Can Community Notes Replace Professional Fact-Checkers?

Anonymous ACL submission

Abstract

Two commonly-employed strategies to combat the rise of misinformation on social media are (i) fact-checking by professional organisations and (ii) community moderation by platform users. Policy changes by Twitter/X and, more recently, Meta, signal a shift away from partnerships with fact-checking organisations and towards an increased reliance on crowdsourced community notes. However, the extent and nature of dependencies between fact-checking and *helpful* community notes remain unclear. To address these questions, we use language models to annotate a large corpus of Twitter/X community notes with attributes such as topic, cited sources, and whether they refute claims tied to broader misinformation narratives. Our analysis reveals that community notes cite fact-checking sources up to five times more than previously reported. Fact-checking is especially crucial for notes on posts linked to broader narratives, which are *twice* as likely to reference fact-checking sources compared to other sources. In conclusion, our results show that successful community moderation heavily relies on professional fact-checking.

1 Introduction

The proliferation of misinformation on social media (Arnold, 2020; Diakopoulos, 2020), along with the rise of generative AI (Augenstein et al., 2024) have led to increasing concerns about its current and future potential harms, (e.g., to health (Clemente et al., 2022)) and threats to democracy and political stability (Reglitz, 2022).

Fact-checkers play a crucial role in combatting misinformation (Graves, 2017), and in recent years, have partnered with social media platforms, e.g., Meta, YouTube, and TikTok, to tackle its spread on these platforms. However, due to the scale of misleading content shared online, community moderation (e.g., options to flag potential misinformation, group/server moderators) is often employed in parallel (Morrow et al., 2022), as a complementary approach (e.g., (Google, 2025); see also the practice of *snoping* (Pilarski et al., 2024)).

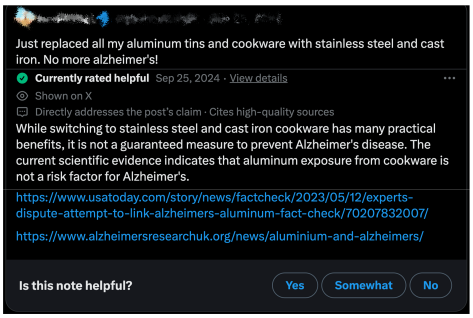

Figure 1: An example of a community note. Notice the fact-checking link and rating.

The expansion of fact-checking projects in the last decade (Lauer and Graves, 2024), alongside their broader initiatives to curb misinformation (e.g., citizen media literacy programmes (Juneja and Mitra, 2022)) have been aided by partnerships with social media platforms such as Meta and Google (Graves and Anderson, 2020), which fund independent fact-checking agencies to fact-check potentially false claims on their platform.[1]

However, political pressure and accusations of bias and censorship, and most recently, Meta's announcement of its plans to end its partnerships with fact-checkers in the U.S. and implement a community moderation model (Meta, 2025), threatens the financial stability of fact-checking organisations, and hence, their ability to keep up with the increasing volume and sophistication of misinformation spread (Stencel et al., 2024; IFCN, 2024).

Meta's recent policy shift also implies that these two strategies (fact-checking and community notes) are independent and in opposition, rather than two complementary strategies of tackling online mis-

[1] Fact-checkers provide a judgment of claim veracity and exert no influence on the platforms' content moderation policies (Catalanello and Sanders, 2025).

information. In this paper, *we examine Twitter/X community notes as a case study to understand how fact-checking is used in community notes*. Specifically, we investigate the following two questions: **(RQ1) To what extent do community notes rely on the work of professional fact-checkers?** and **(RQ2) What are the traits of posts and notes that rely on fact-checking sources?** Studying the relationship between fact-checking and community notes is vital for understanding the shared role of expert and community-driven fact-checking in the global information ecosystem.

We find that at least 1 in 20 community notes rely explicitly on the work of professional fact-checkers, while this reliance is higher still for high-stakes topics such as health and politics. Our experiments also show that fact-checking is vital for debunking misleading content linked to broader narratives or conspiracy theories. These findings imply that high-quality community notes cannot be produced independently of professional fact-checking. They further suggest that the pressure on fact-checkers exerted by platforms and politicians by defunding and discrediting fact-checking organisations will have corrosive effects on the quality of notes and destructive implications for information integrity more widely.

2 Background

Community moderation has been proposed as a means of addressing the scalability (Martel et al., 2024) and cross-partisanship trust (Flamini, 2019) challenges associated with fact-checking. Twitter/X's Community Notes programme (piloted in 2021 and publicly launched in October 2022 (Twitter/X, 2021)) is a notable example of such a system.

Any platform user may volunteer as a Community Notes contributor, although they must achieve a particular 'rating impact score' before they can write notes (Twitter/X, 2024b). Notes that achieve a 'helpful' rating appear underneath the post, explaining why the post is misleading (see Fig. 1). To be rated 'helpful', a note must receive similar levels of helpfulness rating from users with diverse viewpoints (Twitter/X, 2024a).

A small but growing body of work has analysed Twitter/X's Community Notes dataset, focusing on the targets, sources, and limitations of notes.

Targets of notes. Community notes tend to focus on misleading posts from large accounts (Pilarski et al., 2024), focusing on posts that lack important content or present unverified claims as facts (Pröllochs, 2022; Drolsbach and Pröllochs, 2023).

Sources in notes. Analyses have indicated that notes were rated more helpful if they link to 'trustworthy' sources (Pröllochs, 2022) and that the majority of sources cited by notes were 'trustworthy' left-leaning news outlets. A recent study finds that 55% of URLs used in notes were related to news websites, 18% to research, 9% to social media, 9% to encyclopedic sources, but *just 1.2%* to fact-checking sources (Kangur et al., 2024).

Limitations of notes. Only 11% of submitted notes reach 'helpful' status (i.e., shown to users) by achieving a cross-perspective (Renault et al., 2024; Wirtschafter and Majumder, 2023), and the time frame for notes to reach the algorithm's required agreement level (15.5 hours on average) limits its capacity to halt misinformation spread (Renault et al., 2024). Additional concerns about the notes' efficacy highlight their indifference to the expertise needed for certain claims and reliance on subjective helpfulness rather than objective facts, free labour and inadequate support and guardrails regarding explicit content (Gilbert, 2025).

Our work provides novel insights into the targets, sources and limitations of community notes by shedding light on the relationship between notes and professional fact-checking. Namely, we study the extent to which fact-checking sources form the basis of note-writers' efforts to counter misinformation and identify the strategies they employ.

