Request a Note: How the Request Function Shapes X's Community Notes System

Request a Note: How the Request Function Shapes X's Community Notes System

YUWEI CHUAI, University of Luxembourg, Luxembourg
SHUNING ZHANG, Tsinghua University, China
ZIMING WANG, University of Luxembourg, Luxembourg and Chalmers University of Technology, Sweden
XIN YI, Tsinghua University, China
MOHSEN MOSLEH, University of Oxford, United Kingdom
GABRIELE LENZINI, University of Luxembourg, Luxembourg

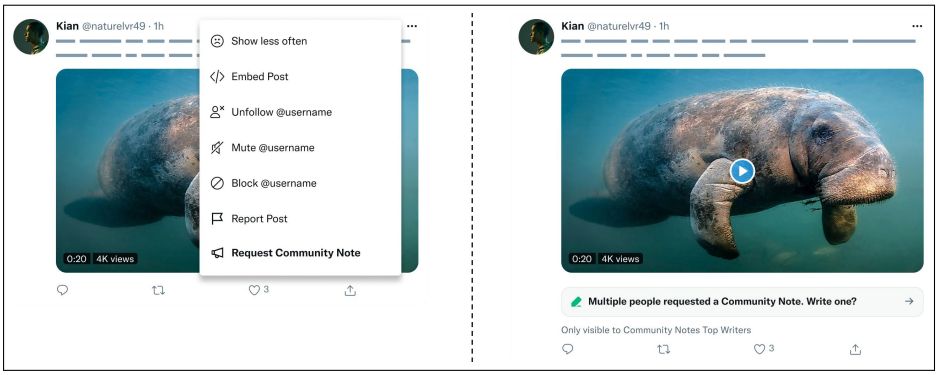

Fig. 1. The introduction of "Request Community Note" feature on 𝕏, sourced from the official Community Notes website [43]. It allows users to request a note on a post they believe would benefit from additional context (see the left part of the figure). When a sufficient number of requests are submitted, a subset of Community Notes contributors (i.e., top writers) will see an alert below the corresponding requested post, and can choose to propose a note (see the right part of the figure).

X's Community Notes is a crowdsourced fact-checking system. To improve its scalability, X recently introduced "Request Community Note" feature, enabling users to solicit fact-checks from contributors on specific posts. Yet, its implications for the system—what gets checked, by whom, and with what quality—remain unclear. Using 98,685 requested posts and their associated notes, we evaluate how requests shape the Community Notes system. We find that contributors prioritize posts with higher misleadingness and from authors with greater misinformation exposure, but neglect political content emphasized by requestors. Selection also diverges along partisan lines: contributors more often annotate posts from Republicans, while requestors surface more from Democrats. Although only 12% of posts receive request-fostered notes from top contributors, these notes are rated as more helpful and less polarized than others, partly reflecting top contributors' selective fact-checking of misleading posts. Our findings highlight both the limitations and promise of requests for scaling high-quality community-based fact-checking.

CCS Concepts: • **Human-centered computing → Social media**; **Empirical studies in collaborative and social computing**; • **Information systems → Crowdsourcing**; • **Security and privacy → Human and societal aspects of security and privacy**.

Additional Key Words and Phrases: Social media, misinformation, crowdsourced fact-checking, Community Notes, request function

Authors' Contact Information: Yuwei Chuai, University of Luxembourg, Luxembourg, yuwei.chuai@uni.lu; Shuning Zhang, Tsinghua University, China, zsn23@mails.tsinghua.edu.cn; Ziming Wang, University of Luxembourg, Esch-sur-Alzette, Luxembourg and Chalmers University of Technology, Gothenburg, Sweden, ziming@chalmers.se; Xin Yi, Tsinghua University, China, yixin@tsinghua.edu.cn; Mohsen Mosleh, University of Oxford, United Kingdom, mohsen.mosleh@oii.ox.ac.uk; Gabriele Lenzini, University of Luxembourg, Luxembourg, gabriele.lenzini@uni.lu.

1 Introduction

Designing effective countermeasures to identify falsehoods and curb the spread of misleading posts is still a challenge. Years ago, the social media platform 𝕏 (formerly Twitter) adopted *Community Notes*, a crowdsourced fact-checking system designed to address misleading posts and provide users with helpful additional context. This initiative marks the first large-scale use of community-based fact-checking by a major social media platform [5, 14]. A subset of users—called *contributors*—are periodically admitted into the program if they meet specific eligibility requirements, such as having no recent violations of 𝕏's rules. Contributors begin by rating the helpfulness of existing notes, and only after demonstrating reliability can they unlock the ability to write notes themselves. All notes are evaluated by a note selection algorithm, and then those deemed helpful by the algorithm are displayed directly on the relevant posts [44]. Inspired by this program, other social media providers—including YouTube, Meta, and TikTok—are beginning to explore similar community-based moderation models [28, 49, 59]. For example, Meta is piloting the Community Notes program across Facebook, Instagram, and Threads in the U.S., while TikTok plans to offer a similar feature called Footnotes to provide context and corrections for misleading posts. Given the rapid expansion of community-based fact-checking system, it is crucial to understand how the system functions in the real world.

Recent work has shown that 𝕏's Community Notes program can reliably attach helpful notes to the corresponding misleading posts [3, 58]. Moreover, community fact-checks are perceived as more trustworthy than expert fact-checks, can effectively reduce the spread of misleading posts, and even prompt authors to delete their problematic content [12, 13, 22, 58]. However, the Community Notes system, in its current implementation on 𝕏, still faces limitations, such as the small number of displayed notes relative to all generated notes and restrictions on who can author them [12].

To foster note writing and create a more interactive and responsible community, 𝕏 recently implemented a key innovation in Community Notes system—the *Request Community Note* feature (as illustrated in Fig. 1), allowing broader registered users to signal posts they believe would benefit from additional context. This mechanism enables non-contributor users to participate in the moderation process while guiding contributors toward posts where notes are likely to be needed. This request feature is expected to increase the involvement and the interest of users in checking facts, which is crucial in the current media landscape, where misinformation can spread rapidly. However, it remains unclear whether this mechanism effectively improves the Community Notes system by helping more misleading posts receive notes and enhancing the quality of the resulting notes.

Research questions. In this study, we examine the newly introduced Request Community Note feature on 𝕏, tracing the process from the submission of requests to the evaluation of the associated notes. The workflow involves requestors flagging posts (i. e., *requested posts*) for note writers, note writers selecting requested posts to propose community notes, and the note selection algorithm evaluating the generated notes. Notably, note writers retain autonomy: they can independently choose any post to fact-check and propose community notes, regardless of whether they see request alerts or not. Moreover, the note selection algorithm evaluates community notes along two key dimensions: *helpfulness*, indicating the consensus of perceived helpfulness among heterogeneous contributors, and *polarization*, capturing the degree of variation in ratings. Accordingly, we organize our investigation around three research questions to explore how note writers select requested posts to fact-check, the extent to which their selections are influenced by requests, and how community notes fostered by user requests (i. e., *request-fostered notes*) differ from other notes in helpfulness and polarization according to the note selection algorithm:

- **RQ1:** Among requested posts, which ones actually receive community notes, and how do their content and author characteristics shape this likelihood?

- **RQ2:** To what extent do user requests foster the creation of community notes, and how do they influence contributors' selection of posts?
- **RQ3:** Are community notes fostered by user requests more or less helpful and polarized than other notes, and what factors explain these differences?

Data and methods. To address our research questions, we collect all available Community Notes data and associated posts that had received at least five requests as of May 20, 2025. We then specify a logistic regression model to estimate the likelihood of a post receiving a community note based on a wide range of characteristics extracted from posts, such as topics and estimated misleadingness, and their authors, such as verified status and misinformation exposure scores [38] (RQ1). In addition, we analyze the extent to which requests foster the creation of community notes and how requests can influence the choices of note writers (RQ2). Finally, we replicate the note selection algorithm to evaluate request-fostered notes compared to other notes and explore factors affecting the evaluation results (RQ3).

Contributions. Overall, our study results in three main findings. (i) The likelihood that a requested post receives a community note is associated with the post's content and its author. For instance, posts related to entertainment or finance, with higher estimated misleadingness, or from accounts with higher misinformation exposure scores, are more likely to receive community notes. Our findings suggest that Community Notes contributors and requestors may have different selection patterns. (ii) With more than half (53.6%) of requested posts receiving community notes, only an estimated 12.1% of those are likely fostered by user requests, showing that the effect of requesting community notes in fostering note generation is, so far, limited. Nonetheless, requests tend to shift contributors' attention toward political content, while still prioritizing posts with higher misleadingness (as measured by GPT-4.1 from OpenAI). (iii) Request-fostered notes generated by top writers are evaluated as more helpful and less polarized than other notes, a pattern that likely reflects writers' selective prioritization of requested posts with higher levels of misleadingness. To our knowledge, our study provides the first empirical evaluation of this novel request mechanism, offering insights into both the opportunities and challenges of scaling crowdsourced fact-checking systems.

2 Background and Related Work

The spread of misleading posts on social media platforms—including 𝕏 (formerly Twitter), TikTok, and Facebook—remains a significant societal problem, with far-reaching and often harmful consequences [23, 30]. For example, the widespread misinformation during political elections is particularly alarming, as exposure to and belief in false narratives during elections can distort public opinion, erode trust in democratic institutions, and destabilize societies [4, 23, 27, 56]. Beyond electoral contexts, misinformation can exacerbate social divisions and intensify polarization in response to global pandemics and claim change [23, 24]. Given these multifaceted risks, social media providers face increasing pressure to adopt transparent content moderation policies and implement effective and scalable interventions to identify misleading posts, correct falsehoods, and ensure users have access to reliable information [9, 20, 21, 30].

2.1 Misinformation Identification and Fact-Checking

To identify misinformation, traditional human-driven fact-checking often replies on professional experts and third-party organizations. However, the scarcity of expert fact-checkers, coupled with the relentless pace of online content, hampers the timely and scalable implementation of professional fact-checking [5, 14]. Furthermore, despite its depth and rigor, professional fact-checking has faced criticism, as experts are sometimes perceived as biased in selecting which claims to review—raising concerns about agenda-setting and potentially eroding public trust in their assessments [16, 31]. To address the limitations of professional fact-checking, crowdsourced (or community-based) fact-checking approaches

have been proposed, leveraging the wisdom of crowds to achieve broader and faster coverage of online content [1]. Although individual assessments may be subject to bias and noise [2], aggregated judgments—even from relatively small crowds—have been shown to be reliable and comparable in accuracy to expert evaluations [6, 25, 47, 52]. Moreover, crowdsourced assessments have the potential to mitigate public trust issues associated with expert fact-checks [1, 22, 60].

