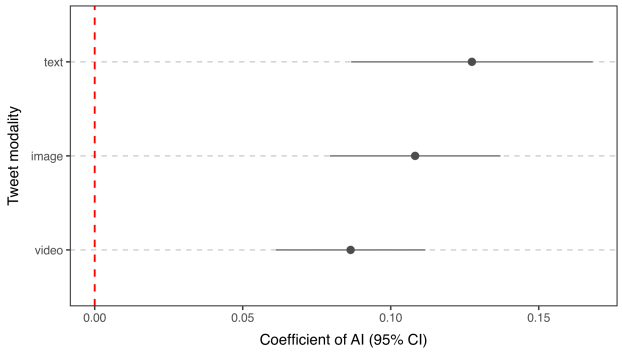

Figure 2. LLM vs. human note rating advantage (AI main-effect coefficient with 95% CI) by tweet modality (text-only, image, video).

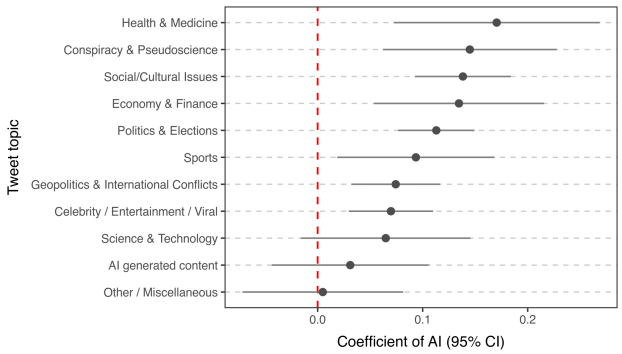

Figure 3. LLM vs. human note rating advantage (AI main-effect coefficient with 95% CI) by tweet topic category.

percentages with 95% confidence intervals. Results are presented in Figure 1. LLM notes have higher average % helpful ratings and lower average % unhelpful ratings than human notes across all rater groups. The largest difference appears among neutral raters, followed by left-leaning raters, and right-leaning raters.

To quantify differences in individual ratings between LLM and human notes, we estimated a linear mixed effects model predicting individual rating scores from LLM writer authorship, rater's political ideology, and their interactions. Table 1 shows that LLM notes receive significantly more positive ratings at the center of the ideology spectrum (AI coefficient $= 0.104$, $p < 0.001$); given that human notes' average helpful ratings achieve around 78.5% from centrist raters, this corresponds to approximately a 10-percentage-point increase. The negative quadratic interaction indicates that this advantage is largest among moderate raters and diminishes toward both ideological extremes, while the linear interaction indicates that the advantage decreases more steeply among right-leaning raters. The estimates remain consistent across alternative specifications (Table 1, Models 2–4).

To explore how the LLM advantage in helpfulness ratings varies across tweet types, we conducted exploratory subgroup analyses by tweet modality and topic. The LLM advantage is consistently positive across all three modality types but varies in magnitude (Figure 2): the point estimate is largest for text-only posts, followed by posts with images and videos. Topic-level heterogeneity is more pronounced (Figure 3). LLM notes receive better ratings on posts about health and medicine and conspiracy theories and pseudoscience claims. Notably, LLM notes show minimal or non-significant advantage when fact-checking AI-generated content.

4.2. Equal-Exposure Note-Level Analysis Confirms the Advantage

The rating-level analysis demonstrates that LLM notes receive more positive individual evaluations across the ideological spectrum. We complement this with a note-level analysis to investigate whether this advantage translates into better note-level outcomes. Using raw platform-calculated outcomes, the comparison shows mixed signals. LLM notes accumulate fewer ratings on average (59.50 vs. 109.94) and have a lower CRH rate (13.07% vs. 18.02%, see Appendix B), while having similar note helpfulness scores (mean: 0.25 vs. 0.24) and a lower rate of reaching Currently Rated Not Helpful status (1.12% vs. 4.13%). However, as discussed in Section 2.2, direct comparison of platform note-level outcomes is confounded by differential exposures that may disadvantage LLM writers.