3 Dataset

We download files containing all community notes and their metadata from the official website,[2] which amounts to 1.5M notes authored between January 28th 2021 and January 6th 2025. Of these, a total of 135K are rated by the community as 'Helpful', 51K are rated 'Not helpful', and 1.3M are unpublished, i.e., did not receive enough community ratings to reach a verdict. See Fig. 5 in App. B for statistics.

We filter the notes as follows. First, we remove 526K non-English notes, which we identify by applying the language detection library fast-langdetect.[3] Then, we further filter 268K "unnecessary" notes—notes attached to tweets that are classified by the community as "not misleading".

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

[3] https://github.com/LlmKira/fast-langdetect

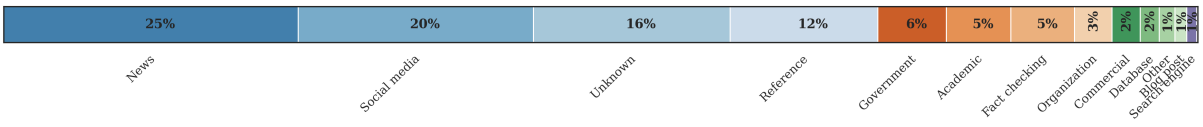

Figure 2: The categories of links used by Community notes' authors as a source.

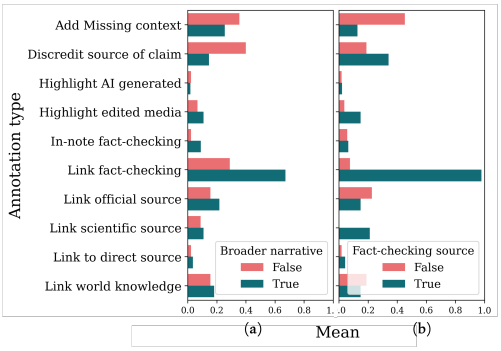

Figure 3: (a) strategies in debunking claims related to broader narratives. (b) the different ways in which fact-checking sources are used to debunk claims.

Conspiracy \ FC source	✓	✗
✓	22%	11%
✗	28%	39%

Table 1: Percentage of samples related to a broader narrative or conspiracy vs. have a fact-checking source.

Finally, to focus only on notes that are used to address misinformation, we filter out 44K notes that contain one of the words "ad", "spam", or "phishing". Following these filtration steps, we are left with a dataset containing 664K notes.

The next step involves categorising the sources that the note authors use to support their claims. First, we use regex to extract all the URLs found in the notes. See Tab. 3 in App. B for a list of the top-100 most common domains. We classify each URL into one of 13 categories (detailed in Fig. 2). In particular, we check the URL against a manually annotated dataset of popular domains and fact-checking agencies. If unsuccessful, we classify it using a language model. The full pipeline is described in App. C.1.

Moreover, we further subsample the notes for performing the in-depth analysis required for answering RQ2 (§4.2). From the notes rated as 'Helpful', we sample 3.5K notes with a "Fact-checking" source and a random sample of 22K additional notes. We then used web crawling to scrape the text of the posts to which these notes were attached. We name this subset $\mathcal{S}_{\text{text}}$ for simplicity.

4 Analysis

4.1 RQ1: To what degree do community notes rely on fact-checkers?

According to Fig. 2, at least 5% of all English community notes contain an external link to professional fact-checkers. This number grows to 7% when only considering notes rated as 'helpful' (Fig. 6 in App. B). Conversely, only 1% of notes rated as 'not helpful' contain a fact-checking source (Fig. 7 in App. B). These figures are significantly larger than what was reported in previous studies (1.2% (Kangur et al., 2024)), possibly because Kangur et al. (2024) utilise a smaller dataset of fact-checking agencies and classify fact-checking divisions of popular journals as "news". The results imply that notes incorporating fact-checking sources are generally considered more helpful.

We further assess whether notes with fact-checking sources are perceived to be of higher quality by analysing individual user ratings of notes both with and without such sources. Specifically, we collect user ratings for a balanced (i.e., including of a fact-checking source or not) sample of 20K notes rated by at least 50 users, and calculated the average ratings for the notes. As can be seen in Fig. 9 in App. B, community notes with fact-checking sources are generally rated higher than their counterparts. Interestingly, while notes with fact-checking links are more likely to be regarded as having a good source (higher *HelpfulGoodSources*), they are also more likely to be rated as *notHelpfulSourcesMissingOrUnreliable*. Tab. 4 in App. B contains a sample of such notes.

4.2 RQ2: What are the traits of posts and notes that rely on fact-checking sources?

We begin by performing a topic analysis, comparing topics of posts whose notes reference fact-checking sources to those citing other sources. To this end, we apply a strong zero-shot text classification model[4] to our $\mathcal{S}_{\text{text}}$ subset by classify-

[4] https://huggingface.co/r-f/ModernBERT-large-zeroshot-v1 with default settings.

ing spans of the form "`Tweet:<POST TEXT>; Note <NOTE TEXT>`" into one of 13 classes. The authors manually evaluated the quality of the classification results and considered it satisfactory. Notably (Fig. 4), fact-checking sources are more likely to be included in posts related to high-stakes issues such as health, science, and scams and less likely to be included in posts on tech or sports.

We then analyse annotations (binary attributes explaining the warrant for the note) by community note authors. Fig. 8 in App. B contains the full breakdown of annotations for notes with and without fact-checking sources. Notes containing a link to fact-checking sources are overrepresented in posts where unverified information is presented as a fact or when the post contains a factual error. Conversely, they are under-represented in posts with outdated information or satirical content. Tab. 5 in App. B contains a sample of such notes.

These results indicate that community note-writers adapt their strategies based on the stakes and scope of the claim, and the depth of research needed to counter misinformation. We hypothesise that they are more likely to rely on external fact-checking when refuting complex or unverifiable claims (Wuehrl et al., 2024), as well as claims related to broader narratives or conspiracy theories which cannot be fully addressed in the scope of a note.[5] Conversely, claims involving misleading media can often be debunked with examples alone, making fact-checking sources unnecessary. To investigate this hypothesis, the authors of this paper manually annotated 400 $<$ post, note $>$ pairs from $\mathcal{S}_{\text{text}}$ with attributes related to the complexity of the claims and how community notes address them. (See App. C.2 for annotation guidelines). The results (Fig. 3.a) support our hypothesis. Claims related to broader narratives or conspiracy theories are much more likely to include a link to a fact-checking source. In contrast, other types of claims are more likely to be addressed by providing missing context or by invalidating the credibility of the claim's source. Additionally, Fig. 3.b depicts the different ways in which fact-checking sources are used to debunk claims. It demonstrates how such sources are rarely used to provide missing context but rather focus on discrediting sources of claims and providing scientific evidence.