Meanwhile, automatic fact-checking powered by machine learning models continues to develop, offering increasingly sophisticated tools for identifying misinformation. Traditional detection models are typically trained on surface-level content or contextual features such as linguistic patterns, semantic similarity, propagation structures, or user metadata [10, 14]. Although these approaches have proven useful in identifying repeated misinformation or content with distinctive stylistic markers, particularly in constrained domains or benchmark datasets, they often struggle with more subtle, novel, or context-dependent cases. In particular, such models lack the ability to reason over evidence, integrate external knowledge sources, or adapt quickly to evolving narratives. As a result, their performance in real-world fact-checking scenarios often falls short, especially when confronted with emerging claims that deviate from the data on which they were trained [26].

Recent advances in Large Language Models (LLMs), such as OpenAI's GPT series, introduce novel opportunities to overcome the shortcomings of traditional machine learning models in fact-checking. LLMs exhibit strong contextual understanding and can simulate human-like reasoning processes, allowing them to evaluate claims not only on textual cues but also in relation to broader world knowledge. They can flexibly incorporate external evidence through retrieval-augmented generation, adapt to diverse domains with minimal task-specific training, and generate explanations that improve transparency in the fact-checking process [10, 57]. Previous work suggests that LLMs can achieve human-comparable performance in misinformation detection tasks [11, 17, 19]. For instance, ChatGPT has been shown to achieve an accuracy of approximately 90% in identifying false headlines, demonstrating the potential of LLMs to support large-scale misinformation detection [19]. At the same time, LLM-based fact-checking is not without challenges. LLMs tend to adopt a more structured evaluation strategy, whereas human annotators often display greater variability in their use of evaluative criteria, particularly for borderline or ambiguous claims [17]. More critically, their influence on user perception can be harmful: when LLMs mislabel true claims as false, they risk reducing belief in accurate information; conversely, when they express uncertainty about false claims, they may inadvertently increase belief in misinformation [19, 35]. These limitations underscore that, while LLMs perform strongly in detecting misinformation, human judgment remains essential for deciding what content to surface to users, helping to mitigate unintended harms.

2.2 Crowdsourced Fact-Checking in Practice: X's Community Notes

Given the necessity of human-centered approaches and the demonstrated promise of crowdsourced fact-checking, 𝕏 introduced Community Notes, a system that engages diverse users in collaboratively annotating potentially misleading content at scale [44]. First, Community Notes contributors can append contextual information or corrections to potentially misleading posts, offering alternative perspectives or clarifications. Contributors are required to cite external sources in their community notes. These notes are continuously evaluated through peer ratings from other contributors. Additionally, 𝕏 develops a bridging algorithm to feature objectively informative community notes that are subjectively perceived to be helpful across heterogeneous contributors [58]. Beyond contributors, all users can now participate indirectly through the recently launched Request Community Note function, which allows any account with a verified phone number to submit requests for posts they believe would benefit from a community note. When enough requests are submitted, *top writers* can see alerts highlighting these posts and can choose to write a note in response (see the illustrations in Fig. 1). This division of roles—between note writers, note raters, and requestors—illustrates how

Community Notes distributes responsibilities across different layers of the user base, thereby expanding participation beyond the core contributor group.

Research has shown that Community Notes system can successfully identify misleading content, reduce its spread, and even pressure authors to delete problematic posts [3, 12, 37, 58]. Additionally, community-based annotations are often perceived as more trustworthy than professional fact-checks [22]. These encouraging findings suggest that Community Notes is promising in addressing misinformation on social media platforms and have sparked growing scholarly interest in understanding the dynamics of the system and identifying opportunities to refine and enhance its design [5]. For instance, despite the relatively large volume of notes generated by contributors, only a small fraction—around 10%—achieve sufficient support from diverse users and are displayed to users [12, 18]. At the same time, the display process is often not fast enough to intervene before misleading posts have already begun to spread widely [12, 14]. To address these limitations, recent work has explored two complementary directions.

First, researchers and practitioners have turned to technological augmentation, using Large Language Models (LLMs) to help scale human judgment by assisting contributors in drafting and refining notes [32]. For example, the Supernotes approach generates AI-synthesized summaries that integrate insights from multiple existing contributor notes with the goal of fostering consensus among diverse raters [18]. Second, 𝕏 has expanded user participation by introducing the request function, which allows any user with a verified account to flag posts that they believe should receive a community note. This structural expansion extends the influence of the wider user base beyond the core contributor group, potentially increasing the coverage and responsiveness of the system. However, despite its significance, this new function has not yet been systematically studied, leaving open questions about its effects on content selection for fact-checking, note generation, and the quality of resulting notes. Addressing these questions represents the primary goal of this study.

3 Data and Methods

In this section, we first describe our data collection process, including the posts, requests, and associated community notes. We then outline the post content and author characteristics extracted for analysis, including GPT–based evaluations of post misleadingness as a proxy for misinformation risk. Next, we examine external source domains cited in community notes, assessing them in terms of information quality and political bias. Since request alerts are only visible to top writers, we also identify top note writers following the criteria provided on the official Community Notes website. Finally, we replicate the note selection algorithm to evaluate the helpfulness and polarization of notes, enabling us to compare request-fostered notes with those generated independently of requests.

3.1 Data Collection

To address our research questions, we downloaded the complete set of available Community Notes data—including requests, notes, ratings, and note status histories—on May 20, 2025, when 𝕏 began releasing request data to the public [40]. Specifically, we collected 5,888,351 requests submitted by 1,689,152 unique users since the launch of the request feature on July 18, 2024. In addition, the dataset includes 1,787,609 notes that have received 149,646,500 ratings since the introduction of Community Notes program.

Subsequently, we used the full-archive search endpoint of the 𝕏 Pro API to collect the requested posts. In total, the submitted requests target 2,427,451 unique posts. However, according to the platform guidelines at the time of data collection, a post must receive at least 5 requests to be surfaced to top writers, indicating a threshold for which posts are eligible for community evaluation [43]. In our dataset, 154,090 posts meet this criterion, representing only 6.3%

Table 1. Overview of requested posts. The three columns are for all requested posts, posts with community notes, and posts without community notes, respectively. Reported are mean values or count numbers for the variables (standard deviations in parentheses).

	(1) All	(2) With notes	(3) Without notes
# Requested posts	98,685	52,862 (53.6%)	45,823 (46.4%)
# Post authors	22,915	15,575	12,779
Post content characteristics			
Sentiments: Positive	0.131 (0.244)	0.134 (0.246)	0.127 (0.241)
Sentiments: Negative	0.460 (0.335)	0.448 (0.335)	0.474 (0.335)
Topics: Politics	37.0%	32.4%	42.4%
Topics: Science & Technology	13.5%	12.8%	14.2%
Topics: Entertainment	26.9%	28.0%	25.6%
Topics: Finance & Business	32.6%	33.9%	31.2%
Content type: Claim	0.666 (0.309)	0.664 (0.315)	0.667 (0.302)
Content type: Opinion	0.522 (0.335)	0.507 (0.336)	0.540 (0.333)
GPT misleadingness	0.428 (0.317)	0.435 (0.323)	0.421 (0.310)
Media	65.4%	68.4%	62.0%
Post author characteristics			
Account type: Blue	72.5%	72.3%	72.6%
Account type: Business	6.6%	6.4%	6.7%
Account type: Government	7.2%	6.5%	7.9%
Account age	3,156.145 (1,997.934)	3,078.018 (1,998.710)	3,246.273 (1,993.259)
Followers	6,828,060	6,561,490	7,135,578
Followees	13,347	11,081	15,960
Misinformation exposure score	0.546 (0.139)	0.555 (0.143)	0.536 (0.135)
Partisan score	−0.035 (0.767)	0.023 (0.768)	−0.095 (0.761)

of all requested posts. However, the requests submitted to these posts account for more than half (52.6%) of the total requests. This suggests that, while users submitted a large volume of requests, attention was concentrated on a small set of posts. Based on the corresponding IDs of these eligible posts, we successfully retrieved 133,351 posts, of which 98,685 (74%) are in English and were authored by 22,915 unique accounts. We restrict our analysis to the collected posts in English and focus on the request feature in the U.S. context. Next, we introduce how to extract the characteristics of posts and their authors for our subsequent analysis (see summary statistics in Table 1).

3.2 Characteristics of Post Content

3.2.1 Sentiments. Sentiments are an important factor that influences the dissemination of online content [15, 51]. We extract sentiments in the posts content using a state-of-the-art sentiment classification model published on HuggingFace. The predicted sentiment categories include positive, neutral, and negative. Specifically, we compute the probabilities of positive and negative sentiments for each collected post using the TwitterroBERTa-base model (2022 updated). It was trained on 124 million posts created from January 2018 to December 2021, and fine-tuned for sentiment analysis. This model achieves superior predictive performance compared to other similar sentiment models [34]. We denote the probabilities of positive sentiment and negative sentiment in each post as *Positive* and *Negative*, respectively. On average, requested posts exhibit significantly higher negative sentiment (mean of 0.460) than positive sentiment (mean of 0.131; $t = 199.208$, $p < 0.001$).

3.2.2 Topics. Online misinformation often leverages sentiments to enhance its virality and tends to concentrate on specific topics—most notably, e. g., politics [15, 56]. 𝕏 has provided domain labels for each post in the collected dataset. Following prior work [14], we extract these domains from the requested posts and group them into four broad topics: Politics, Science & Technology, Entertainment, and Finance & Business. Notably, a single post may be associated with multiple topics. Politics (37%) is the most frequent topic in the requested posts, followed by Finance & Business (32.6%), Entertainment (26.9%), and Science & Technology (13.5%).