To address this, we construct a subset of ratings that equalizes exposure by design and recompute note helpfulness scores by applying the Community Notes scoring algorithm to these ratings alone. This yields a subset of 13,721 ratings across 1,674 notes on 663 tweets. We compare the rater distributions before and after filtering and find that they do not differ substantially from the broader population in political leaning or helpfulness leniency (Appendix D).

Under equal exposure, LLM notes achieve significantly higher helpfulness scores than human notes (mean: 0.21 vs. 0.18; AI coefficient $= 0.019$, $z = 2.944$, adj. $p = 0.010$), confirming that the LLM quality advantage identified in our rating-level analysis carries through to the note-level outcome once the exposure confound is removed. We also derive final rating status from the recomputed helpfulness scores: a higher proportion of LLM notes reach CRH status (2.40% compared to 1.89% for human notes) and a lower proportion reach CRNH status (0.90% compared to 1.39% for human notes), though these differences are not statistically significant. We report further note-level robustness checks in Appendix C, which provide consistent results, including restricting the sample to notes with at least 30 ratings to mitigate the algorithm's penalty on low-rating-count notes and creation-time-matched analyses to control for differences in exposure.

4.3. Differences in Writing and Sourcing Strategies

LLM and human notes differed in both writing and source citation behaviors. LLM notes were substantially longer than human notes (mean 35.8 words vs. 26.9 words, $t = 22.190$, $p < 0.001$). The LLM writer also cited more URLs on average (mean: 1.51 vs. 1.23 URLs; $t = 8.67$, $p < 0.001$). Analysis of cited domains reveals distinct sourcing patterns (Tables 2 and 3).

LLM notes most frequently referenced mainstream news outlets and social media platforms, including reuters.com (7.7% of LLM notes), en.wikipedia.org (7.2%), instagram.com (5.8%), youtube.com (5.3%), x.com (4.2%), bbc.com (4.2%), snopes.com (3.9%), cnn.com (2.4%), facebook.com (2.0%), and yahoo.com (2.0%). In contrast, human notes most commonly cited x.com (18.7%), nearly four times the rate observed in LLM notes (4.2%). Wikipedia appeared at similar rates in both groups (7.6% human vs. 7.2% LLM), but other mainstream news sources appeared less frequently in human notes: reuters.com (1.6% vs. 7.7%), bbc.com (1.1% vs. 4.2%), and snopes.com (0.7% vs. 3.9%). This suggests that the LLM writer relies more heavily on traditional authoritative sources, whereas human writers more frequently reference platform-native content and social media posts. We note that this pattern characterizes our specific implementation.

Table 2. Top 10 domains cited in LLM notes, with comparison to human notes.

Rank	Domain	% LLM	% Human
1	reuters.com	7.7	1.6
2	en.wikipedia.org	7.2	7.6
3	instagram.com	5.8	2.0
4	youtube.com	5.3	3.9
5	x.com	4.2	18.7
6	bbc.com	4.2	1.1
7	snopes.com	3.9	0.7
8	cnn.com	2.4	0.8
9	facebook.com	2.0	0.7
10	yahoo.com	2.0	0.5

Table 3. Top 10 domains cited in human notes, with comparison to LLM notes.

