# Proposal for Evaluating AI and Community Responses to Misinformation

Levi Goldberg

May 2025

## Abstract

This research project aims to evaluate the effectiveness of different methods for countering misinformation on social media, primarily focusing on X's Community Notes program. While Community Notes, a crowd-sourced fact-checking system, has shown promise in reducing engagement with misleading posts, its limitations, such as vulnerability to manipulation and inconsistent quality, necessitate further investigation. This project will compare crowd-sourced Community Notes with responses generated by an AI model prompted to counter the same misinformation and analyze the content of these responses, potentially comparing them to evaluations from professional fact-checkers, to determine the relative effectiveness of crowd-sourced and AI-generated fact-checks. Ultimately, this research seeks to identify the most effective approaches for correcting misinformation on social media platforms

## 1 Introduction

When the Community Notes program was launched on Twitter (now X) under the name of Birdwatch, its goal was to provide a transparent platform for misinformation detection and correction. If X users encountered a post that they thought contained misleading information, they could now add credibility indicators and a context providing note to attempt to fact check the post. These notes would then be rated by other users to determine how beneficial they were as fact checks, with the algorithm behind which notes would be displayed alongside a misleading post being open source. Further in line with their goal of transparency, all the data collected by X regarding Community Notes are available for download on their website.

Since these notes are provided by users, they may not reach the same level of accuracy as those provided by professional fact-checkers, but this system allows for greatly enhanced scalability and relies on "the wisdom of the crowds" - the notion that collective knowledge from a diverse group can exceed the knowledge of any individual [1]. On a platform like X with millions of daily posts, it would be impossible for professional fact checkers to review every one. The idea behind Community Notes is that a large group of individuals can review a much wider range of posts, while relying on their collective knowledge to create efficient fact checks.

Although the open-source nature of Community Notes may build trust with users, it also leaves the platform vulnerable to coordinated manipulation. An investigation by WIRED magazine found that groups of note contributors actively coordinate to upvote certain notes and downvote others to control which notes would be displayed alongside a post [2]. Although some of these groups were devoted to fighting misinformation, others were allegedly Russian trolls aiming to spread dissension.

Community Notes was originally designed to be used in conjunction with professional fact checkers, but since Twitter was rebranded to X, it has become the only method of fact checking on the platform. Although prior research in this problem space has found evidence that Community Notes are effective at detecting misinformation, it remains to be seen how useful they are for correcting it. Detecting fake news can help to limit its dispersion, but it does little to prevent it from being shared again in the future. As other social media sites, like Meta, are in the process of adopting similar crowd-sourced fact checking, it is necessary to evaluate the flaws with Community Notes and analyze potential alternatives.

## 2 Related Works

### 2.1 Prior Research on Community Notes

The natural first step in evaluating the effectiveness of Community Notes is determining if they are able to reduce the spread of misinformation. A study by Slaughter et al. on the diffusion of misinformation on X determined that posts that received context-providing notes had a significantly lower rate of engagement [3]. Engagement was measured based on the number of likes and comments on Community Note receiving posts compared to the estimated number if they had not been noted. This study also determined that there was a significant gap in the effectiveness of context-providing notes on political versus non-political posts, with notes on political posts being seen as more biased and less trustworthy. In this same vein, a paper by Drolsbach et al. found that professional fact checkers were viewed as more biased and less trustworthy than crowd-sourced fact checks [4]. Further, they determined that context-providing notes enabled a human to better identify misinformation than simple credibility-indicators that didn't explain why the post was misleading.

The research thus far is extremely favorable towards the effectiveness of Community Notes at countering misinformation. However, the fact that Community Notes was not designed to be the sole fact checking platform on X is still relevant. Borenstein et al. attempted to determine to what extent Community Notes would be able to replace professional fact checkers [5]. By using an LLM to annotate the hundreds of thousands of notes and the posts they were attached to in the Community Notes data set, they found that the highest-rated notes and notes responding to the most complex misinformation overwhelmingly included references to professional fact checking sources. This suggests that whether or not a professional is involved in fact checking a potentially misleading post, the research provided by these professionals is still crucial to successfully correcting misinformation. One limitation of this study is that the LLM used to annotate the dataset was unable to process X posts that contained non-text media, such as images or videos. This is a major problem for anyone working with the Community Notes data, since the scale of the data set necessitates automating parts of the analysis.