3.2.3 Content types and misleadingness. We distinguish between *factual claims*, which can be independently verified, and *personal opinions*, which are more subjective content to express users' personal feelings. This distinction is important for understanding how requests and community notes may prioritize verifiable content over opinionated posts. To operationalize this, we use GPT-4.1 to classify requested posts as factual claims or personal opinions. In addition, prior work suggests that LLMs can achieve a high performance in identifying misleading information, which is comparable to human annotators [11, 17, 19]. Building on this, we employ GPT-4.1 to assess the misleadingness of collected posts. We provide task-specific prompts to the model and manually validate its outputs through a subset of random posts to ensure reliability (see details in Suppl. S1). This approach allows us to analyze both content types and the likelihood of misinformation in the requested posts. The requested posts are more likely to be claims (mean of 0.666), compared to opinions (mean of 0.522; $t = 84.797$, $p < 0.001$). Additionally, the GPT-estimated misleadingness is, on average, 0.428 in the requested posts.

3.2.4 Media posts. Media elements, such as images or videos, can increase user engagement with associated posts, making media-based misinformation potentially more viral and harmful than text-only content [14, 46]. Additionally, the Community Notes team on 𝕏 implemented a media note feature, which allows helpful media notes to be displayed across all posts that contain the same media [41]. This feature can help curb the spread of misleading posts at the very beginning if they contain the same misleading media annotated by existing community notes, thereby increasing the responsiveness of the Community Notes system. By analyzing the media keys returned by the 𝕏 Pro API, we find that more than half of all requested posts contain media (65.4%).

3.3 Characteristics of Post Authors

3.3.1 Built-in characteristics of post authors. We analyze several built-in features of the post authors obtained from the 𝕏 Pro API. The dataset includes in total 22,915 unique post accounts who authored the requested posts. Each author account is categorized as either verified or unverified. Verified accounts are categorized as *Blue*, *Government*, or *Business*: Blue accounts correspond to individual 𝕏 Premium subscribers, while government and business accounts are officially verified by 𝕏. We find that the requested posts are predominantly authored by blue accounts (72.5%), followed by unverified (13.8%), government (7.2%), and business (6.6%) accounts. In addition, we consider the number of followers (mean of 6,828,060; median of 326,732), the number of followees (mean of 1,334,676; median of 1180), and the account age, defined as the number of days from the account creation to the creation of the corresponding post. The average of the account ages in the requested posts is 3156 days, i. e., ~ 9 years. These features provide insights into the authors' influence, reach, and longevity on the platform.

3.3.2 Misinformation exposure score and partisan score. Previous work developed a method to calculate users' exposure to misinformation from political elites on 𝕏 [38]. They assigned each elite (public figures and organizations) a "falsity score" based on the veracity of their statements, as determined by professional fact-checkers like PolitiFact. Users'

misinformation exposure scores were then computed by averaging the falsity scores of the elites they followed on 𝕏. This approach allows researchers to assess the relationship between users' exposure to elite misinformation and their sharing behaviors, as well as their estimated political ideologies. We estimate the political partisanship and misinformation exposure of the relevant post authors on 𝕏 using the API service provided by this study. Partisanship scores range from −1 (Democrat) to 1 (Republican) and are based on the number of Democratic and Republican public figures followed by each user. The misinformation exposure score, ranging from 0 to 1, reflects the proportion of followed public figures rated as false by PolitiFact. Out of 22,915 post authors, we successfully collect misinformation exposure scores and partisan scores for 10,492 (45.8%) authors. These scores, together with other post and author characteristics, provide key context for analyzing which posts are more likely to attract community notes and how authors' influence and ideological orientations affect note generation. Summary statistics for the characteristics related to posts and their authors are reported in Table 1.

3.4 External Domain References in Community Notes

We extract the external domains cited in the community notes based on string matching. To assess the quality and political bias of external domains referenced in the community notes, we rely on domain ratings compiled by Lin et al. [33]. This study aggregated six sets of expert news domain ratings using an ensemble approach that combined imputation and principal component analysis, resulting in a comprehensive set of quality scores for 11,520 domains. The aggregated ratings provide a reliable measure of news quality, allowing us to evaluate the credibility of the information cited in community notes. In addition, we assess the political bias of external domains using Media Bias/Fact Check, a widely used online resource covering over 10,000 media sources, journalists, politicians, and countries [54]. Analogous to previous research [29, 54], we map bias categories to scores ranging from −1 to 1. Specifically, the "Least Biased" and "Pro-Science" are coded as 0. The "Left" and "Right" are coded as −1 and 1, respectively. The "Left-Center" and "Right-Center" are coded as −0.5 and 0.5, respectively. Finally, we successfully assign domain quality scores and domain bias scores to external domains in 367,378 community notes, accounting for 20.6% of the total notes. Together, these measures allow us to evaluate both the reliability and ideological leaning of the information cited in the community notes.

3.5 Identification of Top Writers

Since only top note writers can view request alerts once the associated posts reach the request threshold [43], it is essential to identify these top writers in order to evaluate whether requests actually drive note writing and coverage. The dataset, however, does not directly indicate top-writer status. To address this, we approximate the top writers using the criteria provided on the official Community Notes website: contributors must have at least 4% of their notes rated "Helpful" [45]. Based on this criterion, we identify 56,113 top note writers, representing 22.2% of all note writers (433,936). Notably, the recognition of top-writer status also requires a writing impact score of 10 or higher, but this metric is not publicly available. Despite this limitation, our approach offers a conservative estimate that avoids underestimating the potential effect of requests on note generation.

3.6 Replication of Note Selection Algorithm

The Community Notes program employs a matrix-factorization-based approach to identify annotations that resonate across heterogeneous user groups [58]. We replicate this note selection algorithm to evaluate both the helpfulness and polarization of community notes [42]. The algorithm produces two key estimates: the *note intercept*, which captures

overall helpfulness indicating how broadly a note is judged to be useful across raters; and the *note factor*, which, by contrast, measures polarization reflecting the extent to which contributor ratings diverge. Together, these metrics enable us to assess not only the quality of community-generated fact-checks but also the degree of consensus surrounding them. To validate our replication, we compare the note statuses generated by the reproduced algorithm with the actual production statuses. Each community note can receive one of the three statuses: Currently Rated Helpful, Needs More Ratings, and Currently Rated Not Helpful. Overall, 99.7% of replicated statuses match their production counterparts, demonstrating the reliability of our reproduction.

4 Empirical Results

Here, we report our empirical results, structured around the three research questions. We begin by examining what types of requested posts are more likely to receive community notes (RQ1), then access the extent to which requests foster note writing (RQ2), and finally compare the quality of request-fostered notes with other notes in terms of helpfulness and polarization (RQ3).

4.1 Likelihood of Receiving Community Notes

Of all the requested posts, 52,862 (53.6%) received at least one community note. To investigate the factors associated with this outcome (RQ1), we estimate a logistic regression model predicting the likelihood that a requested post receives a community note. The estimation results are presented in Fig. 2.

Post content characteristics. We analyze how post content characteristics shape the likelihood of receiving community notes. For sentiments, we find that posts expressing stronger negative sentiment are significantly less likely to receive community notes ($coef. = -0.077$, $p < 0.01$). For topics, we find that posts related to Politics ($coef. = -0.334$, $p < 0.001$) and Science & Technology ($coef. = -0.117$, $p < 0.001$) are less likely to receive community notes compared to posts not associated with these topics. For example, political posts have 28.4% lower odds of receiving community notes compared to non-political posts. In contrast, posts related to Entertainment ($coef. = 0.086$, $p < 0.001$) and Finance & Business ($coef. = 0.073$, $p < 0.001$) are more likely to receive community notes compared to posts not associated with these topics. Additionally, posts containing media are more likely to receive community notes compared to posts without media ($coef. = 0.205$, $p < 0.001$), corresponding to a 22.7% increase in the odds of receiving community notes. For claims vs. opinions, we find that the probability that a post contains a claim is not significantly associated with its likelihood of receiving a community note ($coef. = 0.037$, $p = 0.271$), whereas posts expressing opinions are significantly less likely to receive notes ($coef. = -0.096$, $p < 0.001$). Importantly, we find that posts with higher GPT-estimated misleadingness are significantly more likely to receive community notes ($coef. = 0.308$, $p < 0.001$). This suggests that Community Notes contributors tend to prioritize posts with a higher potential for spreading misinformation.

Post author characteristics. We further evaluate how post author characteristics shape the likelihood of receiving community notes. For verified types of post authors, we find that posts from blue ($coef. = -0.122$, $p < 0.001$) and business ($coef. = -0.203$, $p < 0.001$) accounts are less likely to receive community notes than posts from unverified accounts. Additionally, posts from accounts with a higher number of followees are less likely to receive community notes ($coef. = -0.060$, $p < 0.001$). In contrast, we find that posts from accounts with higher misinformation exposure scores ($coef. = 0.621$, $p < 0.001$) or right-leaning partisanship ($coef. = 0.053$, $p < 0.01$) are more likely to receive community notes. This finding suggests that contributors prioritize posts from accounts with higher risk of spreading misinformation and right-leaning partisanship.