Rank	Domain	% LLM	% Human
1	x.com	4.2	18.7
2	en.wikipedia.org	7.2	7.6
3	youtube.com	5.3	3.9
4	x.com/grok	0.4	2.6
5	instagram.com	5.8	2.0
6	reuters.com	7.7	1.6
7	theguardian.com	1.2	1.6
8	t.co	0.0	1.1
9	bbc.com	4.2	1.1
10	share.google	0.0	1.0

5. Discussion

Our results show that LLM-generated Community Notes can achieve broader cross-ideological acceptance than human-written notes, receiving more positive ratings from raters across the political spectrum. This finding is particularly significant given that Community Notes prioritizes cross-perspective helpfulness over simple accuracy or neutrality. In this sense, Community Notes provides a real-world cooperative information system: contributors propose factual context, but success depends on whether that context can

be accepted by people who may otherwise disagree. The fact that LLMs perform better than human notes on average on this standard suggests they are capable of being broadly helpful rather than merely presenting factual information, a capability with important implications for scaling content moderation while maintaining or even improving the quality of contextual information provided to users.

This has implications for information integrity at scale. Misinformation spreads faster than professional fact-checkers can respond, and while crowdsourced systems distribute this burden, they still rely heavily on voluntary human effort which is limited in volume. Our results suggest that scaling AI contributions can close the gap. At the same time, LLM and human writers have different strengths that are not easily substituted (Li et al., 2025). LLMs can rapidly synthesize widely available information and generate notes at scale, while human contributors are better suited to handling novel events, niche topics, and rapidly evolving events that require domain knowledge and understanding of how posts may be interpreted. Designing AI-augmented fact-checking systems to harness this complementarity, rather than treating AI deployment as a replacement for human community participation, could enhance the overall quality of the information environment.

More broadly, the study provides field evidence for cooperative AI in a civic information setting. Prior work has shown that LLMs can help groups identify common ground in controlled deliberation tasks. We extend this line of work to a naturally occurring, large-scale platform environment where agreement is not elicited in a laboratory but inferred from the judgments of tens of thousands of real users.

Finally, this study highlights the importance of field evaluation for AI systems that interact with human communities. Offline benchmarks, expert labels, and crowdworker ratings are useful, but they cannot fully capture how AI systems perform inside live socio-technical systems. By using the transparent data infrastructure of Community Notes, we evaluate LLM-written notes under natural platform conditions and with organic user feedback. The results suggest that LLM fact-checking can be viable in real-world settings, and LLMs can serve as productive contributors to cooperative information systems.

Impact Statement

This work provides field evidence that LLMs can produce fact-checks accepted by people with diverse political viewpoints, demonstrating a form of AI contribution to cooperative human systems. If scaled responsibly, AI-augmented fact-checking could help close the gap between the speed of misinformation spread and the capacity of human fact-checkers to respond. More broadly, as platforms open participation to AI agents in systems governed by human evaluation, understanding the conditions under which AI contributions are genuinely helpful becomes a critical question for cooperative AI research.

References

Bakker, M. A., Chadwick, M. J., Sheahan, H., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., Summerfield, C., and Botvinick, M. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38176–38189, 2022.

Costabile, L. et al. Assessing the potential of generative agents in crowdsourced fact-checking. *Online Social Networks and Media*, 48:100326, 2025.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

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

Drolsbach, C. P., Solovev, K., and Pröllochs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7):pgae217, 2024.

Konya, A., Chatham, C., Moran, S., Wilkins, M., and Small, C. T. Bridging divides: AI-mediated communication in the Israeli–Palestinian conflict. *arXiv preprint arXiv:2502.02478*, 2025.

Li, H., De, S., Revel, M., Haupt, A., Miller, B., Coleman, K., Baxter, J., Saveski, M., and Bakker, M. Scaling human judgment in community notes with LLMs. *Journal of Online Trust and Safety*, 3(1), September 2025. ISSN 2770-3142. doi: 10.54501/jots.v3i1.255.

Singh, S., Jaidka, K., and Kan, M.-Y. GitSearch: Enhancing community notes generation with gap-informed targeted search. *arXiv preprint arXiv:2602.08945*, 2026.

Slaughter, I., Peytavin, A., Ugander, J., and Saveski, M. Community notes reduce engagement with and diffusion of false information online. *Proceedings of the National Academy of Sciences*, 122(38):e2503413122, 2025. doi: 10.1073/pnas.2503413122.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Kocisky, T., Evans, R., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., and Summerfield, C. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., and Baxter, J. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.