Another flaw with Community Notes is its susceptibility to organized efforts to influence its content. This is, in part, because the algorithm behind determining which notes would be displayed alongside a misleading post is open source, allowing groups of users to "game" the system. One possible solution to this is a new and more opaque algorithm, as posited by De et al. [6]. These researchers created a framework for AI-generated "Supernotes" that synthesize the content of existing notes on a given post to provide a concise, accurate fact check. This system could still be gamed, but the algorithm behind Supernotes includes a simulated jury trained on the data of which Community Notes have been rated the most helpful over time to rate several different Supernote candidates and promote the one that is most likely to be rated the most helpful. De et al.'s analysis of the effectiveness of these Supernotes found that users rated them as more helpful than traditional notes. The primary drawback of this framework for improving Community Notes is that a Supernote cannot be generated until several notes have already been written by human users. Furthermore, another deficiency of the entire Community Notes platform is that even though it is more scalable than professional fact checking, there are simply not enough active contributors to review every single post on X.

## 2.2 The Format of Fact Checks

Having established that Community Notes are generally effective despite some limitations, we now explore potential improvements. Notes include quick-to-provide credibility indicators and a more time-consuming context statement. If credibility indicators alone proved effective in countering the spread of misinformation, the platform's scalability could improve. The previously discussed research from Drolsbach et al. suggested that credibility indicators on their own were not as effective as a context-providing statement, which was further confirmed in a study by Lu et al. [7]. For this project, the researchers used AI to attach credibility indicators to misleading posts to study how they would affect the diffusion of the post. They found that although these indicators helped some users to identify misinformation, they did little to lower the rate of engagement with or the spread of misleading information. They found some evidence that credibility indicators are more effective at changing viewers' beliefs when there is also social influence (i.e., comments from other users stating the post is misleading) present, which further supports the notion that a combination of credibility indicators and context-providing responses is the most efficient way of countering misinformation. This conclusion was also supported by Ecker et al., who investigated whether credibility indicators on their own could backfire and cause misinformation to spread faster, which they found no evidence for [8]. Their research also suggested that short fact checks that succinctly explained why the information is incorrect were more effective at countering misinformation than longer, more in-depth explanations.

Further regarding the format of these context statements, a paper by Burel et al. in which, rather than attaching a note to misleading posts on X, the authors designed a bot to message the post's creator to research how different types of responses would be received by the people spreading misinformation [9]. They found that spreaders of misinformation were most likely to respond positively to being fact checked if the statement was phrased politely.

One format of responding to misinformation used by previously cited papers, including Burel et al. and Ecker et al., is a narrative fact check, in which a story is told to explain why the information is misleading. There was some belief that these may be more effective than non-narrative fact checks, due to the way a narrative format can enhance comprehension and retention. However, a different study by Ecker et al. determined that when the information provided in a narrative and non-narrative refutation has minimal differences, there is no significant difference in the refutation's effectiveness at countering misinformation [10].

Thus, the research suggests that the most effective format for a fact check is a combination of credibility indicators and a refutation of the misinformation that is short and positively phrased, with no difference between narrative or non-narrative formats. A study by Pyreddy et al. into the differences in the emotions and sentiments expressed by humans versus AI found that AI-generated responses to a given prompt are generally more positive, consistent, and concise, whereas human-generated responses had more varied tones, word choice, and length [11]. Hence, one method of improving social media fact checking could be to train an AI to write Community Notes-style responses. Since an AI could consistently write responses in the most effective format and could provide citations to professional fact checkers, it seems likely that AI-generated credibility indicators and refutations could prove a potent method of detecting and correcting misinformation.