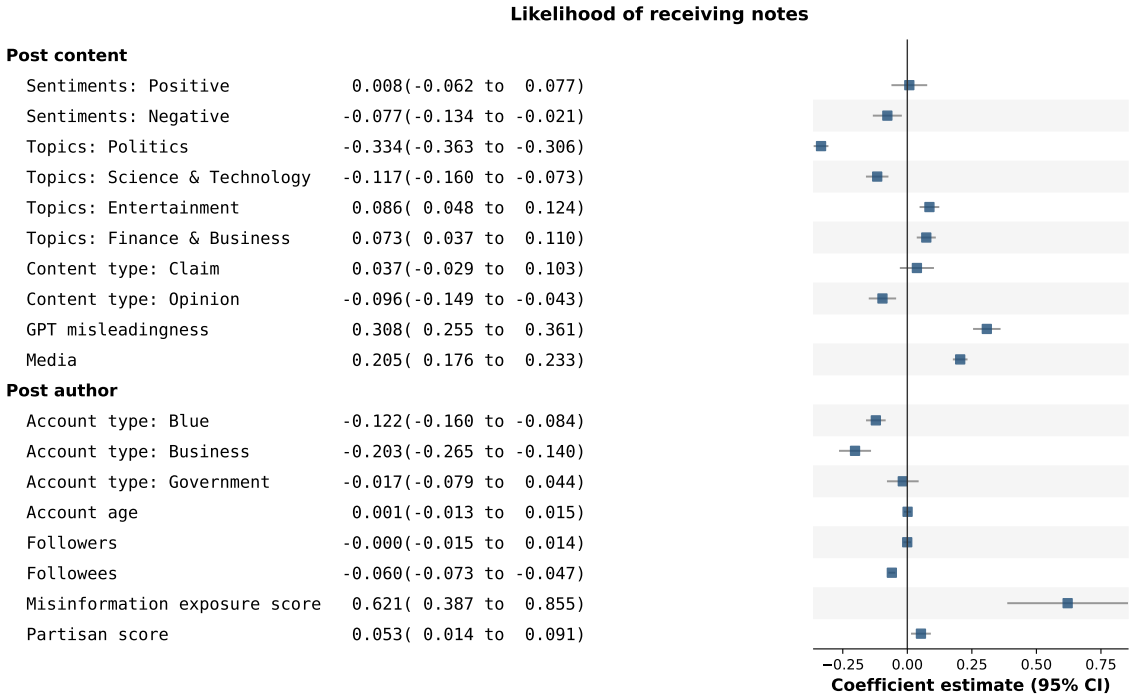

Fig. 2. The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note. Shown are coefficient estimates with 95% Confidence Intervals (CIs). The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

In summary, the likelihood that a requested post receives a community note is shaped by both posts' and their authors' characteristics. Requested posts with negative sentiment, opinionated framing, or authored by blue/business accounts are less likely to attract notes, as are those related to politics or science. In contrast, posts related to entertainment or finance, those containing media, and posts flagged by GPT as more misleading are significantly more likely to receive notes. At the author level, posts from accounts with higher misinformation exposure or right-leaning partisanship are also more likely to be annotated. These findings suggest that Community Notes contributors who write notes and requestors who submit requests may have different selection patterns that are associated with both post features and author attributes.

4.2 Request Timing and Request-Fostered Notes

The request feature allows non-contributors to support the Community Notes program by flagging posts, with the goal of fostering note writing. However, receiving community notes is not necessarily a direct result of the submitted requests, as contributors can also independently select and fact-check posts regardless of requests. To assess whether the request feature meaningfully fosters note creation (RQ2), we examine the timing of requests relative to note creation.

Request timing relative to note creation. As shown in Fig. 3a, the median time from post creation to the submission of first request is 2.72 hours, while it takes significantly longer for a post to accumulate five requests (median of 13.4

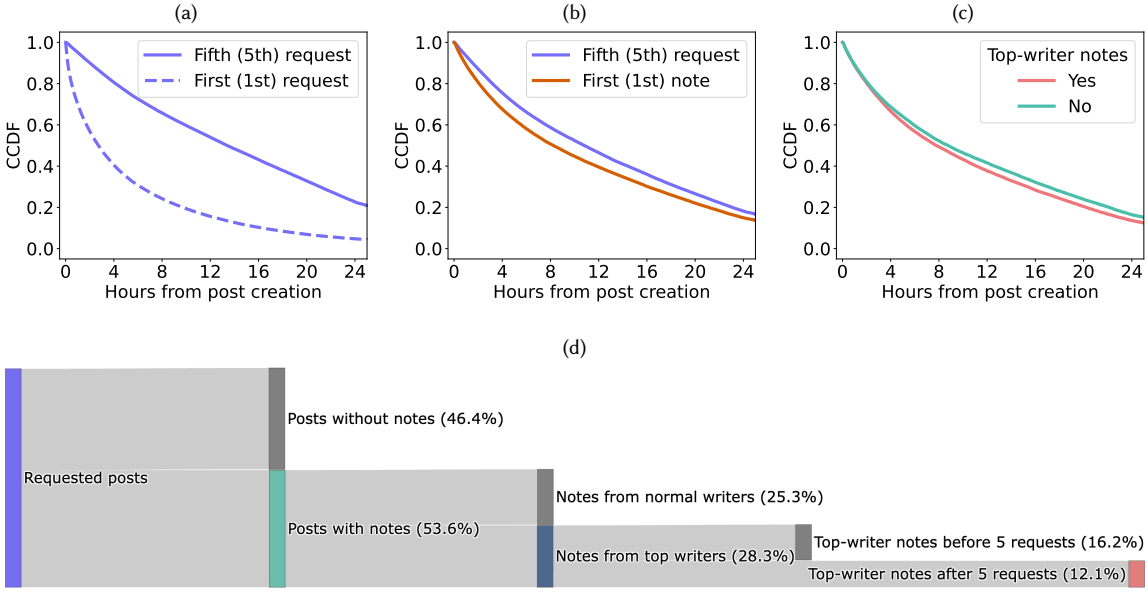

Fig. 3. The statistics for the request timing relative note writing. (a) The Complementary Cumulative Distribution Functions (CCDFs) for the hours from post creation to the submission of first request and to the submission of fifth request. (b) The CCDFs for the hours from post creation to the submission of fifth request and to the generation of first community note, respectively. (c) The CCDFs for the hours from post creation to the notes written by top writers versus other contributors. (d) The sankey plot illustrating the estimated proportion of posts for which community notes are likely fostered by requests.

hours; $p_{MWU} < 0.001$). We find that the response time of community notes (median of 8.2 hours) is significantly shorter than the time it typically takes for posts to receive five requests (Fig. 3b; $p_{MWU} < 0.001$). Furthermore, Fig. 3c shows that top writers (median of 7.78 hours) generate notes almost an hour faster than other contributors (median of 8.76 hours; $p_{MWU} < 0.001$). Taken together, these findings suggest that community notes tend to be written before the fifth request is submitted (i. e., request threshold)—implying that requests may not be the primary driver of note creation. Consistent with the guidelines of the request feature, we find that only an estimated 12.1% of requested posts were likely influenced by the request activity, as their associated notes were written by top writers after the request threshold was reached (Fig. 3d). Therefore, the effect of the request feature in driving note creation is still limited.

Request-fostered notes from top writers. We further employ logistic regression to compare how top writers select posts after seeing request alerts versus when writing notes without such alerts. The estimation results are reported in Fig. 4. The coefficient estimates for Politics ($coef. = 0.247$, $p < 0.001$) and Science & Technology ($coef. = 0.091$, $p < 0.05$) are significantly positive. This suggests that top writers select more political and scientific posts when receiving request alerts. On the contrary, the coefficient estimates for Entertainment ($coef. = -0.097$, $p < 0.01$) and opinionated content ($coef. = -0.102$, $p < 0.05$) are significantly negative. This suggests that top writers choose less entertainment-related and opinionated content after receiving request alerts. These results suggest that requests can shift contributors' attention toward political and scientific content, while diverting attention away from opinionated posts or those related to entertainment. In particular, the coefficient estimate for GPT-estimated misleadingness is significantly positive ($coef. = 0.451$, $p < 0.001$), suggesting that top writers prioritize posts with higher potential misleadingness when guided by request alerts.

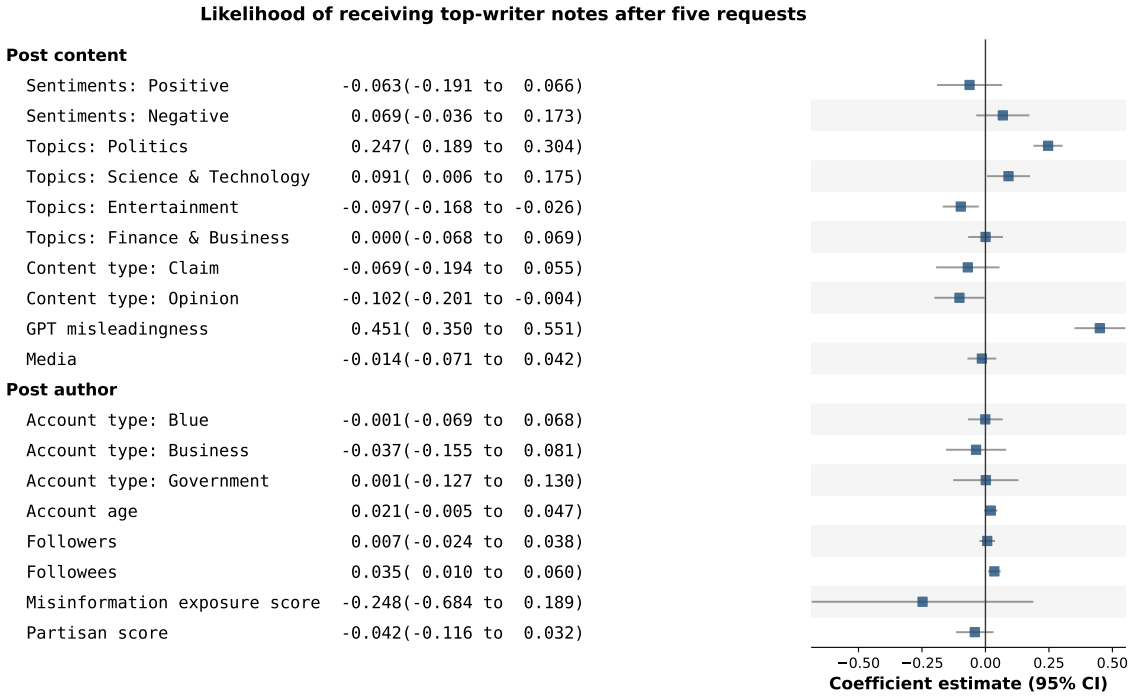

Fig. 4. The estimation results for the logistic regression model predicting the likelihood that a requested post receives a top-writer note after the request threshold, i. e., request-fostered notes. Shown are coefficient estimates with 95% CIs. The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

In summary, out of all requested posts eligible for community evaluation, only 12.1% appear to have been generated by top writers through request alerts. Furthermore, requests typically arrive later than community notes themselves, as the time from post creation to request alerts is significantly longer than the time from post creation to note generation, limiting the request feature's potential to foster note writing. However, by comparing notes written by top writers before and after the request threshold, we find that requests shift contributors' attention toward political content while prioritizing posts with higher potential misleadingness. Thus, although the overall effect of requests on note creation remains modest, the feature shows promise in channeling contributor attention toward high-stakes content at greater risk of misinformation.