Zhou, X., Sharma, A., Zhang, A. X., and Althoff, T. Correcting misinformation on social media with a large language model. *arXiv preprint arXiv:2403.11169*, 2024.

Appendix

A. Within-Rater Pairwise Comparison

As a complement to the rating-level analysis, we conducted a pairwise comparison restricted to raters who evaluated both an LLM and a human note on the same tweet. We constructed 1,186 unique LLM–human note pairs, yielding 21,978 rater–pair observations from 10,345 raters. For each observation, we encoded a win/loss/tie outcome based on whether the rater's rating for the LLM note was better than their rating for the human note.

We observed a high tie rate: 71% of all rater–pair observations resulted in equal scores for both note types, suggesting that these raters found the two notes equivalently helpful. Excluding ties and fitting a Bradley-Terry logistic model with standard errors clustered by rater, we find that the LLM note was preferred in 54.4% of non-tied head-to-head comparisons ($\beta = 0.178$, SE $= 0.035$, $z = 5.10$, $p < 0.001$; OR $= 1.19$), significantly above the 50% chance level.

B. Full-Sample Note-Level Outcomes

Table 4 presents note-level analysis of note helpfulness scores, CRH status, and CRNH status in the full sample. A higher proportion of human notes achieved CRH status than LLM notes (18.02% vs. 13.07%; AI coef. $= -0.058$, $z = -4.556$, $p < 0.001$). And among CRH notes, human notes reached that status faster (median: 5.90 hours vs. 7.38 hours for LLM notes; Mann-Whitney $U = 27{,}979$, $p = 0.054$). However, LLM notes had a lower rate of Currently Rated Not Helpful (CRNH) status (1.12% vs. 4.13% for human notes; AI coef. $= -0.027$, $z = -4.714$, $p < 0.001$) and similar average note helpfulness scores (mean: 0.25 vs. 0.24; AI coef. $= 0.007$, $z = 1.143$, $p = 0.253$). These results should be interpreted in light of substantial platform-specific confounds: the Community Notes algorithm used to compute these outcomes penalizes notes with few ratings, and LLM writers face a submission timing disadvantage that results in accumulating fewer ratings.

To benchmark LLM writer's performance against the distribution of individual human writers who have written community notes, we compare the CRH rate and hit rate. The CRH rate is the fraction of a writer's notes achieving CRH status, and the hit rate is defined as (CRH notes − CRNH notes) / total notes to account for notes reaching unhelpful status. Our LLM writer achieves a CRH rate of 13.07%, outranking 84.7% of human writers on CRH rate (78.4% among human writers who have written at least 30 notes), and a hit rate of 11.96%, corresponding to the 85.4th percentile among human writers (79.0% among human writers who have written at least 30 notes).

Table 4. Note-level comparisons in the full sample. CRH, CRNH, and helpfulness score are analyzed using linear mixed-effects models (LMMs); number of ratings and time to CRH are compared using two-sided Mann–Whitney U tests.

Metric	LLM	Human	Test	AI coef.	Statistic	p
N (total notes)	1,614	1,332	—	—	—	—
CRH rate	13.07%	18.02%	LMM	-0.058	$z = -4.56$	< 0.001
CRNH rate	1.12%	4.13%	LMM	-0.027	$z = -4.71$	< 0.001
Helpfulness score (N with scores: 1,243 / 1,130)	0.25	0.24	LMM	0.007	$z = 1.14$	0.253
Num. ratings, median	22	51	Mann–Whitney U	—	$U = 740{,}798$	< 0.001
Time to CRH, hrs, median (N CRH notes: 211 / 240)	7.38	5.90	Mann–Whitney U	—	$U = 27{,}979$	0.054