## 3 Preliminary Design

The open-source data from Community Notes is stored in Tab-Separated Value (TSV) files, each of which contains information about approximately 100,000 notes. As of May 2025, there are twenty files of data, though more are added with the creation of new Community Notes. Each row in each file contains information about a specific Community Note, including the ID of the note, when it was created, how helpful it was rated by other Community Notes users, and the ID of the original X post that the note was attached to.

The note ID and post ID can both be used to find a JSON webpage containing all of the relevant information about the note/post, including the text string and other media. I will use the Requests library in Python to scrape through these JSON pages and extract the original text of each note and the text of the post it was fact checking.

Once I have all of this data added to the dataset, it will need to be meticulously cleaned to remove any posts consisting mostly of non-text media, such as photos or images, or non-English text strings. This is because these posts would require an additional step of data processing to convert them into a form that would be easily understandable for an LLM, and there is such a large amount of data available ($\sim$ 2 million notes) that this step would simply add to the scale of the project.

Next, I will use the OpenAI Python library to send a query to ChatGPT that includes the text of the note being fact checked and instructions to generate another fact check in the format that has been shown to be the most effective at correcting misinformation. These AI-generated fact checks will then be appended to the dataset alongside the original fact check from Community Notes.

In order to use the OpenAI API, I will need to purchase GPT tokens. One token is approximately four characters, so assuming that the average word is about four characters (five including spaces) and the average post is about 80 words long, I would need approximately 100 tokens per post. Although there are approximately two million notes in the data set, these will not all require their own query for ChatGPT. Many of the notes in the data set were created in response to the same post, and many of the posts will be removed from the data set during the cleaning process. A very rough estimate is that I will need to query ChatGPT about 500,000 posts, which would require approximately 50,000,000 tokens. Per OpenAI's pricing website, their GPT 4.1 model costs $2 per million tokens and their GPT 4.1 mini model costs $0.4 per million tokens. Thus, using GPT 4.1 would cost approximately $100, and using the GPT 4.1 mini would cost approximately $20 in tokens, both of which are affordable.

Finally, I hope to find a professional fact checker to write context providing statements for a curated subset of the posts Community Notes were attached to. This subset will likely be comprised of the posts that had the highest number of notes attached to them. Professional fact checkers are considered to be the optimal method for correcting misinformation, so this will create a metric that the crowd-sourced and AI-generated fact checks can be compared to. This professional will likely also require payment for their services, the amount of which will depend on the number of posts they evaluate. I have contacts at the Toda Peace Institute, which has worked on the issue of misinformation on social media, who may be able to get me in contact with professional fact checkers.

## 4 Analysis of Risks

One major risk is that I will be unable to find a professional fact checker to use as a baseline for comparison. My contingency for this scenario would be to instead compare the AI-generated and crowd-sourced fact checks to the highest-rated Community Note on any given post. This would only be possible on X posts that have had sufficient engagement to have a highest rated note, which would limit the data available, but would still hopefully produce valuable results.

## 5 Preliminary Evaluation

In order to compare the effectiveness of crowd-sourced and AI-generated fact checks, I will need to determine a quantitative measure of effectiveness. Since professional fact checkers are the ideal method for fact checking, but are simply not scalable to an entire platform, one measure of effectiveness would be the similarity between a given fact check from either a Community Note or an AI and a fact check provided by a professional. Gomaa et al. discuss several different methods for comparing the similarity of two strings of text [12]. Of these methods, the most useful one for the issue at hand would likely be a corpus-based similarity method in which word embeddings are created using a large corpus, and then the similarity of two strings is determined from the embeddings. An open-source program for a corpus-based similarity method is available on Github [13]. This program calculates the embeddings for two strings of text and highlights sections of them that contain similar meanings, which will hopefully allow me to calculate the similarity between the professional's fact check and the open-source/AI-generated fact check. The more similar a given fact check is to the professional's, the more successful that fact check will be considered.