We extend the manual annotation to an LLM-based analysis of 8K balanced (post, note) pairs from $\mathcal{S}_{\text{text}}$. We task OpenAI's GPT-4[6] with determining whether a pair relates to a broader narrative or a conspiracy theory. Listing 2 in App. C details the prompt used. To evaluate model accuracy, two authors independently labelled 100 balanced pairs, achieving an agreement rate of 0.88 and resolving disagreements through discussion. The model attained an F_1 score of 0.85—strong performance for this challenging task. The results (Tab. 1) support our hypothesis: pairs related to a broader narrative or conspiracy theory are *twice* as likely to cite fact-checking sources compared to other sources. In contrast, other pairs are nearly 30% less likely to do so. These findings also highlight the prevalence of such claims and further underscore the importance of fact-checking in combating complex misinformation narratives.

Figure 4: Distribution of notes' topics, with and without a fact-checking source.

5 Conclusion

In this work, we annotate a large corpus of Twitter community notes with attributes such as topic, cited sources, and whether they refute claims tied to broader misinformation narratives. We find that effective community moderation depends on professional fact-checking to an extent far greater than previously reported. We find that community notes linked to broader narratives or conspiracy theories are particularly reliant on fact-checking.

Our results reveal that community notes and professional fact-checking are deeply interconnected—fact-checkers conduct in-depth research beyond the reach of amateur platform users, while community notes publicise their work. The move by platforms to end their partnerships and funding for fact-checking organisations will hinder their ability to fact-check and pursue investigative journalism, which community note writers rely on. This, in turn, will limit the efficacy of community notes, especially for high-stakes claims tied to broader narratives or conspiracies.

[5] For example, the claim "Michelle Obama is a male".

[6] Version `gpt-4o-2024-08-06`.

Limitations

The main limitations of our work concern the characteristics of the dataset we analyse. First, we restrict our analysis to notes written in English, excluding over half a million notes in other languages. This decision was made to avoid potential noise and biases arising from the authors' unfamiliarity with public discourse in different regions and reliance on machine translation. In future work, we aim to extend our analysis to other languages.

Moreover, except for a small subset of notes, we did not have access to the original tweets they were written for. Even when the tweet text was available, many contained non-text media, were written in internet vernacular that was challenging to interpret, or lacked important context. These factors limit the accuracy and effectiveness of our models and analysis.

Finally, due to resource constraints, our manual annotation study was limited to a relatively small sample of tweets and notes. In future work, we wish to utilise crowd workers to not only annotate a larger dataset but also increase the diversity and perspective of the annotators.

Broader Impact and Ethical Considerations

Given that this work analyses real-world posts, ethical concerns may arise from using this data for research purposes. Posts from non-protected accounts and Community Notes on Twitter/X are publicly available, however, we acknowledge that they may contain sensitive personal information. To minimise any breach of anonymity and privacy, we anonymised links to individual accounts, and we do not publicly release this information. We do not analyse the posts or notes by individual users, and instead examine aggregated data in the form of topics and sources cited.

Although the Community Notes dataset represents attempts to curb harmful misinformation and conspiracies, given the intense partisanship involved (Allen et al., 2022; Draws et al., 2022), as well as the explicit content of some claims, some instances may be considered offensive. We also acknowledge that our own perspectives and biases as authors shape the impact of our findings in certain ways. For example, as mentioned in the previous section, we were unable to analyse non-English posts in-depth, so our conclusions are likely somewhat focused on discourse in the Anglosphere (e.g., the US, UK, Ireland, Canada, Australia, New Zealand etc.). Furthermore, although we based our criteria for conspiracy theories on well-established sources, e.g., AP News, FactCheck.org, the European Commission, and identified conspiratorial narratives from both left- and right-wing sources, our own perspectives (i.e., as scientists from Western countries) may also have impacted what we considered to be conspiracy theories.

References

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*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Phoebe Arnold. 2020. The challenges of online fact checking: how technology can (and can't) help. Technical report, Full Fact.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, pages 1–12.

Rebecca Catalanello and Katie Sanders. 2025. Meta is ending its third-party fact-checking partnership with us partners. here's how that program works.

Yuwei Chuai, Moritz Pilarski, Gabriele Lenzini, and Nicolas Pröllochs. 2024a. Community notes reduce the spread of misleading posts on X.

Yuwei Chuai, Anastasia Sergeeva, Gabriele Lenzini, and Nicolas Pröllochs. 2024b. Community Fact-Checks Trigger Moral Outrage in Replies to Misleading Posts on Social Media. ArXiv:2409.08829 [cs].

Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024c. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–52.

Suárez Vicente Javier Clemente, Eduardo Navarro-Jiménez, Juan Antonio Simón-Sanjurjo, Ana Isabel Beltran-Velasco, Carmen Cecilia Laborde-Cárdenas, Juan Camilo Benitez-Agudelo, Álvaro Bustamante-Sánchez, and José Francisco Tornero-Aguilera. 2022. Mis–dis information in covid-19 health crisis: A narrative review. *International Journal of Environmental Research and Public Health*, 19(9).

Nicholas Diakopoulos. 2020. Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism. *Digital journalism*, 8(7):945–967.

Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2114–2124, Seoul Republic of Korea. ACM.

Chiara Patricia Drolsbach and Nicolas Pröllochs. 2023. Diffusion of Community Fact-Checked Misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22.

Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. 2024. Community notes increase trust in fact-checking on social media. *PNAS Nexus*.

Daniela Flamini. 2019. Most republicans don't trust fact-checkers, and most americans don't trust the media.

Yang Gao, Maggie Zhang, and Huaxia Rui. 2024. Can Crowdchecking Curb Misinformation? Evidence from Community Notes.

Sarah Gilbert. 2025. Three reasons Meta will struggle with community fact-checking. *MIT Technology Review*.

Google. 2025. Misinformation policies - youtube help.

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3):518–537.

Lucas Graves and C.W. Anderson. 2020. Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. *New Media & Society*, 22(2):342–360.

IFCN. 2024. State of the fact-checkers report. Technical report, International Fact-Checking Network.

Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes. ArXiv:2406.12444 [cs].

Sarawut Kankham and Jian-Ren Hou. 2024. Community Notes vs. Related Articles: Assessing Real-World Integrated Counter-Rumor Features in Response to Different Rumor Types on Social Media. *International Journal of Human–Computer Interaction*, pages 1–15.

Laurens Lauer and Lucas Graves. 2024. How to grow a transnational field: A network analysis of the global fact-checking movement. *New Media & Society*, 0(0):19.

Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G. Rand. 2024. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, 19(2):477–488. PMID: 37594056.

Meta. 2025. More speech and fewer mistakes.

Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or false: Studying the work practices of professional fact-checkers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386.

Moritz Pilarski, Kirill Olegovich Solovev, and Nicolas Pröllochs. 2024. Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1262–1275.

Nicolas Pröllochs. 2022. Community-based fact-checking on twitter's birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 794–805.

Merten Reglitz. 2022. Fake news and democracy. *J. Ethics & Soc. Phil.*, 22:162.

Thomas Renault, David Restrepo Amariles, and Aurore Troussel. 2024. Collaboratively adding context to social media posts reduces the sharing of false news. ArXiv:2404.02803 [econ].

Mark Stencel, Erica Ryan, and Joel Luther. 2024. With half the planet going to the polls in 2024, fact-checking sputters. Technical report, Duke Reporters' Lab.

Twitter/X. 2021. Introducing birdwatch, a community-based approach to misinformation.

Twitter/X. 2024a. Note ranking algorithm.

Twitter/X. 2024b. Rating and writing impact.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Valerie Wirtschafter and Sharanya Majumder. 2023. Future challenges for online, crowdsourced content moderation: Evidence from twitter's community notes. *Journal of Online Trust and Safety*, 2(1).

Amelie Wuehrl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024. What makes medical claims (un)verifiable? analyzing entity and relation properties for fact verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2046–2058, St. Julian's, Malta. Association for Computational Linguistics.

A Extended Background

A.1 Professional fact-checking and community note practices

Although fact-checks and community notes share similarities in how they address misleading claims, they also differ in key elements of practice and techniques of persuasion and communication (Kankham and Hou, 2024). Fact-checking typically involves the analysis and verification of public claims (e.g., statements in news reports and social media). In addition to verifying claims, in recent years many fact-checking organisations have also assumed a wider role in combatting misinformation spread, conducting long-term investigative journalism projects and citizen media literacy programs (Juneja and Mitra, 2022).

Professional fact-checkers in organisations signatory to the International Fact-Checking Network (IFCN) follow a rigorous set of principles and transparency commitments.[7] In contrast, any platform user can contribute to community notes under anonymity, and the rating approach relies on the 'wisdom of crowds', while there appears to be little oversight or transparency regarding the biases of the note-writers.

Numerous studies have documented the structured workflow that fact-checkers follow: (i) claim selection; (ii) collecting evidence; (iii) deciding on a verdict; and (iv) writing the fact-checking article (Graves, 2017; Micallef et al., 2022; Warren et al., 2025). Fact-checking articles, which are subject to multiple rounds of editorial scrutiny, are more formal and standardised in tone and style than community notes, which vary considerably. Fact-checkers must rely on credible sources and evidence to convince the reader, while community note writers may employ a range of persuasion techniques, such as appeals to emotion or other logical fallacies.

Moreover, community notes typically serve as direct rebuttals to misleading posts, while fact-checking articles may address a more general claim than is expressed in a specific post. Finally, fact-checking articles are a one-way exchange, while community notes represent a more horizontal and interactive dialogue between writer and recipient of the fact-check (Kankham and Hou, 2024).

Our work builds on current understanding of the relationship between professional fact-checking and amateur community moderation by examining the extent to which community note writers deploy the work of professional fact-checkers in their notes.

[7] https://www.ifcncodeofprinciples.poynter.org/the-commitments

A.2 Impact of Community Notes on misinformation spread

A growing number of studies have examined the effectiveness of community notes as interventions on the spread of misinformation. Posts identified by community notes as misleading have been found to attain less virality (reposts, quote tweets and replies) than non-misleading posts (Drolsbach and Pröllochs, 2023; Renault et al., 2024). Community notes have also been shown to increase the probability of tweet retractions and deletions and speed up the retraction process (Gao et al., 2024; Renault et al., 2024). However, other studies have found more mixed evidence; for example, that users' followers, likes and engagement increase after their post receives a community note, indicating that there is little impact on the subsequent behaviour of Twitter/X users; (Wirtschafter and Majumder, 2023). Curiously, one study claims that showing community notes on posts reduced the spread of misleading posts by an average of 61% (Chuai et al., 2024a), while a more recent analysis by the same authors found no effect of community notes on engagement with misinformation (Chuai et al., 2024c).

Empirical tests have demonstrated the promise of community notes; people shown either community notes or related news article suggestions were both less likely to to believe and report misleading information compared to a control group. For positive rumours, people shown community notes were less likely to believe them and share them than people shown related articles, however for negative rumours, related articles were more effective in reducing self-reported belief and likelihood of sharing, although these findings were based on responses to a single health-related claim (Kankham and Hou, 2024). People shown community notes alongside misleading social media posts were more accurate in identifying misleading posts, and the notes were judged to be more trustworthy than context-free misinformation flags (e.g., "Checked by fact-checkers" or "Checked by other social media users"), regardless of (US-centric) political beliefs (Drolsbach et al., 2024). On the other hand, crowdworkers have been found to exhibit bias: in other words, the more they liked a

statement's claimant, the more they overestimate the statement's truthfulness (and vice versa). The higher workers' self-reported confidence in their ability to judge the truthfulness of statements was, the less accurate their judgments were, demonstrating overconfidence (Draws et al., 2022). Displaying community notes leads users to post more negative and angry replies to misleading posts (Chuai et al., 2024b).

B Additional Material

This section details additional results or material referenced from the paper's main body.

Fig. 5 A histogram of the number of community notes written every month and their rating (*helpful*, *not helpful*, or *needs more data*).

Fig. 6 The categories of links used by Community notes' authors as a source, filtering for notes rated as "helpful".

Fig. 7 The categories of links used by Community notes' authors as a source, filtering for notes rated as "not helpful".

Fig. 8 Mean scores of community annotations of misleading posts.

Fig. 9 Community ratings of notes with and without fact-checking source.

Tab. 2 List of professional fact-checking organisations and their URLs.

Tab. 2 List of top 100 most common domains found in the community notes dataset, and their categorization.

Tab. 4 Examples of community notes containing fact-checking sources that are rated as having *notHelpfulSourcesMissingOrUnreliable*.

Tab. 5 A sample of tweets, notes, and their community annotations, as well as whether the note contains a fact-checking link.

C Reproducibility

C.1 Source Classification Pipeline

We classify each URL in our dataset of 664K notes into one of 13 categories using the pipeline described below.

1. Check whether the domain name of the URL is found in a manually curated list of domains of professional fact-checking organisations (See Tab. 2 in App. B for the full list). If so, classify the URL as "fact-checking".
2. Otherwise, search for paraphrases of the word "fact-check" in the URL,[8] and classify it as "fact-checking" if a match was found.
3. Otherwise, check whether the domain name is found in Tab. 3, which the authors of this paper manually annotated.
4. Otherwise, use GPT-4[9] to classify the domain name into one of the 13 categories. Listing 1 in App. C details the prompt we used.
5. Finally, if GPT-4 fails or outputs an unknown category, label the URL as "unknown".