C. Robustness Checks for Note-Level Outcomes

In addition to the note-level analysis with equal-exposure raters, we conducted two robustness checks to address potential confounds in the note-level outcomes. To address the concern that notes with fewer ratings have over-penalized note helpfulness scores, we restrict the sample to notes with at least 30 ratings ($n = 1{,}538$ notes; LLM $= 680$, human $= 858$). After filtering, the percentage of CRH notes was no longer statistically different (LLM $= 27.94\%$, human $= 27.04\%$; AI coef. $= -0.006$, $z = -0.317$, $p = 0.751$), and LLM notes continued to show a significantly lower proportion of CRNH notes (LLM $= 0.59\%$, human $= 4.55\%$; AI coef. $= -0.035$, $z = -4.129$, $p < 0.001$). The difference in helpfulness scores increased (AI coef. $= 0.019$, $z = 2.410$, $p = 0.016$). Among CRH notes, human notes still reached CRH status more quickly (median $= 5.78$ hours vs. 7.16 hours; Mann-Whitney $U = 24{,}102$, $p = 0.098$).

To account for differential note exposure due to submission timing, we conduct creation time-matched analyses. We retain human notes created within $\pm$30, $\pm$60, or $\pm$90 minutes of the LLM note creation on the same post, under the assumption that notes written close in time have similar exposure. Because LLM and human notes are often written hours apart, matching rates were low (6.3%, 11.8%, and 15.9%, respectively) and sample sizes were substantially reduced. Within the $\pm$60-minute window ($n = 405$ notes; LLM $= 190$, human $= 215$), the CRH gap was narrow and non-significant (LLM $= 21.58\%$ vs. human $= 22.79\%$; AI coef. $= -0.019$, $z = -0.610$, $p = 0.542$). The difference in average note helpfulness scores was larger but not statistically significant (LLM $= 0.28$ vs. human $= 0.25$; AI coef. $= 0.023$, $z = 1.703$, $p = 0.089$). In this matched sample, LLM notes reached CRH status faster (median $= 4.61$ hours vs. 5.55 hours).

Tables 5 and 6 present the full note-level statistics for the $\geq$30-ratings subset and timing-matched subsets.

Table 5. Note-level comparisons in the subset of notes with at least 30 ratings.

Metric	LLM	Human	Test	AI coef.	Statistic	p
N (total notes)	680	858	—	—	—	—
CRH rate	27.94%	27.04%	LMM	-0.006	$z = -0.32$	0.751
CRNH rate	0.59%	4.55%	LMM	-0.035	$z = -4.13$	< 0.001
Helpfulness score (N with scores: 645 / 798)	0.31	0.28	LMM	0.019	$z = 2.41$	0.016
Num. ratings, median	67.00	91.50	Mann–Whitney U	—	$U = 236{,}530.5$	< 0.001
Time to CRH, hrs, median (N CRH notes: 190 / 232)	7.16	5.78	Mann–Whitney U	—	$U = 24{,}102$	0.098

(a) $\pm$30-minute window. Matched 102/1,614 LLM notes (6.3%).

Metric	LLM	Human	Test	AI coef.	Statistic	p
N (total notes)	102	116	—	—	—	—
CRH rate	21.57%	20.69%	LMM	0.004	$z = 0.09$	0.929
CRNH rate	1.96%	6.90%	LMM	-0.049	$z = -1.74$	0.082
Helpfulness score (N with scores: 96 / 109)	0.28	0.24	LMM	0.041	$z = 2.13$	0.033
Num. ratings, median	64.00	55.00	Mann–Whitney U	—	$U = 6{,}491.5$	0.216
Time to CRH, hrs, median (N CRH notes: 22 / 24)	6.01	5.67	Mann–Whitney U	—	$U = 288$	0.605

(b) $\pm$60-minute window. Matched 190/1,614 LLM notes (11.8%).