4.3 Algorithmic Evaluation of Community Notes

To address RQ3, we downloaded and rerun the source code of the note selection algorithm released by 𝕏. Fig. 5a presents the distribution of note intercepts and note factors from the note selection algorithm. The note intercepts reflect the perceived helpfulness of community notes across raters, with higher intercepts indicating higher helpfulness. The note factors capture the polarization in ratings, with values deviating from 0 indicating greater disagreement among raters regarding the helpfulness of community notes. Here, we categorize all community notes into three groups: *writer-only* (notes on posts that received fewer than five requests and were solely written by contributors), *request-fostered* (notes

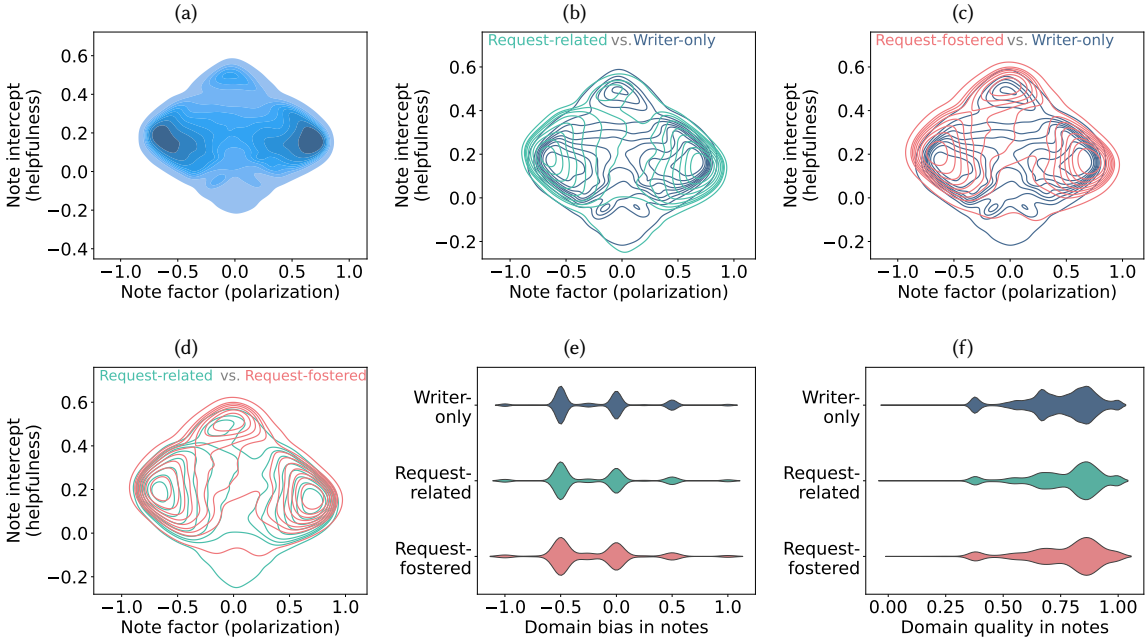

Fig. 5. Overview of note evaluations and source domains in community notes. (a) The distribution of estimated note intercepts (helpfulness) and estimated note factors (polarization) from the note selection algorithm. (b) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-related notes and writer-only notes. (c) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-fostered notes and writer-only notes. (d) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-related notes and request-fostered notes. (d) The violin plots showing the distributions of domain bias in community notes across the three categories: writer-only, request-related, and request-fostered. (e) The violin plots showing the distributions of domain quality in community notes across the three categories: writer-only, request-related, and request-fostered.

authored by top writers after the fifth request, likely influenced by requests), and *request-related* (all remaining notes on posts that received five or more requests but not classified as "request-fostered"). We then analyze the helpfulness and polarization (measured as the absolute value of note factors) of community notes across these categories.

Helpfulness and polarization of request-fostered notes. We examine the distributions of note intercepts and note factors across writer-only notes, request-related notes, and request-fostered notes.

- Request-related notes vs. writer-only notes (Fig. 5b): Request-related notes exhibit lower helpfulness (mean = 0.172 vs. 0.178; $p_{MWU} < 0.001$) and greater polarization (mean = 0.453 vs. 0.343; $p_{MWU} < 0.001$), compared to writer-only notes.
- Request-fostered notes vs. writer-only notes (Fig. 5c): Request-fostered notes are rated as more helpful (mean = 0.231 vs. 0.178; $p_{MWU} < 0.001$) but also more polarized (mean = 0.424 vs. 0.343; $p_{MWU} < 0.001$) than writer-only notes.
- Request-related notes vs. request-fostered notes (Fig. 5d): Request-fostered notes are rated as more helpful (mean = 0.231 vs. 0.172; $p_{MWU} < 0.001$) and less polarized (mean = 0.424 vs. 0.453; $p_{MWU} < 0.001$) than request-related notes.

Taken together, these findings indicate that, although requests in general are associated with less helpful and more polarized notes, request-fostered notes written by top writers tend to achieve higher quality—being both more helpful

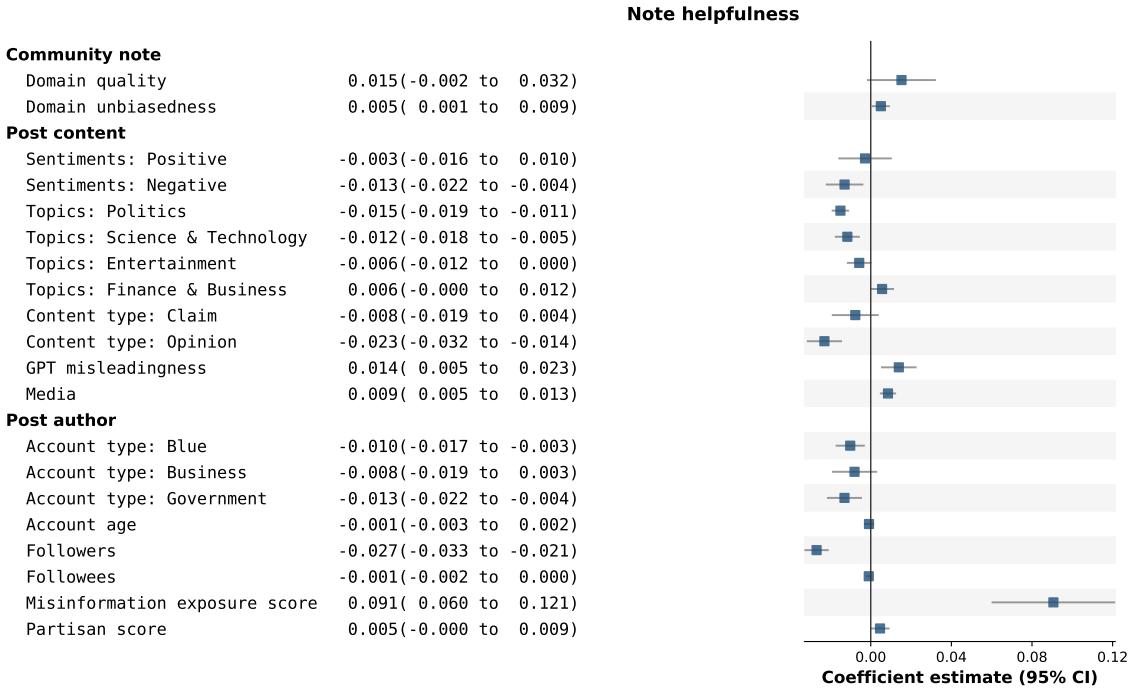

Fig. 6. The estimation results for the linear regression model predicting note helpfulness. Shown are coefficient estimates with 95% CIs. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

and less polarized than request-related notes. To further substantiate this pattern, we estimate a logistic regression predicting whether a note is request-fostered versus request-related using helpfulness and polarization as predictors. The results show a strong positive association for helpfulness ($coef. = 2.528$, $p < 0.001$) and a significant negative association for polarization ($coef. = -0.231$, $p < 0.001$), confirming that request-fostered notes stand out within the request context by being both more helpful and less polarized than other notes. Subsequently, we investigate what factors could explain this quality advantage.

Since contributors are required to cite external sources when writing community notes, we evaluate cited domains in terms of political bias and information quality to probe possible explanations for the higher quality of request-fostered notes than other notes. Overall, we find that community notes tend to cite more left-leaning domains than right-leaning ones (Fig. 5e). Domains in request-related (mean of −0.210) are slightly more left-leaning than those in writer-only notes (mean of −0.185; $p_{MWU} < 0.001$). The political bias of domains in request-fostered notes (mean of −0.214) has no statistically significant difference from request-related notes ($p_{MWU} = 0.352$). In terms of information quality (Fig. 5f), request-related notes cite higher-quality domains (mean of 0.769) than writer-only notes (mean of 0.742; $p_{MWU} < 0.001$), whereas request-fostered notes (mean of 0.767) do not differ from request-related notes ($p_{MWU} = 0.834$). These findings suggest that the higher quality of request-fostered notes is unlikely to be explained by domain choices alone.

Because Community Notes contributors are anonymized, rater bias toward specific writers is unlikely. Observed differences in helpfulness or polarization likely reflect the types of posts and authors contributors choose to fact-check. We therefore examine which notes, posts, and authors are associated with higher helpfulness or polarization.

Factors associated with note helpfulness. The estimation results from a linear regression model predicting note helpfulness are shown in Fig. 6. We find that the unbiasedness (i. e., neither left-leaning nor right-leaning) of external domains cited in community notes is positively associated with note helpfulness ($coef. = 0.005$, $p < 0.05$), suggesting that notes referencing politically neutral sources are perceived as more helpful compared to those citing left- or right-leaning domains. In contrast, the information quality of external domains is not significantly associated with helpfulness ($coef. = 0.015$, $p = 0.080$).

Regarding post content, notes on posts with higher negative sentiment are rated as less helpful ($coef. = -0.013$, $p < 0.01$). Posts related to Politics ($coef. = -0.015$, $p < 0.001$) and Science & Technology ($coef. = -0.012$, $p < 0.001$) tend to receive less-helpful notes compared to posts without these topics. Additionally, posts with higher opinion scores are associated with lower note helpfulness ($coef. = -0.023$, $p < 0.001$). Conversely, community notes on posts containing media ($coef. = 0.009$, $p < 0.001$) or with higher estimated misleadingness ($coef. = 0.014$, $p < 0.01$) are rated as more helpful. With respect to post authors, notes on posts from the blue ($coef. = -0.010$, $p < 0.01$) and government ($coef. = -0.013$, $p < 0.01$) accounts, as well as from authors with more followers ($coef. = -0.027$, $p < 0.001$), are less helpful compared to those from unverified accounts with fewer followers. In contrast, higher misinformation exposure scores of post authors are positively associated with note helpfulness ($coef. = 0.091$, $p < 0.001$).