Using this pipeline, we successfully classify 95% of the URLs to one of the 13 categories.

Listing 1 The prompt used to classify URLs into categories.

Listing 2 The prompt used to classify tweets and notes into broader narratives and conspiracy theories.

C.2 Manual Annotation Setup

We annotate 400 (tweet, note) pairs from $\mathcal{S}_{\text{text}}$ with 12 binary attributes. Each (tweet, note) pair was annotated in a multi-label fashion, i.e., more than one attribute can be selected at the same time. Fig. 10 depict our simple annotation setup, with the 12 attributes being as follows.

Broader narrative Whether the (tweet, note) pair is related to a broader narrative or a conspiracy theory.

Discredit source of claim If the community note describes the source shared by the original post as non-credible.

Add missing context If the community note provides some missing context to refute a claim.

Highlight AI generated If the community note claims that the post shared AI-generated content.

Highlight edited media If the community note claims that the post shared some media that was edited (edited with Photoshop, the clip was cut, etc.).

Link to direct source If the community note shares a link to a source where an entity says that a claim made about them is false.

Link official source If the community note shares a link to an official source such as a government website.

Link scientific source If the community note

[8] These URLs mostly link to the fact-checking divisions of news outlets, e.g., https://apnews.com/article/fact-checking-909101991741

[9] Version `gpt-4o-2024-08-06`.

Name	URL	Language	Region/domain
Lead stories	leadstories.com	English	Global
AFP Factuel	factuel.afp.com	French	Global
AAP FactCheck	aap.com.au/factcheck	English	Australia
Full Fact	fullfact.org	English	Global
Science Feedback	science.feedback.org	English	Science
Politifact	politifact.com	English, Spanish	USA
HoaxEye	hoaxeye.wordpress.com	English	Images
Logically Facts	logicallyfacts.com	Multiple	Europe/India
FactCheckNI	factcheckni.org	English	North Ireland
DFRLab	dfrlab.org	English	Global
FactReview	factreview.gr	Greek	Global
Lupa	lupa.uol.com.br/jornalismo	Portuguese	Global
Check your fact	checkyourfact.com	English	Global
Climate feedback	climatefeedback.org	English	Climate
Factcheck	factcheck.org	English	USA
Health feedback	healthfeedback.org	English	Health
Snopes	snopes.com	English	US
aosfatos	aosfatos.org	Portuguese	Global
Demagog	demagog.org.pl/fake_news	Polish	Poland
FakeReporter	fakereporter.net	Hebrew	Israel
litmus factcheck	litmus-factcheck.jp	Japanese	Japan
Climate Feedback	climatefeedback.org	English	Global
AFP	factcheck.afp.com	English	Global
USA Today	usatoday.com/story/news/factcheck	English	USA
Statesman	statesman.com	English	USA
Dallas News	dallasnews.com/news/politifact	English	USA
Google Fact Check	toolbox.google.com/factcheck	English	Global
MediaBias/FactCheck	mediabiasfactcheck.com	English	Global
MedDMO	meddmo.eu	English, Greek	Greece, Cyprus, Malta
Poynter	poynter.org/fact-checking	English	USA
Newsmeter	newsmeter.in/fact-check	English, Tamil	India
Africa Check	africacheck.org	English	Africa
Fact Crescendo India	english.factcrescendo.com	English	India
Factseeker	factseeker.lk	English	Sri Lanka
Fact Crescendo Thailand	thailand.factcrescendo.com	Thai	Thailand
Fact Crescendo Afghanistan	afghanistan.factcrescendo.com	Persian	Afghanistan
Only Fact	onlyfact.in	English	India
Factly	factly.in	English	India
Fact Crescendo Sri Lanka	srilanka.factcrescendo.com	Sinhala	Sri Lanka
Fact Crescendo Cambodia	cambodia.factcrescendo.com	Cambodian	Cambodia
Becid	becid.eu	Baltic langs	Baltic
Fact Hunt	facthunt.in	English	India
Tec Arp	techarp.com	English	Global (based in Malaysia)
10 news	10news.com/news/fact-or-fiction	English	USA
RMIT Fact Check	rmit.edu.au	English	Australia
Gigafact	gigafact.org	English	USA
Ayupp	ayupp.com/fact-check	English	India
The Journal	thejournal.ie	English	Ireland

Table 2: List of professional fact-checking organisations and their URLs.

Domain	Category	Domain	Category
x.com	social media	thehill.com	news
twitter.com	social media	amp.theguardian.com	news
youtube.com	social media	whitehouse.gov	government
youtu.be	social media	news.sky.com	news
un.org	organisation	merriam-webster.com	reference
u.today	news	techarp.com	news
t.co	social media	cbc.ca	news
snopes.com	fact checking	politifact.com	fact checking
en.m.wikipedia.org	reference	pbs.org	commercial
en.wikipedia.org	reference	telegraph.co.uk	news
google.com	search engine	businessinsider.com	news
instagram.com	social media	time.com	news
britannica.com	reference	justice.gov	government
reuters.com	news	cnbc.com	news
bbc.co.uk	news	wsj.com	news
apnews.com	news	sciencedirect.com	academic
bbc.com	news	msn.com	news
nytimes.com	news	statista.com	reference
theguardian.com	news	business.x.com	commercial
vice.com	news	amp.cnn.com	news
usatoday.com	news	congress.gov	government
factcheck.org	fact checking	factcheck.afp.com	fact checking
cnn.com	news	yahoo.com	search engine
washingtonpost.com	news	timesofindia.indiatimes.com	news
ncbi.nlm.nih.gov	academic	thelancet.com	academic
nbcnews.com	news	hrw.org	organisation
help.twitter.com	reference	healthfeedback.org	fact checking
cdc.gov	government	fda.gov	government
npr.org	news	m.youtube.com	social media
forbes.com	news	law.cornell.edu	academic
newsweek.com	news	medium.com	blog post
fullfact.org	fact checking	healthfeedback.org	fact checking
dailymail.co.uk	news	who.int	organisation
cbsnews.com	news	haaretz.com	news
web3antivirus.io	database	axios.com	news
timesofisrael.com	news	mayoclinic.org	commercial
help.x.com	reference	nejm.org	academic
nypost.com	news	scienceexchange.caltech.edu	academic
aljazeera.com	news	indiatoday.in	news
reddit.com	social media	bloomberg.com	news
independent.co.uk	news	pewresearch.org	academic
usgs.gov	academic	jamanetwork.com	academic
abcnews.go.com	news	leadstories.com	news
nature.com	academic	dictionary.cambridge.org	reference
gov.uk	government	jpost.com	news
web.archive.org	database	archive.ph	database
foxnews.com	news	healthline.com	commercial
tiktok.com	social media	abc.net.au	news
edition.cnn.com	news	france24.com	news

Table 3: List of top 100 most common domains found in the community notes dataset, and their categorization.