Metric	LLM	Human	Test	AI coef.	Statistic	p
N (total notes)	190	215	—	—	—	—
CRH rate	21.58%	22.79%	LMM	-0.019	$z = -0.61$	0.542
CRNH rate	2.11%	6.05%	LMM	-0.039	$z = -1.98$	0.048
Helpfulness score (N with scores: 178 / 193)	0.28	0.25	LMM	0.023	$z = 1.70$	0.089
Num. ratings, median	54.00	57.00	Mann–Whitney U	—	$U = 21{,}064.5$	0.587
Time to CRH, hrs, median (N CRH notes: 41 / 49)	4.61	5.55	Mann–Whitney U	—	$U = 950$	0.662

(c) $\pm$90-minute window. Matched 257/1,614 LLM notes (15.9%).

Metric	LLM	Human	Test	AI coef.	Statistic	p
N (total notes)	257	313	—	—	—	—
CRH rate	19.46%	20.13%	LMM	-0.013	$z = -0.47$	0.638
CRNH rate	1.56%	5.43%	LMM	-0.037	$z = -2.37$	0.018
Helpfulness score (N with scores: 240 / 285)	0.27	0.24	LMM	0.027	$z = 2.37$	0.018
Num. ratings, median	57.00	59.00	Mann–Whitney U	—	$U = 40{,}772$	0.778
Time to CRH, hrs, median (N CRH notes: 50 / 63)	4.96	5.46	Mann–Whitney U	—	$U = 1{,}562$	0.942

Table 6. Note-level comparisons in timing-matched subsets. Each sub-table restricts analysis to LLM notes matched to human notes within the indicated submission-time window.

D. Representativeness of Complete Raters

We compare the distribution of rater characteristics between the full rater population and the subset of "complete raters" who evaluated all notes on a given tweet. The left panel shows the distribution of coreRaterIntercept, which measures a rater's baseline tendency to rate notes as helpful. The right panel shows the distribution of coreRaterFactor1, which captures political leaning inferred from historical rating patterns (negative = left-leaning, positive = right-leaning). Both distributions are highly similar across the two groups, suggesting that complete raters do not systematically differ from the overall population in helpfulness leniency and political leaning (Figure 4). Together, these comparisons indicate that restricting the analysis to raters who rated all notes on a tweet does not substantially alter the distribution of key rater characteristics.