In summary, the helpfulness of community notes is positively associated with (i) citations to unbiased domains, (ii) posts containing media, (iii) post misleadingness, and (iv) misinformation exposure scores of post authors. On the contrary, the helpfulness of community notes is negatively associated with (i) negative sentiment in posts, (ii) political and scientific topics, (iii) opinionated content, (iv) blue and government accounts, and (v) the number of followers.

Factors associated with note polarization. The estimation results for a linear regression model predicting note polarization are shown in Fig. 7. Community notes that cite politically unbiased domains tend to have reduced polarization across their ratings ($coef. = -0.019$, $p < 0.001$), whereas citations to domains with higher information quality are associated with increased polarization ($coef. = 0.043$, $p < 0.01$).

Regarding post content, positive sentiment in posts is negatively associated with note polarization ($coef. = -0.046$, $p < 0.001$). Posts related to Politics are linked to higher polarization ($coef. = 0.053$, $p < 0.001$), whereas posts related to Finance & Business ($coef. = -0.021$, $p < 0.001$) are associated with lower polarization. Community notes on posts containing stronger claims ($coef. = 0.037$, $p < 0.001$) or more opinionated content ($coef. = 0.063$, $p < 0.001$) tend to be more polarized. In contrast, community notes on posts that include media ($coef. = -0.026$, $p < 0.001$) or exhibit higher misleadingness ($coef. = -0.030$, $p < 0.001$) have lower polarization scores. For post authors, posts from all types of verified accounts—blue ($coef. = 0.044$, $p < 0.001$), business ($coef. = 0.041$, $p < 0.001$), and government ($coef. = 0.053$, $p < 0.001$)—tend to receive community notes with higher polarization compared to posts from unverified accounts. Additionally, posts authored by accounts with more followers are associated with greater note polarization ($coef. = 0.057$, $p < 0.001$). However, posts from accounts with higher misinformation exposure scores are more likely to receive community notes with reduced polarization ($coef. = -0.163$, $p < 0.001$).

In summary, the polarization of community notes is positively associated with (i) high quality of external domains, (ii) political topics, (iii) posts containing claims or opinionated content, and (iv) verified accounts with many followers. Conversely, the polarization of community notes is negatively associated with (i) citations to unbiased domains, (ii)

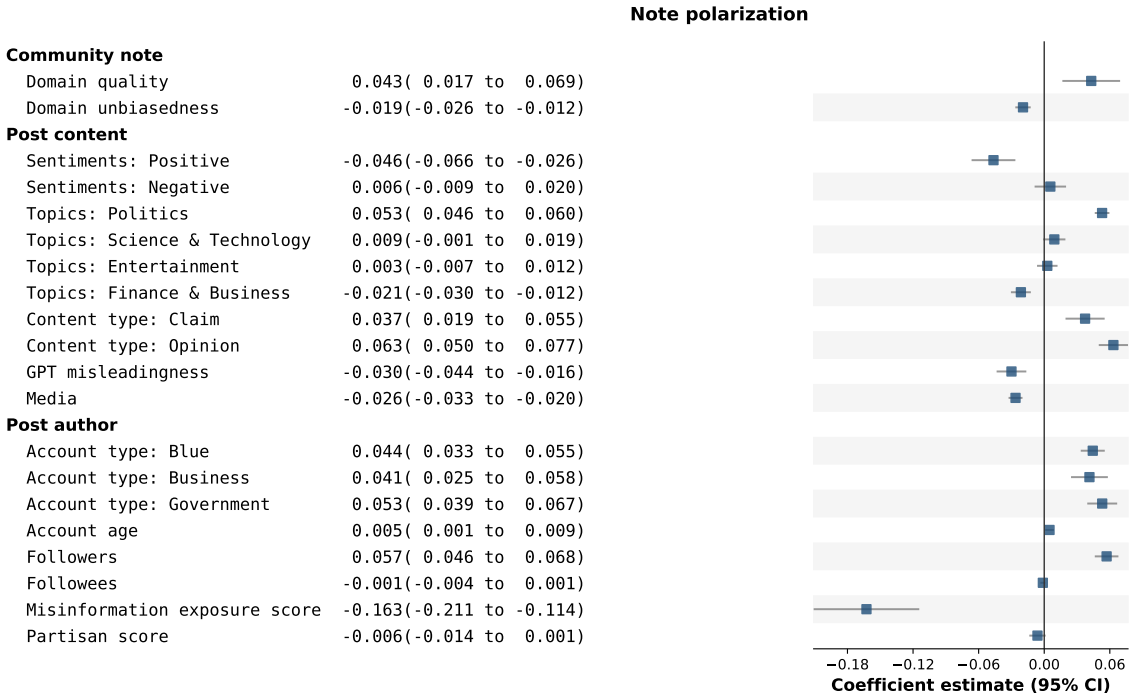

Fig. 7. The estimation results for the linear regression model predicting note polarization. Shown are coefficient estimates with 95% CIs. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

positive sentiment, (iii) posts containing media, (iv) misleadingness of post content and (v) misinformation exposure scores of post authors.

Together with our findings in Section 4.2, which show that requests shift top writers' attention toward political content while prioritizing posts with higher potential misleadingness, the analysis of factors affecting note helpfulness and polarization provides additional insight. We find that post misleadingness is positively associated with note helpfulness but negatively associated with note polarization, whereas political content exhibits the opposite pattern. Consequently, the higher helpfulness and lower polarization of request-fostered notes can be partially interpreted as reflecting top writers' selective fact-checking of misleading posts.

4.4 Summary of Main Findings

In this study, based on the request feature implemented on 𝕏 and associated requests, posts, and community notes, we conduct a comprehensive analysis on the Community Notes system with respect to content selection, contributor behavior, and note evaluation. Overall, our analysis highlights three key findings.

- Requested posts are more likely to receive notes when they are associated with entertainment, finance, media, high misleadingness, or authors with high misinformation exposure, but less so for political or opinionated content. This suggests that Community Notes contributors and requestors have distinct selection patterns.

- Only an estimated 12.1% of the requested posts appear to receive request-fostered notes from top writers, with many notes written independently of the request feature. Nevertheless, requests can shape contributors' behavior by directing their focus toward politically salient content, while still prioritizing posts with higher risk of misinformation.
- Request-fostered notes exhibit higher helpfulness and lower polarization compared to notes generated prior to the request threshold or by non–top writers, a pattern that may partly reflect top writers selectively fact-checking posts with higher misleadingness.

5 Discussion

Although the Community Notes system has emerged as a promising model of crowdsourced fact-checking, its speed and coverage remain limited, constraining its scalability on social media platforms [12, 18]. To address these challenges, 𝕏 recently introduced the Request Community Note feature, which opens a new channel for fact-checking participation and enables a broader set of users to actively solicit notes on specific posts. In this study, we analyze a large-scale dataset of 98,685 posts surfaced through the request feature, along with their associated notes, to examine how this feature influences 𝕏's Community Notes system. Our findings shed light on the role of requests in shaping what content is fact-checked, how contributors engage, and the quality of resulting notes.

5.1 Research Implications

We find that more than half of the requested posts (53.6%) ultimately receive community notes. Yet, the processes of submitting requests and generating notes are largely independent, with only 12.1% of posts receiving annotations that can be potentially attributed to the request feature. This suggests that requests do not strongly determine contributor behavior. Instead, the relative independence of the two processes makes requests a useful lens for examining how Community Notes contributors select posts for fact-checking and how their choices align—or diverge—from the priorities expressed by users who submit requests (i. e., requestors).

Content selection by contributors. Our results show that contributors' decisions to fact-check requested posts are shaped by both post content and author characteristics. Specifically, contributors are less likely to write notes for posts with negative sentiment or opinionated content, authored by blue or business accounts, or focused on politics or science. In contrast, posts related to entertainment or finance, those containing claims or media, and those flagged by GPT as more misleading are significantly more likely to receive notes. At the post author level, posts from accounts with higher misinformation exposure or right-leaning partisanship are also more likely to be annotated.[1] The misalignment between the types of posts surfaced by requestors (e. g., political content) and those contributors prioritize for annotation underscores a divergence in fact-checking selection.

On the one hand, Community Notes contributors tend to fact-check claim posts with higher estimated misleadingness and from authors with higher misinformation exposure scores. This suggests that, compared to requestors, contributors adopt stricter standards, prioritizing content with greater potential for misinformation and harm. On the other hand, previous research finds that contributors often fact-check posts that are relatively straightforward to verify or contain claims already addressed in earlier expert fact-checks [7]. Similarly, LLMs perform strongly when processing logically structured claims but struggle with more complex ones [17]. This suggests that, although contributors appear to focus their attention on posts with higher GPT-estimated misleadingness, their fact-checking may simultaneously be biased

[1]To validate our findings and rule out confounding from request-driven activity, we exclude posts with request-fostered community notes (as identified in Section 4.2) and repeat our analysis. The results remain robust and consistent (see details in Suppl. S2). Notably, in this robustness check, the presence of claims becomes positively associated with the likelihood of receiving a community note, further reinforcing the role of substantive content in shaping contributors' fact-checking decisions.

toward content that is easier to assess. Such a focus can improve efficiency and support the production of high-quality notes, but it also risks overlooking politically salient or ambiguous posts that requestors perceive as urgent. Addressing this gap may require hybrid models of collaboration, where professional fact-checkers complement contributors by targeting content that is more ambiguous, emergent, or difficult to verify [5].

Partisan selection biases between contributors and requestors. Notably, the human-centered fact-checking approaches—whether expert-based or crowdsourced—cannot fully avoid political bias in the selection of targets [16, 48, 53]. Our results reveal significant differences in how requestors and contributors approach political content and partisan cues. Previous work has shown that users preferentially challenge content authored by those with opposite partisan leanings, and Republicans are flagged more often than Democrats for sharing misinformation on 𝕏's Community Notes [2, 29, 50]. Our findings echo this pattern: contributors are more likely than requestors to write notes on posts authored by Republicans, whereas requestors are comparatively more likely to surface posts authored by Democrats. This divergence raises an important interpretive challenge. On the one hand, it may reflect partisan selection biases, with anonymized contributors and requestors each applying their own subjective judgments when deciding what to fact-check [36]. On the other hand, it could mirror underlying asymmetries in the misinformation landscape, where different political groups are disproportionately represented in the spread of misleading content. The coexistence of such tendencies highlights the complexity of maintaining balance and fairness in decentralized fact-checking systems like Community Notes.