ID	summary
0	This claim ruled mostly false. https://www.politifact.com/factchecks/2020/may/07/facebook-posts/facebook-post-cites-doctors-widely-disputed-calcul/
1	The RedState article claims “the shots do not stop transmission of the virus. This is false. ”“Vaccines provide significant protection from ’getting it’ – infection – and ’spreading it’ – transmission – even against the delta variant.” Source: https://www.usatoday.com/story/news/factcheck/2021/11/17/fact-check-covid-19-vaccines-protect-against-infection-transmission/6403678001/
2	There is no proof of this, the photo is real, it’s not the last photo of the child. But snoops say there is a tenuous link the parents used the same law firm to represent them as Maxwell https://www.snopes.com/fact-check/ghislaine-maxwell-jonbenet-ramsey/
3	unfounded https://www.snopes.com/fact-check/ashley-biden-diary-afraid/
4	The mRNA vaccine does not cause cancer: https://www.factcheck.org/2024/05/still-no-evidence-covid-19-vaccination-increases-cancer-risk-despite-posts/
5	Many of the details in this popular essay are inaccurate and too numerous to list here. The essay was fact checked by Snopes in 2005: https://www.snopes.com/fact-check/the-price-they-paid/
6	POLITIFACT - rates False. The report analysed a small sample of 128 temp stations out of several thousand volunteer-run stations, then extrapolated results. NOAA uses 2 programs to record daily temps. The report did not look at the 900 more sophisticated automated stations. https://www.politifact.com/factchecks/2022/aug/19/facebook-posts/fact-checking-talking-point-about-corrupted-climat/
7	There is no verifiable evidence of campaign espionage in either the 2020 or the 2016 presidential elections. https://www.snopes.com/fact-check/obama-spying-trump-campaign/ https://www.washingtonpost.com/politics/2019/05/06/whats-evidence-spying-trumps-campaign-heres-your-guide/
8	Ladapo did get caught altering COVID vaccine study findings. Ladapo replaced the language from an earlier study draft that found no significant risk from COVID vaccines, to then state there was a high risk https://healthfeedback.org/claimreview/analysis-florida-department-health-surgeon-general-joseph-ladopo-contains-multiple-methodological-problems-covid-19-mrna-vaccines/ https://healthexec.com/topics/clinical/COVID-19/florida-surgeon-general-altered-covid-19-study-findings

Table 4: Examples of community notes containing fact-checking sources that are rated as having *notHelpfulSourcesMissingOrUnreliable*.

Tweet	Note	misleadingUnverifiedClaimAsFact	misleadingOutdatedInformation	misleadingFactualError	misleadingSatire	Fact Checking source
The NASA War Document is absolutely terrifying https://t.co/...	misrepresenting a presentation by NASA scientist Dennis Bushnell, The lecture was not detailing plans by NASA to attack the world it was a lecture for defense industry professionals, and how defense tactics might rise to meet evolving threats in the future. https://leadstories.com/hoax-alert/2021/06/fact-check-the-future-is-now-is-not-a-nasa-war-document-plan-for-world-domination-and-phasing-out-of-humans.html	✓	✗	✗	✗	✓
BREAKING NEWS: International Criminal Investigation calls on every public citizen to recommend indictments for Bill Gates, Anthony Fauci, Pfizer, BlackRock, Tedros and Christian Drosten for pushing everyone to receive the ineffective highly dangerous lethal experimental vaccines...	Video has been fact-checked by USA Today, was found to be misleading, and promotes a conspiracy theory about COVID ... https://ca.movies.yahoo.com/movies/fact-check-viral-video-promotes-204414488.html	✓	✗	✗	✗	✓
1) California is RED. It is just because of the MASSIVE Election Fraud that stupid, brainwashed people believe Calif. is blue. Joe Biden won only in the SFO Bay area ...	The map shows the results of Reagan's reelection in 1984, not Biden's election in 2020. https://en.wikipedia.org/wiki/1984_United_States_presidential_election_in_California	✗	✓	✗	✗	✗
Davis blows up $100,000 fireworks in Kai Cenat setup During the Mr Beast Stream ...	The second photo is from a house fire in Atlanta in 2019. https://www.11alive.com/article/news/local/woodland-brook-drive-cause-of-house-fire/85-ecb7df9b-5f65-44e9-bf9d-8c162d36c334	✗	✓	✗	✗	✗
@cnviolations I swear community notes are the only good thing Elon added since he bought Twitter.	Community notes was first launched under former Twitter CEO Jack Dorsey in 2021 under the name of "Birdwatch". The only thing Elon Musk did was that he renamed the feature to community notes. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation https://www.reuters.com/article/factcheck-elon-birdwatch-idUSL1N31Z2VG/	✗	✗	✓	✗	✓
Thailand will become the first country to make the contract null and void, meaning that Pfizer will become responsible for all vaccine injuries ...	Thailand has no plans to void its Pfizer COVID vaccine contract, an official with the country's National Vaccine Institute said. Thailand's Department of Disease Control also rejected the claims as "fake news." ... https://apnews.com/article/fact-check-covid-vaccine-pfizer-thailand-203948163859	✗	✗	✓	✗	✓
Hilarious tweets by footballers, A thread: 1. Virgil Van Dijk [Current Liverpool Captain] https://t.co/...	Virgil Van Dijk did not tweet this, the tweet was made by a fan account in his name. https://www.pinkvilla.com/sports/fact-check-did-virgil-van-dijk-really-root-for-man-u-because-no-one-likes-liverpool-in-resurfaced-viral-tweet-1287250	✗	✗	✗	✓	✓
Rob Reiner announces he's on the Epstein Client List and Epstein Flight logs. What a fool! When a lawyer tells me to STFU, I STFU! https://t.co/...	This is a digitally altered photo that might be misinterpreted even if used as a joke. The name Rob Reiner is misspelled, and the text is not on Reiner's X timeline. https://twitter.com/robreiner?t=iqu43-NszIW5oOM_KqRSpw	✗	✗	✗	✓	✗

Table 5: A sample of tweets, notes, and their community annotations, as well as whether the note contains a fact-checking link.