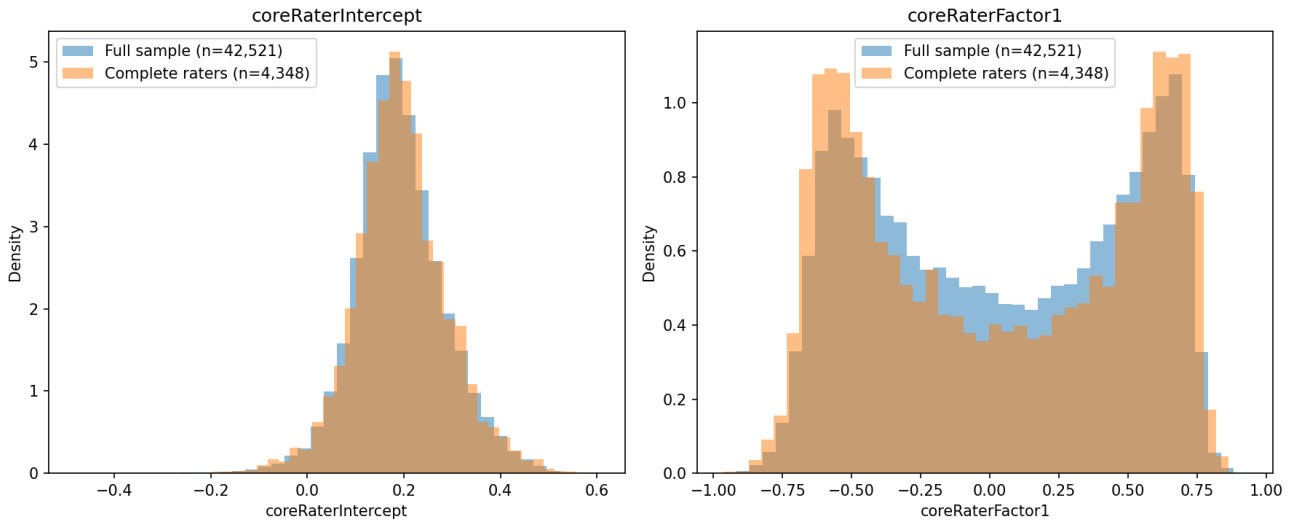

Figure 4. Distribution of rater characteristics for the full rater population vs. complete raters who evaluated all notes on a given tweet. Left: coreRaterIntercept captures baseline helpfulness leniency. Right: coreRaterFactor1 captures political leaning (negative = left-leaning, positive = right-leaning). The close overlap indicates that complete raters are not systematically different from the overall rater population.

E. LLM Writing Pipeline Prompts

E.1. Evidence Retrieval Prompt

```
"""Investigate whether the X post below is misleading. Images or video previews
from the post are provided if they exist.

Step 1: Understand the post and its context
- Find the post by post id {post_id} on X. If not found, search the post's main
  text on X.
- Identify the post author. Also note any signals that the account is
  parody/satire (e.g. bio, handle, verification).
- Fetch the thread and top comments. Understand how others interpret the post
  (e.g. joke/satire, potentially misleading, expressing an opinion).
- Summarize the post context in 1-2 sentences without rewording the post itself.
  Include who the OP is and how others interpret the post, especially if comments
  suggest it's a joke or satire, or provide potential fact-check directions.
  For example: "The OP is Sen. X. Many comments say his claim about X is
  unverified." or "The OP is a parody account, and commenters are laughing and
  treating it as a joke."

Step 2: Search for evidence
1. Search both the web and X for factual sources that refute or confirm the
   post's claims. Use the post context to guide your search if it could provide
   potential fact-check directions.
2. Aim for {target_url_count} pieces of evidence / URLs if possible.
3. For each source, include the URL and a brief note describing how it verifies
   or challenges the post. Include the publication date of the source if
   available.
4. Cover outlets across the ideological spectrum (left, center, right).
   Overlapping reasoning is acceptable when it comes from different publishers.
5. Prioritize evidence that is relevant, solid, and up to date.

Target post (ID: {post_id}):
{post}

Your response should be returned as a JSON object with the following structure:
```
{{
```
```

```
   "post_context": "one/two-sentence summary of the post context",
   "research": [
       {{"url": "url1", "description": "how the content of the URL fact-checks the post
    "}},
       ...
   ]
}}
```
If you cannot find sufficient evidence to fact-check the post, return an empty
research array."""

E.2. Note Triage Prompt

System prompt:

```
"""Decide whether a post needs a Community Note based on the provided evidence.

Returns:
    - "WRITE NOTE" if a note should be written
    - "NO NOTE NEEDED" if the post doesn't need a note
    - "NOT ENOUGH EVIDENCE" if there's insufficient evidence
"""
```