Response of top writers to requests. Although the direct impact of the request feature on fostering note writing and coverage remains modest, requests appear to shift the attention of top writers: they are more likely to engage with political posts and focus on content with higher estimated misleadingness. By concentrating on posts with greater misinformation potential, request-fostered notes from top writers are evaluated as more helpful and less polarized than other notes on requested posts. In this sense, contributors play a crucial role as gatekeepers, directing community fact-checking toward high-risk content.

At the same time, the request feature indirectly amplifies the influence of these contributors in shaping what gets fact-checked. A small elite of writers (22.2%) is responsible for nearly half of all community notes (49.7%), underscoring the uneven distribution of labor characteristic of volunteer-based systems. This concentration of fact-checking activity raises concerns about scalability and resilience: if the system relies disproportionately on a narrow core of expert-like participants, its long-term sustainability may be vulnerable to burnout, disengagement, or shifts in contributor incentives [5]. Moreover, while the reliance on elite contributors enhances quality and consensus, it also complicates the platform's vision of broad, community-driven participation. Rather than democratizing fact-checking, requests may consolidate authority in the hands of a few, blurring the line between peer production and expert review.

Reliability and vulnerability of the note selection algorithm. We provide a comprehensive analysis of how the helpfulness and polarization of community notes are shaped by both content and source factors. Posts with higher estimated misleadingness and authored by accounts with greater misinformation exposure are more likely to elicit helpful annotations. This finding supports the view that Community Notes can function as a reliable mechanism for surfacing accurate fact-checks and correcting misinformation [3].

However, our findings on polarization also highlight a critical vulnerability. Notes on political or opinionated content tend to provoke divided evaluations, even when they cite high-quality domains, which resonates with previous research [55]. Such polarization does not necessarily imply inaccuracy, but can reflect partisan disagreement over credibility, legitimacy of sources, or interpretive framing [8]. Therefore, the Community Notes system faces a significant challenge: factually accurate annotations may still fail to gain consensus if they are evaluated through polarized lenses.

Addressing this limitation points to the need to improve the note selection algorithm to better balance factual accuracy with the practical requirement of fostering consensus among heterogeneous communities.

5.2 Practical Implications

Our findings provide several actionable insights for the design of community-based fact-checking systems on social media platforms, such as 𝕏's Community Notes.

Reducing the delay of request alerts. Requests currently function less as a mechanism to increase overall note volume and coverage, partly due to differences in selection patterns between contributors and requestors, and partly due to delays in reaching the request alert threshold relative to note generation. Although the first request may appear early in a post's diffusion, it often takes considerably longer to accumulate enough requests to trigger the alert. To reduce these delays, the request mechanism could be adapted to customize thresholds based on requestor reputation, granting lower thresholds to high-reputation users while maintaining safeguards for others. For example, 𝕏 is already testing a system that computes helpfulness scores for requestors, such that users with higher scores require fewer co-requests before their note requests are surfaced to contributors [43].

Leveraging LLMs to mitigate contributor bias and workload. Given recent advances in LLMs, these models could be incorporated into the Community Notes system to automatically write notes for requested posts with high estimated misleadingness. Such an integration would help mitigate the potential selection bias in note writers and also alleviate their workload. For instance, 𝕏 has opened the AI Note Writer API to developers, enabling the design of tools that assist in writing notes on the reuqested posts [39]. Beyond writing notes on posts requested by users, LLM-based tools can be expanded to provide proactive support across the platform—for example, issuing warnings to authors when they attempt to publish potentially misleading posts, reviewing posts after publication, and surfacing content with high misleadingness for community evaluation. Together, these applications would help broader and more uniform coverage of online content, complementing the request feature and strengthening the scability of community-based fact-checking.

Addressing polarization and improving consensus. The current note selection algorithm faces a significant challenge: community notes, even when factually accurate, on posts related to politics and from high-influence users (e. g., verified accounts with many followers) tend to provoke polarized evaluations and fail to reach consensus among heterogeneous communities [8]. To address this challenge, platforms could consider two directions for improvement. (i) Introducing an review layer by experts or contributors could help access highly polarized but potentially helpful notes, guiding decisions on which annotations to display [5]. In addition, the algorithm could assign higher weights to ratings from top contributors when aggregating evaluations. (ii) Our findings indicate that notes citing unbiased sources are evaluated as more helpful and less polarized, consistent with prvious research [54]. Given this, platforms could provide contributor guidelines or incentives to encourage the use of neutral sources. Furthermore, LLMs can be leveraged to synthesize existing community notes, improve clarity, and then foster consensus-building across diverse user communities [18].

Taken together, these actionable insights can inform the design and enhancement of community-based fact-checking systems. In particular, LLMs have promise for transforming the Community Notes system in streamlining request processing and assisting in note generation, thereby helping the system remain both reliable and responsive to user demands [32].

5.3 Limitations and Future Work

Our study has several limitations that future research could address. First, our analysis is based on observational data from 𝕏, which constrains causal inferences. While we examine various characteristics—related to posts, their authors, and associated notes—and identify their associations with note outcomes, we cannot definitively determine whether certain factors directly cause higher helpfulness or lower polarization. Second, while GPT-based measures provide scalable assessments of potential misinformation, they may not perfectly address emerging and complex misleading content. This could affect interpretations regarding contributors' priority strategies. Third, given that the top-writer status is not available in our dataset, we cannot perfectly distinguish request-fostered notes from contributor-driven notes. Nevertheless, our estimation method ensures that the effect of the request feature in fostering note writing is not underestimated. Finally, contributors and requestors are anonymized in the Community Notes system, preventing the analysis of individual-level behavior, motivation, or expertise. While this reduces rater bias, it also limits understanding of how personal experience or identity may influence content selection and note quality.

6 Ethics Statement

This research has received ethical approval from the Ethics Review Panel of the University of Luxembourg (ref. ERP 23-053 REMEDIS). All analyses are based on publicly available data. We declare no competing interests.

Acknowledgments

This research is supported by the Luxembourg National Research Fund (FNR) and Belgian National Fund for Scientific Research (FNRS), as part of the project REgulatory Solutions to MitigatE DISinformation (REMEDIS), grant ref. INTER_FNRS_21_16554939_REMEDIS.

References

[1] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), eabf4393.

[2] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[3] Matthew R Allen, Nimit Desai, Aiden Namazi, Eric Leas, Mark Dredze, Davey M Smith, and John W Ayers. 2024. Characteristics of X (formerly Twitter) Community Notes addressing COVID-19 vaccine misinformation. *JAMA* 331, 19 (2024), 1670–1672.

[4] Sinan Aral and Dean Eckles. 2019. Protecting elections from social media manipulation. *Science* 365, 6456 (2019), 858–861.

[5] Isabelle Augenstein, Michiel Bakker, Tanmoy Chakraborty, David Corney, Emilio Ferrara, Iryna Gurevych, Scott Hale, Eduard Hovy, Heng Ji, Irene Larraz, et al. 2025. Community moderation and the new epistemology of fact checking on social media. *ArXiv* (2025).

[6] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[7] Nadav Borenstein, Greta Warren, Desmond Elliott, and Isabelle Augenstein. 2025. Can Community Notes replace professional fact-checkers?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 535–552.

[8] Paul Bouchaud and Pedro Ramaciotti. 2025. Algorithmic resolution of crowd-sourced moderation on X in polarized settings across countries. *ArXiv* (2025).

[9] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2023. New threats to society from free-speech social media platforms. *Commun. ACM* 66, 10 (2023), 37–40.

[10] Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine* 45, 3 (2024), 354–368.

[11] Eun Cheol Choi and Emilio Ferrara. 2024. FACT-GPT: Fact-checking augmentation via claim matching with LLMs. In *Companion Proceedings of the ACM Web Conference 2024*. 883–886.

[12] Yuwei Chuai, Moritz Pilarski, Thomas Renault, David Restrepo-Amariles, Aurore Troussel-Clément, Gabriele Lenzini, and Nicolas Pröllochs. 2024. Community-based fact-checking reduces the spread of misleading posts on social media. *ArXiv* (2024).

[13] Yuwei Chuai, Anastasia Sergeeva, Gabriele Lenzini, and Nicolas Pröllochs. 2025. Community fact-checks trigger moral outrage in replies to misleading posts on social media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[14] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024. Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–52.

[15] Yuwei Chuai and Jichang Zhao. 2022. Anger can make fake news viral online. *Frontiers in Physics* 10 (2022), 970174.

[16] Yuwei Chuai, Jichang Zhao, Nicolas Pröllochs, and Gabriele Lenzini. 2025. Is fact-checking politically neutral? Asymmetries in how U.S. fact-checking organizations pick up false statements mentioning political elites. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 403–429.

[17] Luigia Costabile, Gian Marco Orlando, Valerio La Gatta, and Vincenzo Moscato. 2025. Assessing the potential of generative agents in crowdsourced fact-checking. *ArXiv* (2025).

[18] Soham De, Michiel A Bakker, Jay Baxter, and Martin Saveski. 2025. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM on Web Conference 2025*. 3751–3761.

[19] Matthew R DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* 121, 50 (2024), e2322823121.

[20] Joan Donovan. 2020. Social-media companies must flatten the curve of misinformation. *Nature* (2020).

[21] Chiara Patricia Drolsbach and Nicolas Pröllochs. 2024. Content moderation on social media in the EU: Insights from the DSA Transparency Database. In *Companion Proceedings of the ACM on Web Conference 2024*. 939–942.

[22] Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. 2024. Community notes increase trust in fact-checking on social media. *PNAS Nexus* 3, 7 (2024), pgae217.

[23] Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. 2024. Misinformation poses a bigger threat to democracy than you might think. *Nature* 630, 8015 (2024), 29–32.

[24] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (2022), 13–29.

[25] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[26] Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5916–5936.

[27] Jon Green, William Hobbs, Stefan McCabe, and David Lazer. 2022. Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election. *Proceedings of the National Academy of Sciences* 119, 34 (2022), e2115900119.