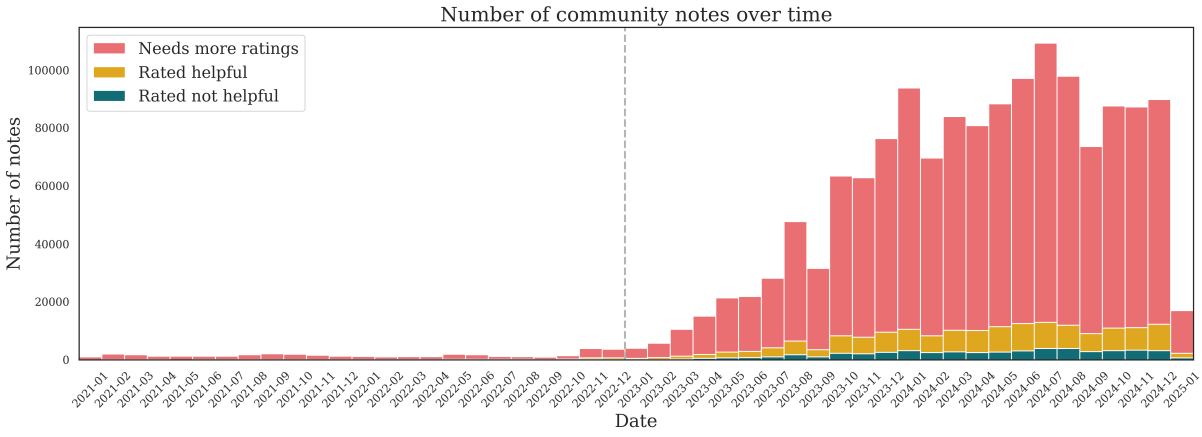

Figure 5: A histogram of the number of community notes written every month and their rating (*helpful*, *not helpful*, or *needs more data*. The grey vertical line (December 2022) indicates the date when the community notes became visible worldwide.

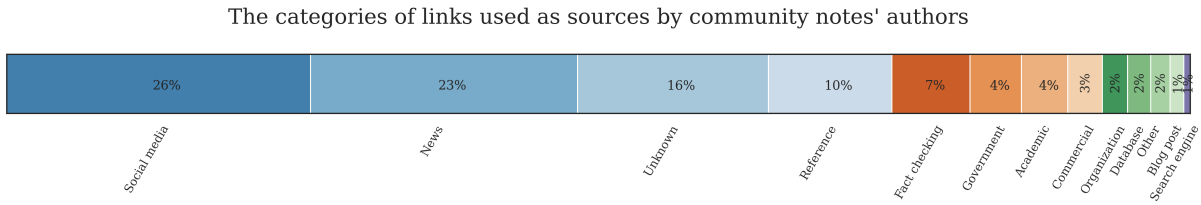

Figure 6: The categories of links used by Community notes' authors as a source, filtering for notes rated as "helpful".

shares a link to some scientific article or website.

Link world knowledge If the community note shares a link to some reference resources such as Wikipedia.

Link fact-checking If the community note shares a link to a professional fact-checking organisation.

In-note fact-checking If the community note performs an in-note fact-check by cross-referencing several sources and constructing a compelling argument.

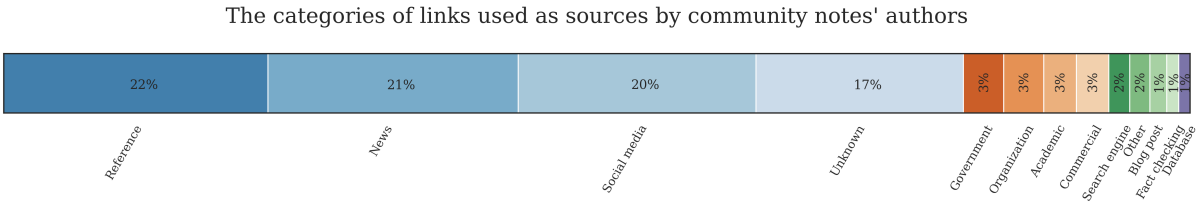

Figure 7: The categories of links used by Community notes' authors as a source, filtering for notes rated as "not helpful".

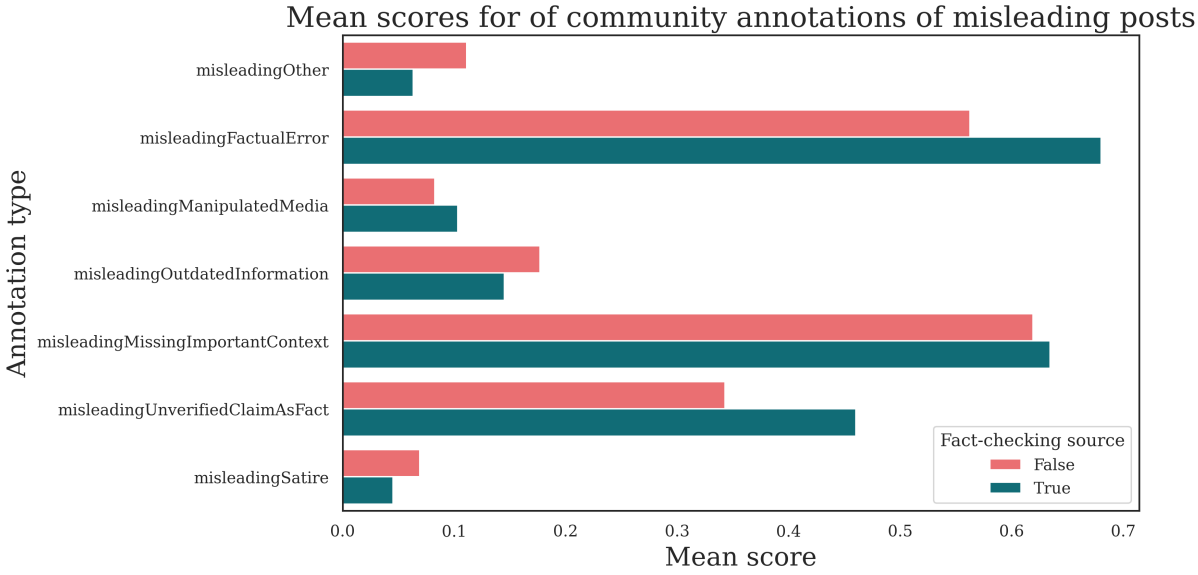

Figure 8: Mean scores of community annotations of misleading posts.