## 6 Expected Contributions

The primary result of this research will be an analysis of the similarity of crowd-sourced and AI-generated fact checks to those created by experts. This will provide insight into the effectiveness of crowd-sourced fact checking platforms, like Community Notes, and potentially provide proof of concept for using AI to generate fact checks in a specific style. Auxiliary contributions could include a trend analysis for Community Notes, investigating things like what type of posts (i.e., political, entertainment, etc.) have the highest number of notes, the most effective and least effective fact checks on average, and the highest similarity between crowd-sourced and AI-generated fact checks.

## 7 Proposed Timeline

This project will be completed over a 15-week semester. The primary deliverables will be a technical report, a poster presentation, and a project demonstration video. Each of these will summarize my project in different ways that will include a description of the scope of my project, the methods involved, and the results (including data visualizations), as well as a data architecture diagram and a graphical abstract of the processes. These will all be in development while I am working on the various research aspects of my project, with a potential timeline shown in the table below.

As for the actual research I will be conducting, I aim to have all of the data preprocessed, extracted, and processed in the first 6 weeks of the semester. The next 6 weeks would then be dedicated to analyzing the data and producing visualizations. This would leave me with 3 additional weeks at the end of the semester to tie up any loose threads and finalize my deliverables.

### Deliverables Breakdown

| **Week** | **Deliverables** |
|---|---|
| 1 | |
| 2 | Version 0 of the Technical Report |
| 3 | Version 0 of the Data Architecture Diagram |
| 4 | Version 0 of the Graphical Abstract |
| 5 | Version 1 of the Data Architecture Diagram and Graphical Abstract |
| 6 | Version 1 of the Technical Report |
| 7 | Version 2 of the Data Architecture Diagram and Graphical Abstract. Version 0 of the Poster |
| 8 | Version 0 of the Project Demonstration Video |
| 9 | Version 2 of the Technical Report |
| 10 | Version 1 of the Poster |
| 11 | Version 3 of the Data Architecture Diagram and Graphical Abstract |
| 12 | Version 1 of the Project Demonstration Video |
| 13 | Version 3 of the Technical Report |
| 14 | Version 2 of the Poster and Project Demonstration Video |
| 15 | Final Versions of All Deliverables |

# 8 Works Cited

[1] Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday.

[2] Elliott, V., & Gilbert, D. (2023). Elon musk's main tool for fighting disinformation on x is making the problem worse, insiders claim.

[3] Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). Community notes moderate engagement with and diffusion of false information online. *arXiv preprint arXiv:2502.13322*.

[4] Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS nexus*, *3*(7), pgae217.

[5] Borenstein, N., Warren, G., Elliott, D., & Augenstein, I. (2025). Can community notes replace professional fact-checkers? *arXiv preprint arXiv:2502.14132*.

[6] De, S., Bakker, M. A., Baxter, J., & Saveski, M. (2024). Supernotes: Driving consensus in crowd-sourced fact-checking. *arXiv preprint arXiv:2411.06116*.

[7] Lu, Z., Li, P., Wang, W., & Yin, M. (2022). The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27.

[8] Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British journal of psychology*, *111*(1), 36–54.

[9] Burel, G., Tavakoli, M., & Alani, H. (2024). Exploring the impact of automated correction of misinformation in social media. *AI Magazine*, *45*(2), 227–245.

[10] Ecker, U. K., Butler, L. H., & Hamby, A. (2020). You don't have to tell a story! a registered report testing the effectiveness of narrative versus non-narrative misinformation corrections. *Cognitive Research: Principles and Implications*, *5*, 1–26.

[11] Pyreddy, S. R., & Zaman, T. S. (2025). Emoxpt: Analyzing emotional variances in human comments and llm-generated responses. *arXiv preprint arXiv:2501.06597*.

[12] Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, *68*(13).

[13] tanzir5. (n.d.). Tanzir5/alignment$_{t}ool$2.0: Align two text sequences. https://github.com/tanzir5/alignment_tool2.0