[28] Joel Kaplan. 2025. More speech and fewer mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/. [Accessed: 2025-05-25].

[29] Simon Fox Kuuse, Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2025. Crowdsourced fact-checking or biased commentary? Analyzing political bias in Twitter's Community Notes. In *Companion Proceedings of the ACM on Web Conference 2025*. 2661–2669.

[30] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[31] Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee. 2023. "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review* (2023).

[32] Haiwen Li, Soham De, Manon Revel, Andreas Haupt, Brad Miller, Keith Coleman, Jay Baxter, Martin Saveski, and Michiel A Bakker. 2025. Scaling human judgment in community notes with LLMs. *ArXiv* (2025).

[33] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* 2, 9 (2023), pgad286.

[34] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 251–260.

[35] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2025. Understanding the effects of AI-based credibility indicators when people are influenced by both peers and experts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[36] Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G Rand. 2025. Political motives help rather than hinder crowdsourced fact-checking. *OSF* (2025).

[37] Cameron Martel and David G Rand. 2024. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour* 8, 10 (2024), 1957–1967.

[38] Mohsen Mosleh and David G Rand. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications* 13, 1 (2022), 7144.

[39] X Community Notes. 2025. AI note writers. https://communitynotes.x.com/guide/en/api/overview. [Accessed: 2025-09-09].

[40] X Community Notes. 2025. Downloading data. https://communitynotes.x.com/guide/en/under-the-hood/download-data. [Accessed: 2025-05-20].

[41] X Community Notes. 2025. Media matching. https://communitynotes.x.com/guide/en/under-the-hood/media-matching. [Accessed: 2025-09-09].

[42] X Community Notes. 2025. Open-source code. https://communitynotes.x.com/guide/en/under-the-hood/note-ranking-code. [Accessed: 2025-05-20].

[43] X Community Notes. 2025. Request a note. https://communitynotes.x.com/guide/en/under-the-hood/note-requests. [Accessed: 2025-05-20].

[44] X Community Notes. 2025. Signing up. https://communitynotes.x.com/guide/en/contributing/signing-up. [Accessed: 2025-05-20].

[45] X Community Notes. 2025. Top contributors. https://communitynotes.x.com/guide/en/contributing/top-contributors. [Accessed: 2025-05-20].

[46] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha WT Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[47] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[48] Moritz Pilarski, Kirill Olegovich Solovev, and Nicolas Pröllochs. 2024. Community Notes vs. Snoping: How the crowd selects fact-checking targets on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1262–1275.

[49] Adam Presser. 2025. Testing a new feature to enhance content on TikTok. https://newsroom.tiktok.com/en-us/footnotes. [Accessed: 2025-05-25].

[50] Thomas Renault, Mohsen Mosleh, and David G Rand. 2025. Republicans are flagged more often than Democrats for sharing misinformation on X's Community Notes. *Proceedings of the National Academy of Sciences* 122, 25 (2025), e2502053122.

[51] Claire E Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J Van Bavel, and Stefan Feuerriegel. 2023. Negativity drives online news consumption. *Nature Human Behaviour* 7, 5 (2023), 812–822.

[52] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at Twitter: How does the crowd compare with experts?. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1736–1746.

[53] Jieun Shin and Kjerstin Thorson. 2017. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication* 67, 2 (2017), 233–255.

[54] Kirill Solovev and Nicolas Pröllochs. 2025. References to unbiased sources increase the helpfulness of community fact-checks. *Scientific Reports* 15, 1 (2025), 25749.

[55] Keisuke Toyoda, Tomoki Fukuma, Koki Noda, Yoshiharu Ichikawa, Kyosuke Kambe, Yu Masubuchi, Hiroshi Someda, and Fujio Toriumi. 2025. Understanding and mitigating polarization in Community Notes: Factors and strategies for improved consensus. In *Companion Proceedings of the ACM on Web Conference 2025*. 2694–2698.

[56] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[57] Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[58] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. 2022. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *ArXiv* (2022).

[59] YouTube. 2024. Testing new ways to offer viewers more context and information on videos. https://blog.youtube/news-and-events/new-ways-to-offer-viewers-more-context/. [Accessed: 2025-05-25].

[60] Yixuan Zhang, Yimeng Wang, Nutchanon Yongsatianchot, Joseph D Gaggiano, Nurul M Suhaimi, Anne Okrah, Miso Kim, Jacqueline Griffin, and Andrea G Parker. 2024. Profiling the dynamics of trust & distrust in social media: A survey study. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

Supplementary Materials for "Request a Note: How the Request Function Shapes X's Community Notes System"

S1 Evaluation of GPT Outputs

Previous work has shown that Large Language Models (LLMs) can effectively identify claims and perform well in truthfulness classification, underscoring their potential to support crowdsourced fact-checking [11, 17, 19]. Building on this, we employ GPT-4.1, the state-of-the-art GPT model at the time of this study, to assess whether posts contain factual claims and/or personal opinions, and to estimate the misleadingness of each post. The GPT outputs are based on the following prompt:

GPT prompt for content type and misleadingness

You are an assistant that evaluates tweet content for two independent dimensions: content type and misleadingness.
For each input tweet, return a single-line JSON string with the following structure:
{
"type_scores": {
"claim": float (0 to 1), # Degree to which the tweet presents a factual/verifiable claim
"opinion": float (0 to 1) # Degree to which the tweet expresses a subjective opinion
},
"misleadingness": float (0 to 1) # How misleading the tweet is, where 0 = not misleading, 1 = extremely misleading
}
Guidelines:
- "claim" and "opinion" scores are independent and do not need to sum to 1.
- Only return the JSON string. Do not include any explanation or additional text.

Notably, GPT-generated annotations are not treated as definitive ground truth. Instead, they function as standardized proxies that allow for large-scale comparisons across posts. We manually validate a random subsample of 200 posts and evaluate the performance of the GPT outputs. Additionally, to balance workload, six trained research assistants each independently evaluate a set of 100 posts, with each post reviewed by three assistants. Specifically, for claims and opinions, assistants rate the extent to which a tweet presents a factual or objectively verifiable claim or expresses a subjective opinion, using a scale from 0 to 1. For misleadingness, assistants rate the accuracy of the GPT-generated explanation on the same 0–1 scale. The explanation is generated by GPT based on the following prompt:

GPT prompt for misleadingness explanation

You are an assistant that previously predicted the misleadingness of a given tweet, returning a score between 0 and 1.
Your task now is to provide a concise and factual explanation that justifies the specific misleadingness score you gave to the tweet.
Do not re-score the tweet. The explanation must be no more than 100 words.

We examine the Pearson correlation coefficients between the scores assigned by research assistants and those generated by GPT for claims and opinions. For each post, we average the ratings from the research assistants and compute the correlation with the corresponding GPT scores. The correlations are 0.403 ($p < 0.001$) for claims and 0.620 ($p < 0.001$) for opinions, indicating a substantial alignment between human judgments and GPT outputs.

To further assess reliability, we transform the scores into binary classes using a threshold of 0.5. Specifically, a post is labeled as containing a claim on the human side if at least two of the three assistants assign it a score ≥ 0.5. Using this criterion, GPT achieves an accuracy of 63% in classifying claims and 73.5% in classifying opinions. In addition, our manual check shows that GPT can provide accurate misleadingness explanations for 90% posts.

S2 Likelihood of Receiving Community Notes

To further examine how Community Notes contributors differ from requestors in post selection and mitigate confounding from request-driven activity, we exclude posts with request-fostered notes identified in Section 4.2 and repeat our analysis on the likelihood of receiving community notes for the requested posts. The estimation results are reported in Fig. S1, and they remain robust and consistent with our main findings.

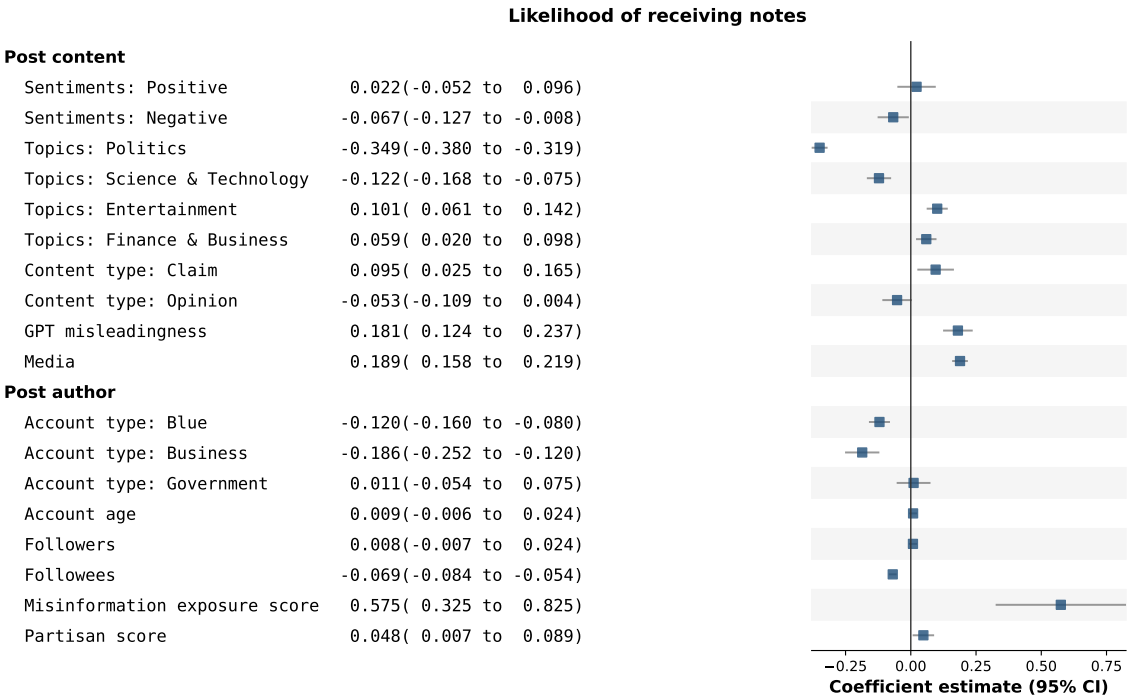

Fig. S1. The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note. Shown are coefficient estimates with 95% Confidence Intervals (CIs). Posts with request-fostered notes are omitted during estimation. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.