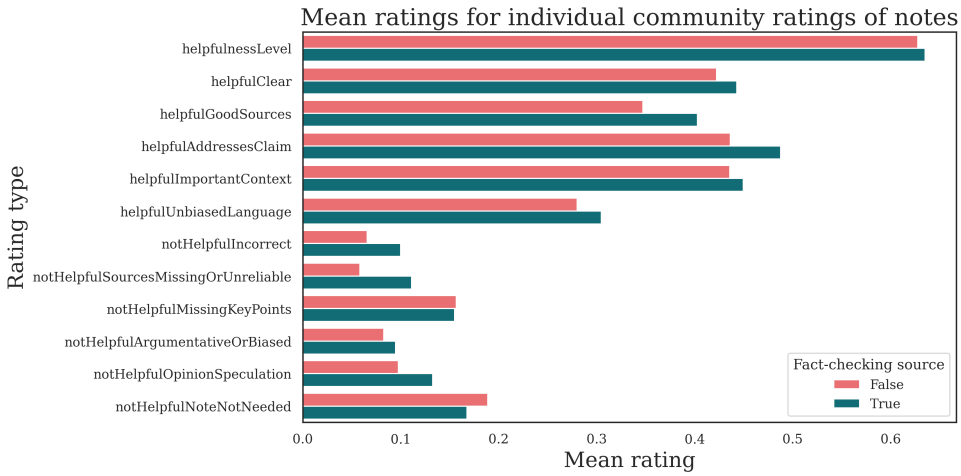

Figure 9: Community ratings of notes with and without fact-checking source.

```
SYSTEM PROMPT
You are a professional IT system who has a vast knowledge of the internet and its
    content. Your goal is simple, but very important: Classify URLs into
    categories. Choose only from the provided categories!


USER PROMPT
Read the following URLs.
Your goal is to categorize each url into one of the pre-defined categories.

Chose from the following list of categories:
Categories =
[
    "social media",  # Social media sites like Facebook, Twitter, Youtube etc.
    "news",  # Websites of news outlets or other organisations that report current
    events, such as the nytimes, the guardian, etc.
    "government",  # Government agencies and organisations, as well as websites
    related to policies and guidelines,  such as the CDC, department of education,
    FDA, etc.
    "academic",  # Academic sources, journals, and magazines, such as pubmed,
    nature, sciencedirect, etc.
    "blog post",  # Independent blog posts about various topics, including cooking,
    travel, home improvement, fandom, reviews, etc.
    "fact checking",  # professional fact checking organisations
    "database",  # Public databases such as google drive, archive.com, dropbox, etc.
    "commercial",  # Webpages of commercial organisations such as BMW, Delta, Nike,
    etc.
    "reference",  # Public resources such as encyclopedias, dictionaries, advocacy
    sources, guides, DIYs, statistics, religious sources, travel information, usage
    guidelines, Q&As, terms of services, etc.
    "organisation",  # non-commercial and non-government organisations such as WHO,
    the UN, Greenpeace, LA-Lakers, etc.
    "other",  # Any other website that does not fit into one of the previous
    categories.
    "unknown",  # if it is impossible to determine the category of the webpage.
]

Output format example:
[
    {
        id: <ID>,
        url: <URL>,
        category: <CATEGORY>,
    }
]


URLs:
<URLS>
```

Listing 1: The prompt used to classify URLs into categories.

```
SYSTEM PROMPT
You are a professional fact-checker who specializes in analyzing misinformation
    spread on social media.
Your goal is to analyse a tweet and a community note written about the tweet and
    decide whether the tweet spread misinformation related to a known conspiracy
    theory or a misleading wider narrative, and if so, which one is it.



USER PROMPT
Read the following tweets and community notes written about them.\nYour goal is to
    analyse them and decide whether each tweet spread misinformation related to a
    known conspiracy theory or a similar misleading wider narrative, and if so (and
    only if so!), which one.
Include your reasoning. Output the results as a json file. If a tweet does not
    relate to a conspiracy theory or a misleading wider narrative, output "none" in
    the json.

- Tweets *do not* discuss a wider narrative if the misleading information is tied
    to a specific singular event that is not connected to major topics on the
    public discourse.
They do discuss a wider narrative if the misleading information is tied to a known
    conspiracy theory or to major topics on the public discorse.

Chose from the following list of theories and wider narrative:
CONSPIRACY_THEORIES =
[
    September 11,
    October 7,
    the great replacement,
    COVID was intentionally spread,
    the COVID outbreak is fake,
    2020 election fraud,
    vaccines cause autism,
    5G towers,
    Russian invasion of Ukraine,
    flat earth,
    chemtrails,
    Q-Anon and deep state,
    Epstein files,
    Barack Obama was not born in the USA,
    Michelle Obama is a man,
    LGBT grooming,
    fluorite in the water,
    climate change,
    Holocaust denial,
    Hunter Biden and Ukraine,
    other,
]

Output format example:
[
    {
        id: <ID>,
        is_related_to_conspiracy: <True/False>,
        conspiracy: <CONSOIRACY (or None)>,
        reasoning: <REASONING>\
    }
]

    Tweets and notes:
    <TWEETS_AND_NOTED>
```

Listing 2: The prompt used to classify tweets and notes into broader narratives and conspiracy theories.

Original tweet	Note	Broader narrative	Discredit claim	Add Missing context	Highlight AI generated	Highlight edited media	Link to direct source	Link official source	Link scientific source	Link world knowledge	Link fact-checking	In-note fact-checking
No, not Ukraine It is Palestine, keep scrolling. https://t.co/Yt7R2Uwqda	The Palestinian woman pictured was arrested after at	☑	☐	☑	☐	☐	☐	☐	☐	☐	☐	☐
There have been many cancer cures, and all have been ruthle	The claim that cancer cures have been ruthlessly sup	☑	☐	☐	☐	☐	☐	☑	☑	☐	☑	☐
The Minnesota Vikings have named Sam Darnold starting QB	There is no verifiable source for this claim. He was on	☐	☑	☑	☐	☐	☐	☐	☐	☐	☐	☐
RIP Toby Keith. The latest victim of the Cancer Epidemic that began with the rc	Toby Keith was diagnosed with cancer in June 2022.	☑	☐	☑	☐	☐	☐	☐	☐	☐	☑	☐
https://t.co/hucThO2qcV	False: mixing primary and general election data The	☑	☐	☐	☐	☐	☐	☑	☐	☐	☑	☐
ELDERLY MAN CARRIES HIS 125 YEAR OLD GRANDFATHE	No human being has ever verifiable lived longer than	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐
If you don't understand calculus and differential equations , yo This is what it takes to generate proprietary trading signals to li There is no substitute for hard work https://t.co/PFe8GRCYtV	There is no calculus or differential equation on this pa	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fluoride is toxic waste leftover from processing aluminum... Long term exposure lower the average IQ by about 10 points...												

Figure 10: Our annotation setup.