User prompt:

```
"""You are a Community Notes writer. Your job is to decide if the target post
could be perceived as misleading and whether it needs a community note to address
its issues. The output should be one of: "WRITE NOTE", "NO NOTE NEEDED", or
"NOT ENOUGH EVIDENCE".

Task rules
- Focus on the main claims of the post, not trivial errors.
- If media is included, use any legible text, recognizable logos/landmarks, and
  clearly identifiable public figures as part of the claim.
- If the target post is a quote/reply, evaluate the target post, using the
  quoted/replied post only as context.
- Use the context information (author info and audience reactions) to inform your
  decision. They are helpful for understanding whether the post is likely to be
  perceived as misleading.
- If you are unsure, err on the side of NO NOTE NEEDED.

Decision logic
Output NO NOTE NEEDED if:
- The OP is a satire/parody account, or the post is joking/sarcastic/exaggerated
  to be ironic, and commenters are interpreting it that way (with no strong signs
  of misunderstanding); or
- The post is mostly opinion, subjective takes, or personal experience; or
- The post contains no major factual claims, or the claims are not verifiable; or
- The post contains factual claims, but the provided evidence indicates those
  claims are accurate or not meaningfully misleading.

If the post contains major factual, verifiable claims and the evidence is
relevant:
- Output NOT ENOUGH EVIDENCE if the evidence is weak, mixed, or insufficient to
  confidently verify or refute the main claims. When unsure about the
  strength/sufficiency of the evidence, err on NOT ENOUGH EVIDENCE.
- Output WRITE NOTE only if the evidence clearly shows that the post's main
  claims are false or misleading in a way that could misinform a reasonable
  reader.

Post:
{post}

Additional context information about the post:
```

```
{post_context}

Evidence:
{evidence}

Output only one of: "WRITE NOTE", "NO NOTE NEEDED", or "NOT ENOUGH EVIDENCE".
Err on the side of NO NOTE NEEDED if unsure."""
```

E.3. Note Generation Prompt

System prompt:

```
"""You are a helpful fact-checking assistant.
Your goal is to write good Community Notes that would be approved helpful by
people with different viewpoints.
Do not invent facts or make claims that are not supported by the provided
evidence."""
```

User prompt:

```
"""Task: Write a community note for the target post below. Images or video
previews from the target post may be provided; if present, analyze any legible
text (OCR), recognizable logos/landmarks, and confidently identifiable public
figures to inform the note. Additional context provides post author details and
audience reactions. If the target post quotes/replies to another post, use it
only for context and focus on the target post.

Hard Constraints:
1. The note is written to explain why the post is misleading and add additional
   context to the post. Focus on primary claim(s) of the post rather than
   trivial details.
2. The note must be grounded in the provided evidence and should cite the URL of
   the evidence it uses. At least one URL must be cited.
3. Keep the note strictly under 280 characters. Stay neutral and clear.
4. No hashtags, emojis, unnecessary words. No markdown, brackets, or parentheses
   around URLs. Do not mention "this note" or "the prompt."

Target post:
```
{post}
```

Additional context about the post:
```
{post_context}
```

Allowed evidence sources:
```
{evidence}
```

Output only the final note (at most 280 characters)."""
```

E.4. Topic Classification Prompt

```
"""Classify the following X/Twitter post into exactly one of these categories:

- Politics & Elections: U.S./global elections, politicians, voting, partisan
  claims, policy debates.
- Geopolitics & International Conflicts: Wars, foreign policy, terrorism,
  diplomacy, country-specific events (e.g., Ukraine, Israel-Palestine,
```

```
  China-Taiwan).
- Health & Medicine: Diseases, treatments, vaccines, public health policies,
  medical advice, COVID.
- Social/Cultural Issues: Abortion, gender/LGBTQ+, race/DEI, guns,
  crime/justice, education, religion.
- Economy & Finance: Inflation, jobs, taxes, crypto, markets, inequality.
- Science & Technology: Climate change, AI, space, gadgets (non-health tech).
- Conspiracy & Pseudoscience: General conspiracies, election fraud claims not
  tied to active politics, QAnon-style, flat earth, etc.
- Celebrity / Entertainment / Viral: Non-political hoaxes, celebrity drama.
- Sports: Sports events, sports scandals, and relevant discussion.
- AI generated content: AI generated / modified content.
- Other / Miscellaneous: Everything else (weather, personal, ads, neutral news
  without controversy).

Post text:
{tweet_text}
{context_section}
Answer with only the exact category name from the list above
(e.g., "Politics & Elections" or "Other / Miscellaneous")."""
```